MECHANISM OF FOREIGN DNA RECOGNITION AND DEGRADATION BY THE TYPE I CRISPR SYSTEM IN *ESCHERICHIA COLI*

by
Sabin Mulepati

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
April, 2014

Abstract

The Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) immune system is used by bacteria and archaea to gain immunity from mobile genetic elements like phage DNA and plasmids. In *Escherichia coli*, small RNA derived from its CRISPR loci (crRNA) are integrated into a large ribonucleoprotein complex called Cascade, which is then used as a surveillance complex to find foreign DNA based on sequence complementary. Previous studies suggested that Cascade recruits an additional nuclease-helicase protein called Cas3 to silence foreign DNA in order to gain immunity. To understand the roles of Cascade and Cas3, we carried out structural and biochemical studies on both of these essential components of the CRISPR immune system.

Here, we report the crystal structure of the Cascade complex from *E. coli* bound to its target DNA to 3.03 Å. The structure reveals a DNA-RNA hybrid at the core of the complex, forming a heavily distorted, discontinuous, arched-ladder that locally forms short A-form-like duplexes. Bases in both strands of the hybrid are flipped out at regular intervals due to the organization of the protein subunits in the complex. The structure presented here shows how Cascade-like complexes have evolved to form a distorted hybrid that is likely primed for recruitment of Cas3 for further degradation of the invasive DNA.

We also report the crystal structure of the HD nuclease domain of Cas3 from *T. thermophilus*, and characterize its nuclease active site. Based on additional biochemical analysis, we show that the HD nuclease likely uses a two-metal-ion-dependent cleavage mechanism. Furthermore, using individually purified Cascade and Cas3 from *E. coli,* we reconstituted CRISPR-mediated plasmid degradation *in vitro*. Analysis of this

reconstituted assay suggests that Cascade recruits Cas3 to a single-stranded region of the DNA target exposed by Cascade binding. Cas3 then nicks the exposed DNA. Recruitment and nicking is stimulated by the presence, but not hydrolysis, of ATP. Following nicking, and powered by ATP hydrolysis, the concerted actions of the helicase and nuclease domains of Cas3 proceed to unwind and degrade the entire DNA target in a unidirectional manner.

Taken together, the results of our study explain how foreign DNA is identified and degraded by the CRISPR immune system, and provide a solid framework for future studies.

Advisor: Scott Bailey

Reader: Daniel J Leahy

Acknowledgement

I am very fortunate to have had such a challenging yet exciting experience as a graduate student at Johns Hopkins. I am grateful to the numerous people who helped with my research and shaped my path, and would like to offer particular thanks to the following individuals.

Firstly, I would like to thank my advisor, Scott Bailey, for taking me as his first graduate student and letting me work on a challenging project. He was supportive and patient throughout, and motivated me without being overbearing. I am also indebted to my thesis committee members—Daniel Leahy, Rachel Green, and Jon Lorsch—and Jie Xiao and Greg Bowman, who later joined my committee. Thank you for your invaluable support throughout my graduate career, and for all the suggestions regarding my future after Hopkins.

Many thanks to Jennifer Kavran, who taught me a lot during one of my first-year rotations, and also introduced me to my eventual advisor. I would like to thank Gabriel Brandt and Brian Learn for their day-to-day advice regarding experiments, and Justin Smith for helping me initially get settled in the lab. I am thankful to Jürgen Bosch for generously providing the heavy-atom clusters that were useful in my crystallography experiments, and Irimpan Mathews at the Stanford Synchrotron Radiation Lightsource for collecting X-ray diffraction data from the SeMet Cascade crystals. I am also thankful to all the members of the Bailey and Bosch labs over the years, and am especially indebted to members who worked directly with me, and in turn, taught me invaluable lessons. I thank members of my PMB class (Bobby Trachman and Andrew Buller) for numerous discussions about experiments and life in general.

Lastly, I would not be where I am today without my family. Thank you to my parents, Sharmila and Rabindra, for supporting me and letting me choose my own path. Thanks to my brother Nabin, and in-laws, Robert and Karen Stauffer, for sharing many wonderful times together and always being eager to hear about my research. Finally, I would not be here without my lovely wife, Janessa Mulepati, for all her love and support. My graduate career was not without its ups and downs, and her ever-cheerful attitude certainly helped me through some of its hardest times. Thank you for always being there.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

CRISPR immune system

Bacteria and archaea constitute a major portion of the earth's biomass, and are found in almost all of its habitats (Breitbart et al., 2005). These unicellular prokaryotes can easily evolve and adapt to changing environmental circumstances through their ability to exchange genetic material by a process called horizontal gene transfer (Koonin and Wolf, 2008). This exchange of genetic material can occur through three different processes: conjugation (from plasmids), transformation (between species), and transduction (from bacteriophages) (Juhas et al., 2009; Koonin and Wolf, 2008). As a result, bacterial and archaeal populations are extremely diverse. Their rapid diversification is evident in the low overall sequence conservation among different strains of *Escherichia coli*. Among the genomes of 61 different *E. coli* strains sequenced, only about 10% of genes are conserved, with the rest of the genes being constantly exchanged (Lukjancenko et al., 2010).

While rapid exchange of genetic material is evolutionarily beneficial, it also makes the recipients constantly prone to harmful genetic elements like phages and plasmids. This is even more alarming, considering that the majority of phages infect bacteria (Breitbart et al., 2005). One way that bacteria and archaea have evolved to deal with such foreign genetic elements is by means of a recently discovered immune system called CRISPR (Mojica et al., 2000; Barrangou et al., 2007; Sorek et al., 2013). CRISPR,

which stands for Clustered Regularly Interspaced Short Palindromic Repeats, is found in ~ 90% of archaeal and ~40% of bacterial species (Makarova et al., 2006).

CRISPR is a RNA-based immune system, and works by creating a genetic record of past infections that can be retrieved upon reinvasion by phages and plasmids to counteract their harmful effects. These records are stored at a particular CRISPR locus in the host genome as 20-50 nucleotide long, invader-derived sequences called 'spacers'. These spacers are interspaced by identical 'repeat' sequences that are about 20-40 nucleotides long. A typical arrangement of these elements in a CRISPR locus is shown in figure 1.1.

The number of CRISPR loci in a single chromosome and their spacer content is variable and can range between 1-18 (Pourcel et al., 2005), with the longest CRISPR locus harboring as many as 374 repeat-spacer sequences (Marraffini and Sontheimer, 2010). CRISPR loci are not static and change over time under constant selection pressure, causing the loci to have variable spacer makeup (Pourcel et al., 2005). The dynamics of the CRISPR loci are described in detail later in this chapter.

The identical repeat sequences that make up a CRISPR locus were first noticed by Ishino and colleagues in *E. coli* in 1987. However, the conservation of such a feature among different bacteria and archaea was not appreciated until much later, as more genome sequences became available (Mojica et al., 2000). Soon after, it was realized that some of the spacer sequences were homologous to DNA segments from known plasmids and phages, suggesting their extra-chromosomal origins (Mojica et al., 2005; Pourcel et al., 2005; Bolotin et al., 2005). This led to the hypothesis that the CRISPR system is an adaptive immune system against foreign invaders, which was later

experimentally proven in the case of *Streptococcus thermophilus*. This study found that, when challenged with new phages, new spacer sequences that were complementary to short segments of the phage genome (called protospacer) were actively incorporated into the CRISPR locus (Barrangou et al., 2007). While spacers are complementary to foreign genetic elements, it has been shown that a small subset of spacers (<5%) is also homologous to sequences in the host chromosome (Stern et al. 2010).

CRISPR loci are most often flanked by a diverse set of CRISPR-associated (*cas*) genes. The Cas proteins encoded by these genes are essential to mounting a proper CRISPR response (Makarova et al., 2006). Brouns and colleagues showed that, in *E. coli,* the CRISPR locus is transcribed and processed by a set of Cas proteins and then packaged into a large surveillance complex (405 kDa) called CRISPR-associated complex for antiviral defense (Cascade). This complex is able to silence foreign DNA in the presence of an additional protein called Cas3 (Brouns et al., 2008). These results and earlier bioinformatics studies suggested that the CRISPR system, although bearing no sequence homology, is functionally analogous to the RNA interference (RNAi) systems found in higher eukaryotes (Makarova et al., 2006; Carthew et al. 2009; Marraffini and Sontheimer, 2010).

The CRISPR immune response has been divided into three distinct stages: (i) adaptation, (ii) CRISPR RNA biogenesis, and (iii) interference (Figure 1.2) (Sorek et al., 2013). In the adaptation stage, the system is able to incorporate new spacer sequences into the CRISPR loci from phages. Upon reinvasion by the same phage, the CRISPR loci are then transcribed and processed into mature crRNA (CRISPR RNA) in the second stage to form surveillance complexes like Cascade. In the interference stage, the crRNA

is used to locate and silence foreign DNA by means of sequence complementarity. Each of the three stages needs to be carried out for a proper CRISPR response.

Of the three CRISPR stages, my thesis is mainly focused on the third (interference) stage. The findings by Brouns and colleagues in 2008—that the Cascade surveillance complex is formed to silence invader DNA—formed the basis of my thesis. The two main questions that I have focused on as part of my dissertation are:

1. How are the different components of the CRISPR system able to recognize foreign DNA?
2. What is the fate of the foreign DNA, once recognized by the CRISPR system?

Since I started my thesis work in 2009, the CRISPR field has seen rapid development (Sorek et al., 2013). We now know that, in addition to DNA, certain CRISPR subtypes exclusively target RNA (Hale et al., 2009; Zhang et al., 2012; Bailey, 2013). Other subtypes, while being more simplistic in their protein makeup, utilize additional RNA components to carry out DNA targeting (Deltcheva et al., 2009). In addition, CRISPR has quickly evolved into a versatile genome-engineering tool (Mali et al., 2013; Jinek et al., 2013; Wang et al., 2013). I summarize some of the recent developments below.

Types of CRISPR systems

As expected in a prokaryotic system, CRISPR loci are extremely diverse in terms of their spacer-repeat makeup and their related Cas proteins (Makarova et al., 2006). Cas proteins have many predicted functions, and consist of RNases, DNases, helicases, integrases, polymerases, and RNA-binding proteins (Jansen et al., 2002). Based on the

overall conservation of the different *cas* genes, the CRISPR system has been divided into three main types—I, II and III (figure 1.3) (Makarova et al., 2011). The signature genes of the three types are *cas3*, *cas9,* and *cas10* respectively, and the three types are not mutually exclusive within a particular species. Some prokaryotes have one or more type present within a single chromosome, but the interplay between the types is not understood.

While the adaptation stage and the proteins involved are conserved in all three CRISPR types, the crRNA biogenesis and interference stages (and their respective proteins) are quite diverse, and can be further subdivided into 10 subtypes as shown in Figure 1.4. Details concerning the three main types are as follows.

The type I CRISPR system, the most prevalent of the three types, is present in both bacteria and archaea. Cas3, the conserved protein of this type, is involved in the interference step, and is thought to silence invader DNA (Brouns et al., 2008; Jore et al., 2011; Semenova et al., 2011). Cas3 usually consists of an N-terminal nuclease domain and a C-terminal helicase domain. The type I CRISPR system can be further divided into six subtypes (subtypes IA-IF), and the two domains of Cas3 also exist as separate proteins in some of the subtypes. Besides Cas3, additional Cas proteins are involved in the targeting of invader DNA, since Cas3 is not able to scan for protospacers. In all of the subtypes, different sets of Cas proteins process the CRISPR locus, and form large ribonucleoprotein surveillance complexes that are able to recognize foreign DNA. In the type IE subtype found in *E. coli,* this complex is called Cascade, and both Cas3 and Cascade are crucial for DNA targeting.

The type II CRISPR system is exclusive to bacteria, and is the most simplistic type in terms of its protein make-up. Like the type I CRISPR system, this type targets DNA as well, but Cas9 is the signature protein. Cas9 is a large protein (~130 kDa) with conserved HNH and RuvC domains (Makarova et al., 2011), and is involved in the processing of the CRISPR locus as well as the targeting of DNA. The type II system is unique in that, in addition to the CRISPR locus and Cas9, it also requires a trans-activating crRNA (tracrRNA) and endogenous RNase III in the CRISPR RNA biogenesis stage (Deltcheva et al., 2011). Cas9 first forms a duplex between crRNA and tracrRNA and then uses the crRNA to locate and cleave the target DNA. Furthermore, its HNH and RuvC nuclease domains cleave the complementary and the non-complementary strand of the target DNA respectively (Jinek et al., 2012). The type II CRISPR system has been further divided into three subtypes (Figure 1.4).

The type III CRISPR system is most prevalent in archaea, and its two subtypes have been shown to cleave both DNA (type III-A) and RNA (type III-B) (Hale et al., 2009; Marraffini et al., 2008). Cas10 is the signature protein of this type, and has been implicated in target interference (Makarova et al., 2011). Cas6, a metal-independent endonuclease, is also present in both subtypes, and is involved in the initial processing of crRNA. Electron microscopy studies suggest that the Cas proteins in the type III CRISPR also form larger Cascade-like complexes found in the type I CRISPR system. (Rouillon et al., 2013; Staals et al., 2013).

Three stages of CRISPR

As mentioned previously, the CRISPR system has three stages, and each carries out a specific function as described below (Figures 1.2 and 1.3). While the adaptation stage is conserved among the three CRISPR types, the second (CRISPR RNA biogenesis) and third (Interference) stages are quite different due to the diversity of the Cas proteins involved. Since my dissertation is primarily concerned with the interference stage of the type I CRISPR system, I focus on this type in the following sections.

Adaptation

The adaptation step is the most conserved, yet the least understood among the three CRISPR stages. In this step, foreign DNA is recognized upon invasion, and short fragments of the invader DNA (termed protospacer) are incorporated into the host genome at its CRISPR locus (Barrangou et al., 2007). The addition of spacer sequences always occurs in the region directly downstream of the AT-rich 'leader' sequence (figure 1.1). This results in a polarity within a CRISPR locus, where the most recently incorporated spacers are next to the leader sequence, while the older spacers are gradually shifted away from the leader (Pourcel et al., 2005). With such an arrangement of spacers, CRISPR loci act as a chronological history of pathogenicity for a particular host. However, the CRISPR locus is not an infinite list. The older, less frequently used spacers are removed under selection pressures, resulting in a degenerate end at the opposite side of the locus (away from the leader sequence).

The leader sequence also consists of promoter elements that control expression CRISPR/Cas components. While transcription of the CRISPR locus is not required for

the adaptation step, elements within the leader are essential for spacer acquisition. Also required during this step are two conserved proteins, Cas1 and Cas2. Both of the genes encoding these proteins are hallmarks of the CRISPR system, and are conserved in all the 3 CRISPR types. Cas1 is a non-specific, metal-dependent, double-stranded DNase that generates ~80 bp products (Wiedenheft et al., 2009; Babu et al., 2011). Cas2, on the other hand, has both metal-dependent, double-stranded, DNase and single-stranded RNase activities (Nam et al., 2012; Beloglazova et al., 2008). Cas1 and Cas2 are most probably involved in the processing of protospacers that are incorporated as spacer sequences.

During spacer uptake, the first repeat sequence proximal to the leader sequence serves as a template, and is duplicated for every spacer acquisition (Swarts et al., 2012; Yosef et al., 2012; Díez-Villaseñor et al., 2013). A single repeat sequence was shown to be sufficient to initiate spacer incorporation (Swarts et al., 2012; Yosef et al., 2012).

Protospacers are very short (only 30-50 nucleotides) compared to the much larger genomic landscapes of invaders that need to be surveyed by the CRISPR adaptation machinery during spacer uptake (~40 Kb in the case of T7 phage). This raises critical questions regarding discriminations that the CRISPR system needs to make during protospacer selection. *In silico* analysis showed that 2-5 nucleotide sequences adjacent to protospacers in the invader genome are conserved within a species (Mojica et al., 2009). This sequence, called the protospacer adjacent motif (PAM), has been experimentally shown to be essential for spacer uptake (Yosef et al., 2012). In *E. coli* K12 strain, robust spacer uptake was demonstrated from plasmids containing a protospacer and a PAM sequence (5'-CWT-3') with only the Cas1 and Cas2 proteins overexpressed (Yosef et al.,

2012). As I describe later, the PAM sequence also plays a crucial role during the interference step.

Besides the PAM, additional DNA motifs have recently come to light from bioinformatics approaches (Yosef et al., 2013). These studies have shown that the presence of conserved 2-nucleotide sequences termed Acquisition Affecting Motif (AAM) on the opposite side of the protospacer (with respect to the PAM) results in higher efficiency during spacer uptake.

In additional to scanning for different sequence motifs, there is increasing evidence that spacer acquisition is synchronized to the other CRISPR stages. Besides Cas1 and Cas2, additional Cas proteins are not required for spacer uptake (Yosef et al., 2012). However, the presence of other CRISPR components like Cascade and Cas3 (involved in the interference stage) has shown to stimulate spacer acquisition from targets with a protospacer (Datsenko et al., 2012). In *E. coli* K12 strain, spacers are preferentially acquired from plasmids that already contain a 'known' protospacer (Datsenko et al., 2012; Swarts et al, 2012). Replicons containing *cas* genes have also been shown to be more efficient spacer donors (Díez-Villaseñor et al., 2013). This results in multiple spacers against the same invader, which may result in increased immunity.

CRISPR RNA biogenesis

CRISPR RNA biogenesis involves the expression and processing of CRISPR loci upon subsequent attack by known invaders (homologous to spacer sequences). The CRISPR loci are expressed as single RNA transcripts consisting of spacers against different invaders. The related *cas* genes are translated as well during this step. The

expression of the CRISPR locus is driven by the leader sequence (Brouns et al., 2008; Pul et al., 2010). The leader sequence is generally AT-rich and consists of promoter elements needed for expression of the *cas* genes (Pul et al., 2010).

What stimulates CRISPR expression is still unclear and seems to vary among species. It most cases, a basal level of CRISPR expression seems to be always present and is induced under stressful conditions (Agari et al., 2010; Juranek et al., 2012). However, in the case of *E. coli* K-12, the promoter elements of the CRISPR locus are strictly repressed by a global transcriptional repressor called histone-like nucleoid protein (H-NS) (Pul et al., 2010) and are activated only upon stress conditions. Although it is known that LeuO, a transcriptional activator, neutralizes repression by H-NS, the exact signals that lead to H-NS de-repression by the action of LeuO to achieve CRISPR expression is not clearly understood. Some of the known signals that lead to CRISPR expression are envelope stress (membrane protein unfolding), ionic strength, and UV light (Perez-Rodriguez et al., 2010; Sorek et al., 2013).

In the type I-E CRISPR subtype, the CRISPR transcript (pre-crRNA in Figure 1.5) is processed by a set of Cas proteins into a mature crRNA that is 61 nucleotides in length (Figure 1.5). In this subtype, the *cas* gene cassette encodes eight proteins, and five of these proteins (Cse1, Cse2, Cas7, Cas5, and Cas6e) are involved in the processing step (Figure 1.4). Throughout this manuscript, these five proteins in the type I-E system (found in *E. coli*) will be referred to as CasA, CasB, CasC, CasD, and CasE respectively (Figure 1.5). Brouns and colleagues (2008) showed that these five proteins process the pre-crRNA, such that, each crRNA contains a single spacer-repeat sequence (figure 1.5). CasE is the endonuclease responsible for the recognition of the hairpin structure formed

by the repeat sequence and their subsequent cleavage at the base of each hairpin (Brouns et al., 2008; Haurwitz et al., 2010). Upon cleavage, CasE stays bound to the repeat element of the crRNA and possibly recruits the other proteins (Cas A-D) to form a 405-kDa Cascade complex (Haurwitz et al., 2010; Jore et al., 2011). Cascade consists of CasA, CasB, CasC, CasD, CasE, and crRNA in a defined stoichiometry of 1:2:6:1:1:1 (Jore et al., 2011).

Cryo-EM structures of *E. coli* Cascade have revealed the organization of the five subunits around a core formed by the crRNA (Figure 1.6) (Wiedenheft et al., 2012). Overall, Cascade has a sea horse-shaped architecture. The crRNA forms the spine of such a structure with the six CasC subunits wrapped around the crRNA in a helical filament. The head is capped by CasE subunit at the 3'-end of the crRNA, and CasA and CasD subunits cap the tail at the 5'-end of the crRNA. Two copies of CasB sit on the inner surface of the CasC-crRNA spine, directly connecting the head (CasE) and the tail of the complex (CasA and CasD). At the end of this stage, different Cascade complexes, each with a different crRNA, are formed. Theses stable complexes are now able to scan for foreign DNA based on sequence complementarity to their spacer sequences.

The CRISPR biogenesis step is not conserved, and there are significant differences between the three CRISPR types since different sets of Cas proteins are used during processing of the long pre-crRNA. In the type II system, the cas9 protein is involved in the processing of the CRISPR locus with the help of an additional RNA component called tracrRNA and endogenous RNaseIII nuclease. In case of the type III system, Cas6 recognizes and cleaves the repeat sequences in CRISPR loci. Cas6,

however, is not part of a Cascade-like Cmr/Csm complexes that scan for foreign RNA/DNA in this CRISPR type.

Interference

The interference stage in the type I CRISPR system is the primary focus of my thesis. In the type I-E CRISPR system, both Cascade and Cas3 are involved in targeting foreign DNA for degradation. Cascade is crucial to this step, as it is able to find protospacers based on complementarity to its crRNA. Scanning force microscopy experiments show that Cascade readily recognizes and locates to sites of the protospacers on the invader DNA (Westra et al., 2012). Upon binding to a dsDNA, Cascade melts the target and interacts extensively with both its complementary and non-complementary strands to form an R-loop (Jore et al., 2011; Wiedenheft et al., 2011). The crRNA forms Watson-Crick base-pairing with the complementary strand while part of the non-complementary strand is exposed as a single-stranded region, possibly acting as a signal for Cas3 recruitment (Jore et al., 2011). Cascade-DNA base-pairing has been proposed to nucleate at the proximal 5'-end of the protospacer spanning nucleotides 1-5, 7, and 8 before proceeding over the complete protospacer (Semenova et al., 2011; Wiedenheft et al., 2011). This binding event is ATP-independent (Jore et al., 2011).

As in the adaptation stage, PAM is essential during the interference stage as well. Single mutations in the PAM abolish CRISPR interference *in vivo* (Semenova et al., 2011). *In vitro* binding experiments have also shown that mutations in the PAM alone severely disrupt Cascade binding to a protospacer. The CRISPR loci themselves do not have a PAM, and hence escape CRISPR-based immunity. Recent single-molecule studies

in the type II CRISPR system have shown that PAM recognition is the first 'obligate' step in target interference (Sternberg et al., 2014). Thus, the PAM is crucial in distinguishing genomic (no PAM) from extra-chromosomal target DNA (with PAM) (Mojica et al., 2009; Semenova et al., 2011).

Cryo-electron microscopy structure of Cascade bound to a complementary RNA shows that Cascade undergoes a concerted, conformational change involving CasA, CasB, and CasE, upon target binding (Wiedenheft et al., 2011). Despite these advances in our understanding of Cascade, the roles of many of its subunits remain unclear. This led us to investigate the possible roles of the different subunits of Cascade, and the results are presented in chapters 2 and 5.

Cascade lacks any detectable DNase activity (Jore et al., 2011) but is thought to recruit the Cas3 nuclease upon target recognition (Brouns et al, 2008; Jore et al., 2011; Semenova et al., 2011). Like Cascade, Cas3 is also indispensable for the phage-resistant phenotype of *E. coli* (Brouns et al., 2008). In the type I-E CRISPR subtype in *E. coli*, Cas3 is composed of an N-terminal HD nuclease domain followed by a super-family-2 DEXH-helicase domain (Haft et al., 2005; Makarova et al., 2006). In some subtypes, the HD nuclease (Cas3'') and the helicase domains (Cas3') are encoded separately but are still expected to have synchronized activities resulting in the destruction of the target DNA. Cas3 does not interact with target DNA in the absence of Cascade, and both its HD nuclease and helicase domains are essential for target interference (Westra et al., 2012).

The HD domain catalyzes the ssDNA endonuclease activity of Cas3, and the helicase domain catalyzes the ATP-dependent unwinding of dsDNA and RNA-DNA duplexes (Sinkunas et al., 2011). In contrast, *Sulfolobus solfataricus* Cas3" has distinct

substrate specificity as it cleaves double-stranded but not single-stranded DNA or RNA (Han and Krauss, 2005). These discrepancies led us to investigate the structural and biochemical properties of the Cas3 nuclease domain ($Cas3^{HDdom}$) from *T. thermophilus* (chapter 3), and the mechanism by which *E. coli* Cascade and Cas3 come together to foreign DNA degradation in the interference stage (chapter 4).

Figure 1.1. A representative CRISPR locus. The repeat sequences are shown as black diamonds, and the spacer sequences as yellow rectangles. The spacers are numbered from the latest (S1) to the oldest (S9) integrated sequences. The leader sequence, which controls CRISPR transcription, is proximal to spacer S1.

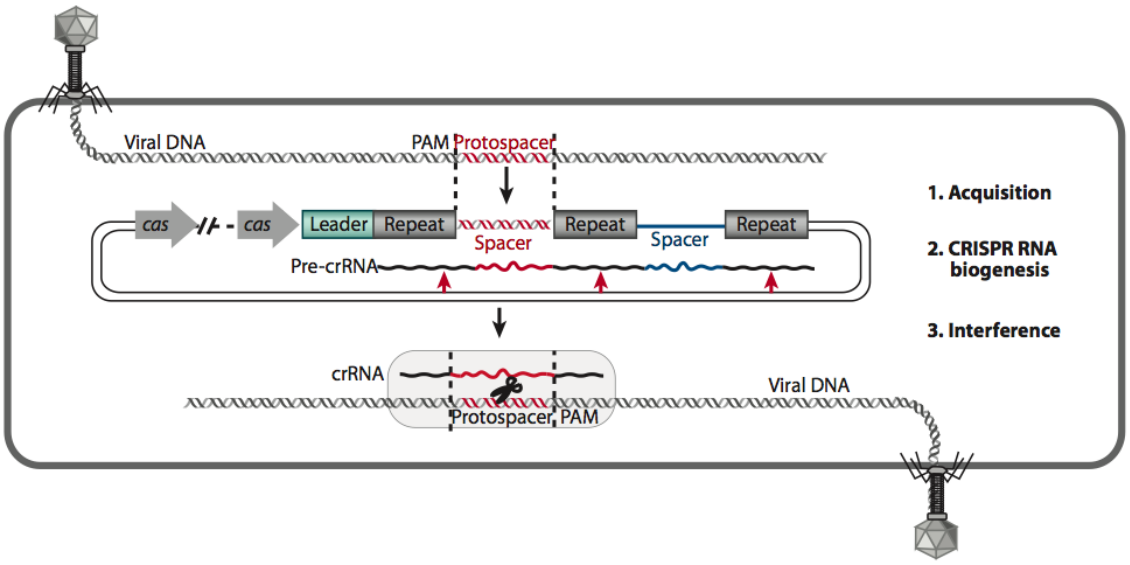


Figure 1.2. The RNA-based CRISPR immune system progresses in three distinct stages (Sorek et al., 2013). 1) Spacers complementary to protospacer sequences are incorporated into the host chromosome during acquisition. 2) Upon subsequent attack by the same invader, the CRISPR locus is transcribed, processed by Cas proteins, and packaged into ribonucleoprotein complexes during CRISPR RNA biogenesis. 3) Invader DNA is identified based on crRNA-target sequence complementarity and is eventually silenced during interference.

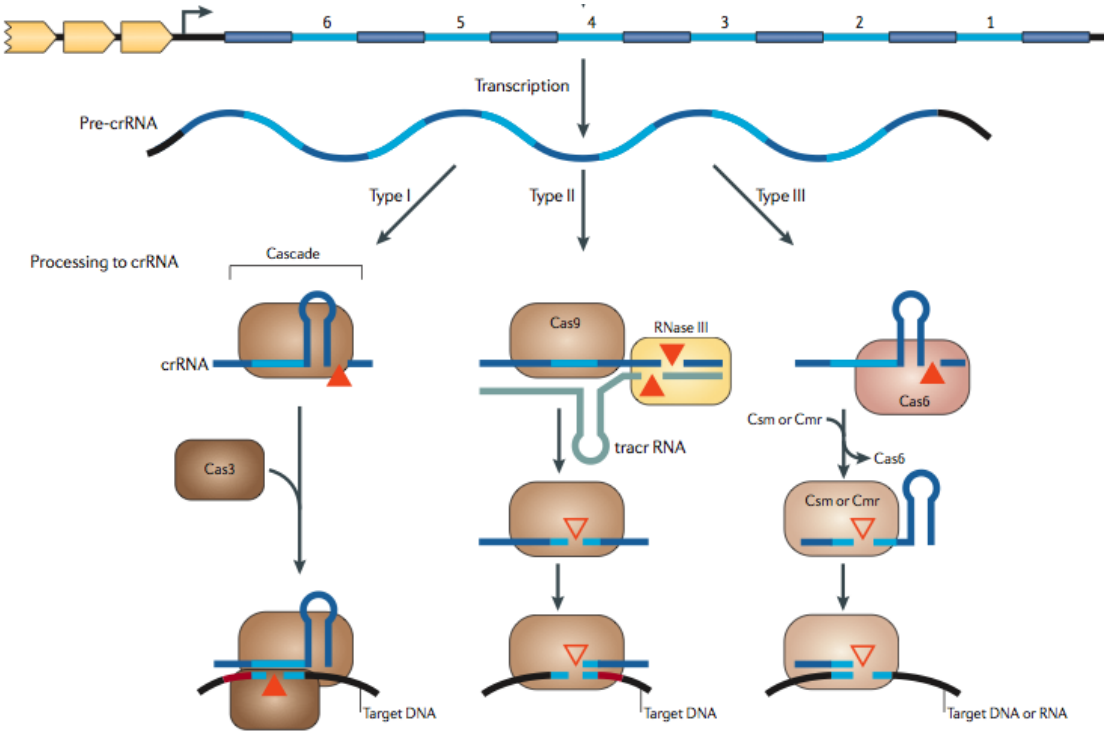


Figure 1.3. The three types of CRISPR systems (Makarova et al., 2011) based on conservation of *cas* genes. The characteristic proteins of the type I, II, and II CRISPR system are Cas3, Cas9, and Cas10 respectively.

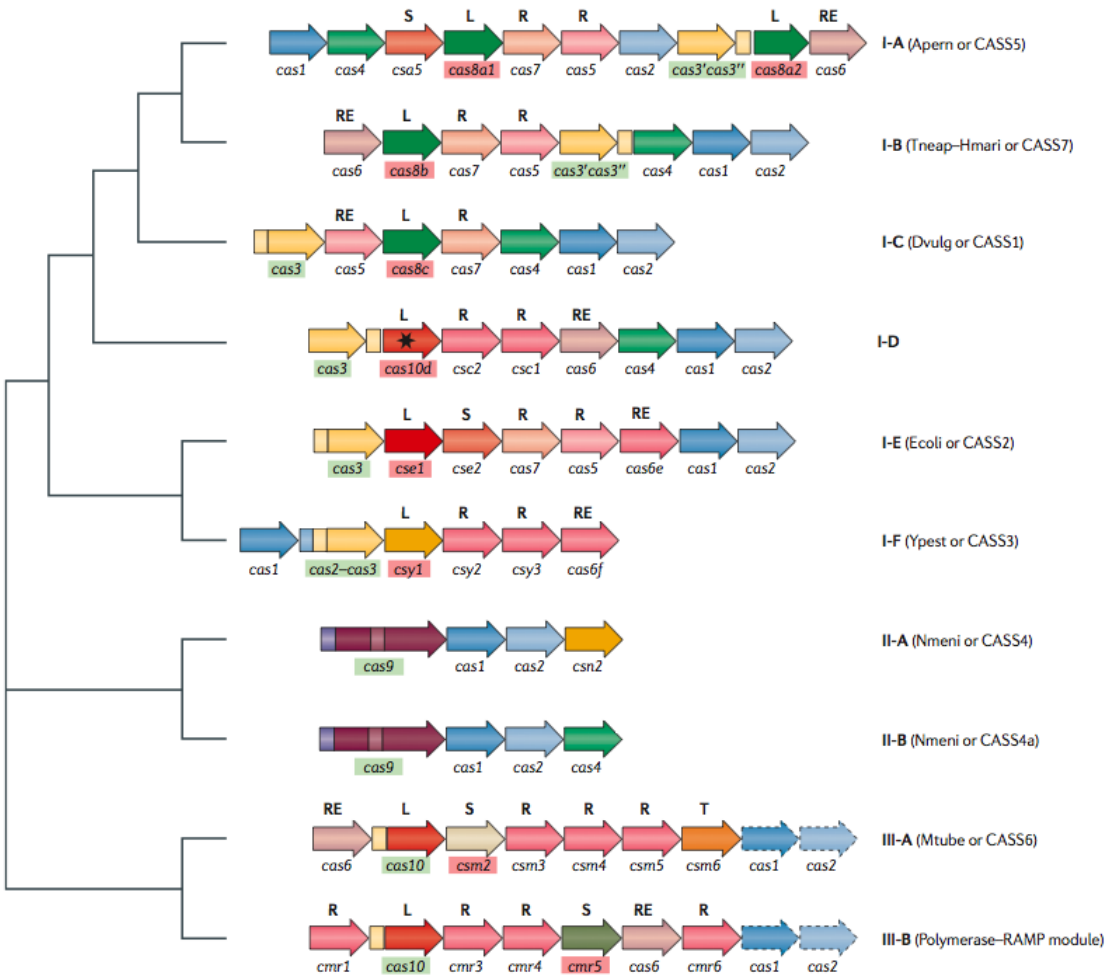


Figure 1.4: Diversity of CRISPR-associated proteins (Makarova et al., 2011). The type I, type II, and type III CRISPR systems are further divided into different subtypes.

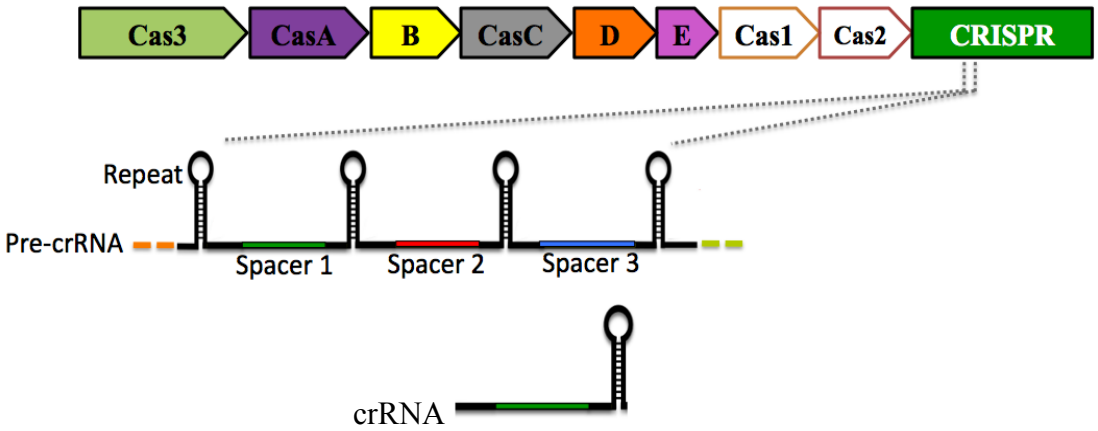


Figure 1.5: Diagram of the type I-E CRISPR/Cas system in *E. coli*. The CRISPR locus is transcribed into a long pre-crRNA and is further processed by five Cas proteins (A-E) into shorter crRNA. Each crRNA is 61 nucleotides in length and consists of an 8-nucleotide 5'-handle, 32-nucleotide spacer, and a 21 nucleotide repeat element.

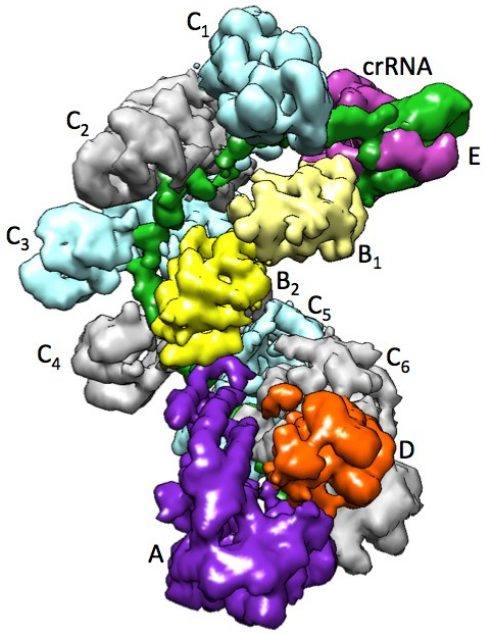


Figure 1.6: Cryo-electron microscopy structure of the Cascade complex from *E. coli* showing the overall organization of the protein subunits. The CasA (purple), CasB (yellow), CasC (cyan and grey), CasD (orange), casE (magenta) subunits interact with different parts of the crRNA (green).

Chapter 2

Crystal structure of the largest subunit of the Cascade complex and its role in target DNA binding

Introduction

Cryo-electron microscopy (cryo-EM) structures of the 405 kDa Cascade complex to ~9 Å from *E. coli* had just been reported (Wiedenheft et al., 2011). While the structures clearly outlaid the general organization of the individual subunits with respect to each other, the roles of the individual subunits were not obvious. One of the approaches that I took to investigate the mechanism of target DNA binding by Cascade was to try and crystallize its individual subunits on their own. Of the many subunits or smaller complexes of Cascade that we were able to purify, we succeeded in crystallizing the CasA subunit from *Thermus thermophilus* and in determining its crystal structure. Based on the *T. thermophilus* CasA crystal structure and the cryo-EM structures of *E. coli* Cascade, Amberly Orr conducted the binding studies of Cascade and different Cascade sub-complexes as part of her Master's thesis. The material presented in this chapter has been previously published and reprinted here from Mulepati, S., Orr, A., and Bailey, S. (2012) Crystal structure of the largest subunit of a bacterial RNA-guided immune complex and its role in DNA target binding. *J. Biol. Chem.* **287**, 22445-22449, with permission from American Society for Biochemistry and Molecular Biology.

Results

Crystal Structure of CasA

To gain a more detailed understanding of CasA (also known as CRISPR-subtype *E. coli 1* (Cse1) or YgcL), we determined its crystal structure. Initial attempts to crystallize the *E. coli* protein were unsuccessful. We therefore expressed and purified the homolog from *T. thermophilus* HB8 (TthCasA). We chose this organism because the sequence of TthCasA has 50% similarity with *E. coli* CasA, and crystal structures of both TthCasB (Agari et al., 2008) and TthCasE (Ebihira et al., 2006) have been determined. Crystals of TthCasA were obtained by vapor diffusion using a precipitant solution containing sodium acetate. The crystals belonged to the space group $P2_1$ (a = 93.9 Å, b = 47.9 Å, c = 129.2 Å, and $\alpha = \gamma = 90°$, $\beta = 97.52°$) and contain two monomers in the asymmetric unit. The structure was determined by single isomorphous replacement, utilizing platinum-soaked crystals, and the structure was refined to 2.4 Å resolution with an R_{work} of 19.4% and an R_{free} of 25.3%. Additional data collection, phasing, and refinement statistics are given in Table 2.2. The final model displayed good geometry and contained all of the TthCasA sequence with the exception of the first 4 N-terminal and last 6 C-terminal residues, as well as two internal loops formed by residues 129–142 (N-loop) and residues 405–409 (C-loop). The structures of the two monomers of TthCasA in the asymmetric unit are virtually identical with a root mean square deviation of 0.1 Å over 473 Cα atoms.

Overall the structure of TthCasA can be divided into two domains corresponding to the N- and C-terminal parts of the polypeptide chain (Fig. 2.1B). The two domains are arranged in a chair-like conformation, with the N-domain forming the seat and the C-

domain forming the backrest (Fig. 2.1B). The larger N-domain includes residues 1–364 and is composed of 11 β-strands and 9 α-helices. A search of the structural database using the DALI server found no significant matches, suggesting that the N-domain has a novel fold. The smaller C-domain includes residues 365–502 and is formed by five α-helices. Four of these α-helices form an up-down-up-down four-helix bundle. The fifth, smaller α-helix is located in a flexible loop (the C-loop) separating the first and second helix of the bundle.

Docking the Crystal Structure of TthCasA into the Cryo-EM maps of Cascade

To gain further insight into the role of CasA, we rigid-body fit the crystal structure of TthCasA into the cryo-EM maps of Cascade with and without bound protospacer target (Wiedenheft et al., 2011). The crystal structure aligned well into both maps, and α-helices in the crystal structure aligned with the corresponding rods of density in the cryo-EM maps (Fig. 2.1, C and D). The quality of the fit into both cryo-EM maps suggests that there is no significant change in the relative orientation of the two domains of CasA, observed in the crystal structure, upon binding to Cascade.

In the cryo-EM map of Cascade with no bound target, the CasA N-domain sits adjacent to CasD, and contiguous cryo-EM density suggests that the N-loop of CasA contacts with the 5'-end of crRNA (Fig. 2.1C). The C-domain of CasA contacts the fifth and sixth CasC subunits as well as the neighboring CasB subunit (Fig. 2.1C). Upon binding to protospacer target, the CasA, CasB, and CasE subunits undergo a concerted conformational change (Wiedenheft et al., 2011). In the cryo-EM map of Cascade bound to protospacer target, the movement of CasA is such that the N-loop appears to no longer

interact with the crRNA, whereas the C-loop now makes new contacts with the crRNA-protospacer duplex (Fig. 2.1D).

DNA Binding by Cascade

Binding of Cascade to nonself target relies on the recognition of a PAM (Semenova et al., 2011). Previous studies on the role of CasA in this binding were performed before the *E. coli* PAM was identified (Jore et al., 2011). We therefore examined the role of CasA in Cascade binding to an 85-bp dsDNA target containing protospacer and functional PAM sequences. A fixed concentration of dsDNA target was incubated with increasing concentrations of either Cascade or CasBCDE, and complex formation with dsDNA was analyzed by native gel electrophoresis (Fig. 2A). CasBCDE did not bind dsDNA target, whereas Cascade did. The amount of complex formed between dsDNA target and Cascade exhibited a sigmoidal dependence on Cascade concentration (Fig. 2B). There are at least two possible explanations for the sigmoidal binding curve, either (i) cooperative binding between multiple sites on Cascade or the dsDNA or (ii) the existence of two equilibriums, one between Cascade and dsDNA and the other between a single subunit of Cascade and the rest of the complex. Because the stoichiometry between Cascade and dsDNA target is thought to be 1:1 (Jore et al., 2011, Wiedenheft et al., 2011) and CasA was seen to dissociate from Cascade during competitive ssDNA binding experiments (Jore et al., 2011), the sigmoidal dependence is more likely the result of dissociation of CasA from Cascade at low concentrations. To confirm this hypothesis, we repeated the above binding experiments in the presence of saturating concentrations of CasA (250 nM). Under these conditions, the amount of

complex formed between Cascade and dsDNA target exhibited a hyperbolic dependence on Cascade concentration (Fig. 2.2B) with an apparent dissociation equilibrium constant (K_d) of 0.54 ± 0.1 nM. The addition of a saturating concentration of CasA to CasBCDE rescued binding of this complex to dsDNA target (Fig. 2.2A) and also displayed a hyperbolic dependence on CasBCDE concentration with a K_d indistinguishable from Cascade in the presence of saturating concentration of CasA (Fig. 2.2B). In control experiments, CasA alone was not able to bind dsDNA target (13) (Fig. 2.2A).

Discussion

The crystal structure of TthCasA reveals a two-domain protein with a novel N-terminal fold and a C-terminal four-helix bundle. A prominent feature of this structure is two disordered loops, one in the N-domain and another in the C-domain, termed the N-loop and C-loop, respectively. Docking the crystal structure of TthCasA into the cryo-EM maps of Cascade suggests that these loops become ordered when CasA binds Cascade and that they make significant contacts with the crRNA and the protospacer target. In the absence of target, the N-loop makes contact with the 5'-end of the crRNA, but upon Cascade binding to protospacer target, the N-loop disengages from the crRNA (Fig. 2.1D) (Wiedenheft et al., 2011). The C-loop makes little or no contacts with the crRNA in the absence of target but does make extensive contacts with the crRNA-protospacer duplex when Cascade is bound to protospacer target (Fig. 2.1D). Thus, both of these loops appear to make key contributions to the specific structural states that correlate with Cascade target binding.

The PAM plays a critical role in self versus nonself recognition (Semenova et al., 2011; Deveau et al., 2008; Gudbergsdottir et al., 2011; Marfaffini and Sontheimer, 2010;

Mojica et al., 2009). The PAM is found in nonself DNA targets but not in the host sequence, CRISPR loci. Recent DNA binding experiments have demonstrated that mutations in the PAM sequence decrease the affinity of Cascade for DNA target (Semenova et al., 2011), suggesting a direct interaction between Cascade and the PAM. The N-loop of CasA may mediate this critical interaction. If a longer nonself target, including the PAM sequence, were modeled onto Cascade, the projected path of the target would position the PAM adjacent to the site where the N-loop of the crystal structure of CasA docks into the EM map (Fig. 2.1D).

Our DNA binding experiments show that CasA is essential for specific binding of Cascade to nonself target (Fig. 2.2). Taken together with the observation that CasA dissociates from the complex at low concentrations, this suggests that CasA expression levels may provide an opportunity for regulation of the activity of Cascade within the cell. Cascade would not be able to bind dsDNA target at low expression levels of CasA, but at high expression levels, Cascade could bind DNA target and signal its destruction by Cas3. Confirmation of this model will require measurement of the cellular concentrations of the individual Cascade subunits.

In summary, we have shown here that the CasA subunit of Cascade is essential for nonself target binding. We present the crystal structure of CasA and its fit into cryo-EM maps of Cascade bound and unbound to protospacer target. This structural analysis reveals two loops in CasA that are likely key sensors for dsDNA target binding.

While this manuscript was in preparation, a similar analysis of CasA was published by Doudna and colleagues (Sashital et al., 2012). This manuscript independently presents similar results but also experimentally confirms the role of the N-

loop in both PAM binding and additionally in the control of nonspecific DNA binding by Cascade.

Methods

Cloning and Protein Expression

The cloning and expression strategy was similar to that described previously (Brouns et al., 2008; Jore et al., 2011). Thus, all genes were amplified from genomic DNA (American Type Culture Collection) and directionally cloned into a series of expression vectors (Table 2.1). An *E. coli* CRISPR array consisting of seven identical spacers (sequence: 5'-CCAGTGATAAGTGGAATGCCATGTGGGCTGTC-3') was synthesized by GeneArt. All proteins were overexpressed in the T7Express strain of *E. coli* (New England Biolabs). Cells were grown in LB medium, supplemented with the appropriate antibiotic(s) (Table 2.1), at 37 °C to an A_{600} of 0.3–0.5, and subsequently protein expression was induced with 0.2 mM isopropyl β-D-1thiogalactopyranoside overnight at 20 °C.

Purification of E. coli Proteins

E. coli CasA, Cascade, and the CasBCDE-crRNA subcomplex were all purified using the same protocol. Harvested cells were lysed in buffer L (20 mM Tris-HCl, pH 8.0, 100 mM NaCl and 10% glycerol), clarified, and then loaded onto a 5-ml immobilized metal affinity chromatography column (Bio-Rad). The column was then washed with 10 mM imidazole before the protein of interest was eluted with 250 mM imidazole. N-terminal tags were removed by treatment with tobacco etch virus (TEV) protease

overnight at 4 °C. Samples were then desalted to remove imidazole and then reapplied to immobilized metal affinity chromatography resin to remove the His-tagged TEV protease, any cleaved tag, or any remaining tagged protein. Samples were then concentrated and loaded on a HiLoad 26/60 S200 size-exclusion column (GE Healthcare) pre-equilibrated with Buffer A (20 mM Tris-HCl, pH 8.0, 200 mM NaCl, and 1 mM tris(2-carboxyethyl)phosphine). As seen previously, all proteins eluted as symmetrical peaks at their expected molecular weights (Jore et al., 2011).

Purification of Thermus thermophilus CasA

Harvested cells were lysed in buffer L, and the clarified lysate was heat-treated at 70 °C for 10 min. Following centrifugation, the sample was adjusted to 1.5 M ammonium sulfate and loaded onto a 5-ml Fast Flow Phe column (GE Healthcare) pre-equilibrated with 40 mM Tris-HCl, pH 7.5, 1.5 M ammonium sulfate, 10% glycerol. Protein was eluted with a linear gradient of 1.5– 0 M ammonium sulfate. The relevant fractions were pooled, and the protein was further purified over a 5-ml Fast Flow Q column (GE Health-care) before finally being loaded on a HiLoad 26/60 S200 column (GE Healthcare) pre-equilibrated with Buffer A. The final purified protein was concentrated to ~30 mg/ml using Ultracel 10K centrifugal filter unit (Millipore).

Crystallization of T. thermophilus CasA

Crystals of *T. thermophilus* CasA were obtained by the sitting-drop vapor diffusion method, mixing 1 μl of CasA at ~30 mg/ml with 1 μl of precipitant solution of 0.1 M MOPS, pH 7.4, and 2.3 M sodium acetate. For stabilization and cryoprotection,

crystals were transferred to a solution of 0.1 M MOPS, pH 7.4, and 3.5 M sodium acetate. Crystals were flash-frozen in liquid nitrogen. Platinum derivatives were obtained by soaking crystals in a solution containing 0.1 M MOPS, pH 7.4, 3.5 M sodium acetate, and 20 mM K_2PtCl_4 for 3 h.

Structure Determination

X-ray diffraction data were collected at either beamline 9.2 at the Stanford Synchrotron Radiation Light Source (SSRL) or beamline X25 at the National Synchrotron Light Source (NSLS). Data were processed with HKL2000 (Otwinowski and Minor, 1997). SHELX (Sheldrick, 2008) was used to find the positions of the platinum sites. Phases were calculated using SOLVE (Terwilliger, 2004) and improved by solvent flattening and noncrystallographic symmetry averaging in RESOLVE (Terwilliger, 2004). Iterative model building and refinement were carried out in COOT (Emsley and Cowtan, 2004) and PHENIX (Adams et al., 2010).

Cryo-electron Microscopy Map Fitting and Preparation of Figures

Rigid-body docking of the *T. thermophilus* CasA crystal structure into the cryo-electron microscopy density of *E. coli* Cascade was performed with Chimera (Goddard et al., 2007). All structure panels were generated using PyMOL (Delano, 2010) or Chimera (Goddard et al., 2007).

DNA Binding Experiments

Binding assays contained 20 mM Tris-HCl, pH 8.0, 100 mM NaCl, and 10% glycerol. All oligonucleotides were gel-purified. dsDNA was made by annealing oligonucleotide A (5'-TCAATCTACAAAATTGAGCAAATCAGACAGCCCACATG-GCATTCCACTTATCACTGGCATTGCTTTCGAGCTTGCCGATCAGCTT-3") with oligonucleotide B (5'-AAGCTGATCGGCAAGCTCGAAAGCAATGCCAGTGATAA-GTGGAATGCCATGTGGGCTGTCTGATTTGCTCAATTTTGTAGATTGA-3'). Trace amounts (5–200 pM) of 5'-end ^{32}P-labled dsDNA were incubated with an increasing concentration of Cascade or CasBCDE for 1 h at 37 °C, prior to electrophoresis through a 5% polyacrylamide gel. In experiments with saturating CasA, 250 nM was confirmed to be saturating as repeating these experiments with 1 μM CasA (data not shown) gave the same results. DNA was visualized by phosphorimaging and quantified using Image Gauge (Fuji). As described before (Semenova et al., 2011), fraction of DNA bound was plotted versus protein concentration and fit to a one-site binding isotherm, using the GraphPad Prism software. Reported K_d values are the average of three replicates.

Acknowledgments

We thank Annie Héroux for help with data collection at the NSLS and Jennifer M. Kavran for critical reading of the manuscript.

Table 2.1. Plasmids used in the CasA studies

Vector		Relevant Features
pRSFDuet-1b (Kan^R)[a]		Two multiple cloning sites (mcs1 and mcs2)
pBAT4 (Amp^R)[c]		No tags
pHAT4 (Amp^R)[c]		Encodes for an N-terminal His-tag, followed by a TEV protease site
pMAT11 (Amp^R)[c]		Encodes for an N-terminal His-MBP (maltose-binding protein) tag, followed by a TEV protease site
Clone	**Vector**	**Genes**
pCRISPRd	pRSFDuet-1b	mcs1: CRISPR (7x spacer) mcs2: empty
pCRISPR-A[d]	pRSFDuet-1b	mcs1: CRISPR (7x spacer) mcs2: CasA
pBCDE[d]	pHAT4	*E. coli casB-casC-casD-casE*
pTthCasA	pBAT4	*T. thermophilus* CasA
pEcoCasA	pMAT11	*E. coli* CasA

[a] From Novagen.

[b] mcs1 can encode for an N-terminal His tag, however all cloning here removes this tag.

[c] From Peränen et al., 1996.

[d] Cells used to express Cascade were made by transformation with pCRISPR-A and pBCDE. Cells used to express CasBCDE+crRNA were made by transformation with pCRISPR and pBCDE.

Table 2.2. Data collection, processing, and phasing statistics

Data collection	**Native**	**Pt soak**
Resolution (Å)[a]	2.37 (2.48-2.37)	2.79 (2.89-2.79)
R_{sym}[a, b]	10.1 (63.4)	6.6 (31.7)
I/σ[a]	11.2 (2.2)	11.1 (2.4)
Redundancy[a]	3.7(3.6)	2.0 (1.9)
Completeness (%)[a]	98.5 (97.6)	98.7 (96.7)
Wilson B factor	50.70	
Mean figure of merit	0.32	
Heavy atom sites		6
Refinement		
Resolution (Å)	46-2.37	
R_{work}[c]	19.4	
R_{free}[c]	25.3	
r.m.s.d. bond (Å)[d]	0.01	
r.m.s.d. angle	1.2	
No. of atoms	7828	
B-factors	63.5	
Ramachandran plot		
Most favored (%)	90.3	
Additional allowed (%)	9.7	

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

[b] R_{sym} is $\Sigma|I_o - I|/\Sigma I_o$, where I_o is the intensity of an individual reflection, and I is the mean intensity for multiple recorded reflections.

[c] R_{work} is $||F_o - F_c||/F_o$, where F_o is an observed amplitude, and F_c is the calculated amplitude; R_{free} is the same statistic calculated over a subset of the data that has not been used for refinement.

[d] r.m.s.d., root mean square deviation.

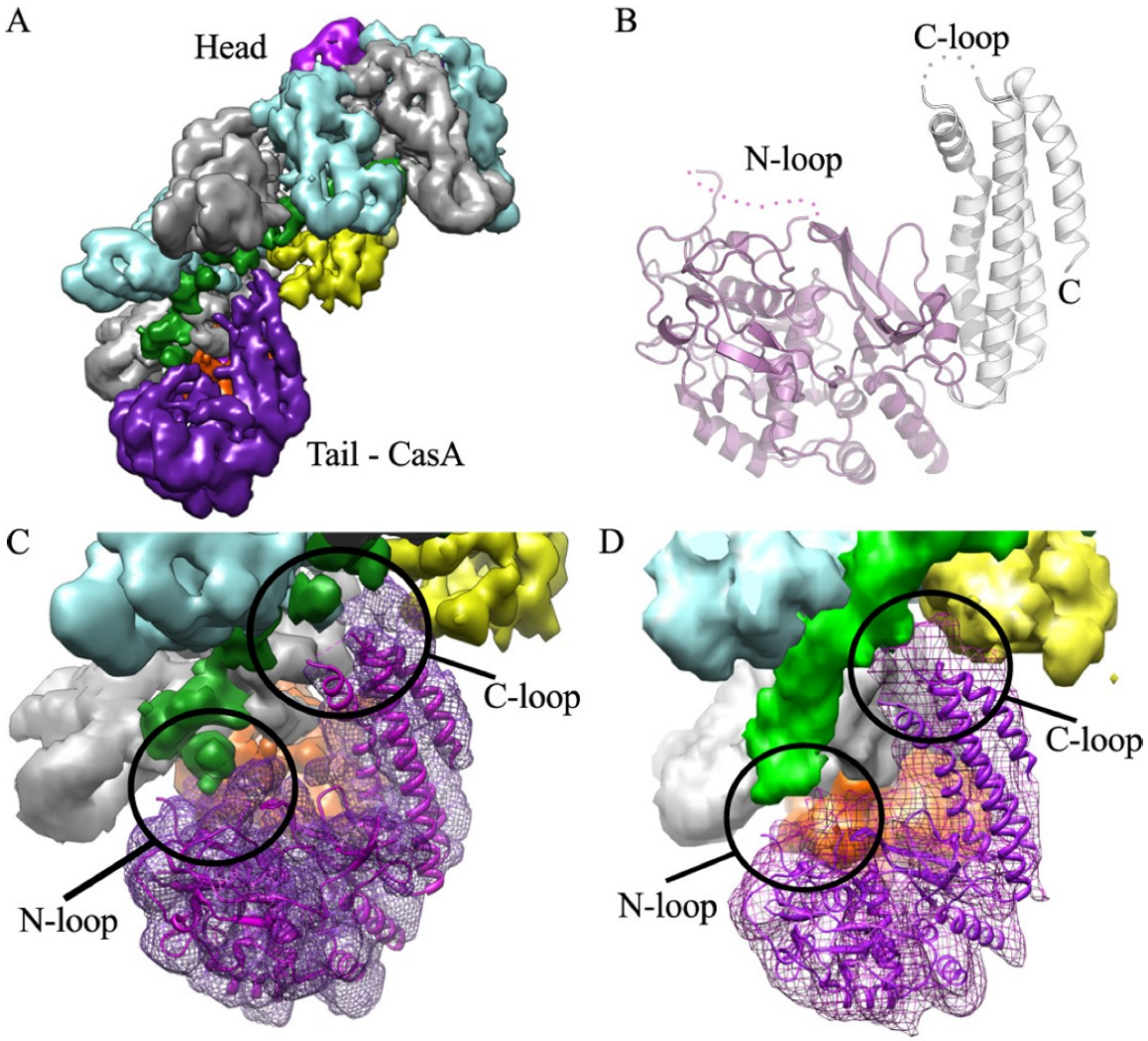


Figure 2.1. Structure of TthCasA. *A*, cryo-EM reconstruction of Cascade. Coloring is as in Wiedenheft et al., 2011 as follows: *magenta*, CasA; *yellow*, CasB; *cyan* and *gray*, CasC; *orange*, CasD; and *pink*, CasE. CasA is located at the tail of the structure. *B*, ribbon representation of the crystal structure of CasA. The N-domain is colored magenta, and the C-domain is colored white. The N- and C-loops are labeled. *C*, fit of the crystal structure of CasA into the cryo-EM map of Cascade. *D*, fit of the crystal structure of CasA into the cryo-EM map of Cascade bound to protospacer target. *C* and *D* are colored as in *A*.

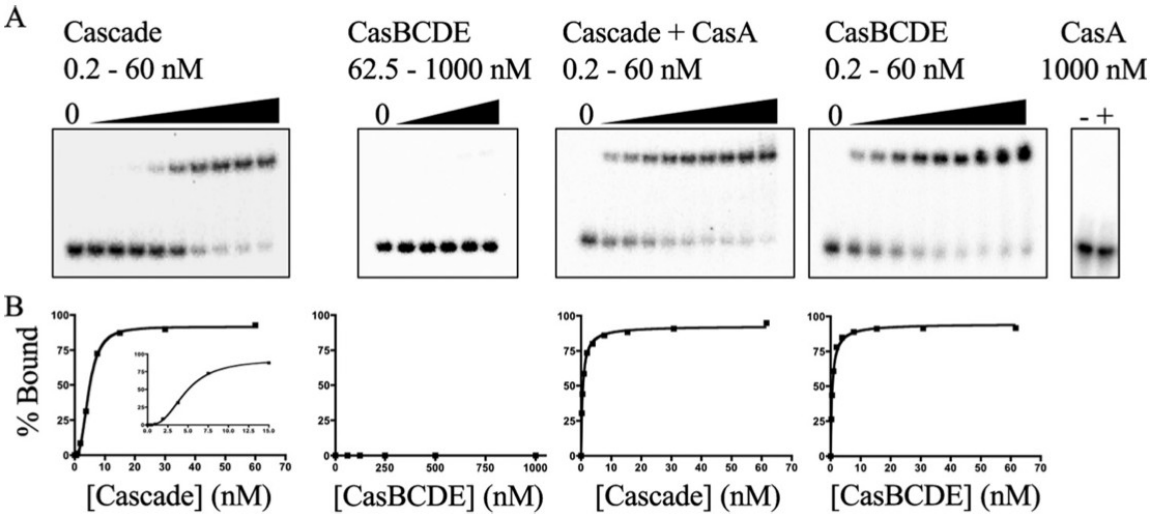


Figure 2.2. Double-stranded DNA target binding by Cascade. *A*, representative gel shift assays for Cascade and CasBCDE in the absence and presence of saturating concentrations (250 nM) of CasA. Also shown is a gel shift assay of CasA alone (1000 nM). The concentrations of the titrated species are given above each gel. *B*, binding curves measured from each of the assays in *A*. For the Cascade titration in the absence of CasA, the inset is a zoomed-in view of the binding curve, highlighting it sigmoidal character.

Chapter 3

Structural and biochemical analysis of the HD nuclease domain of Cas3 protein from *Thermus thermophilus*

Introduction

Cas3 is the signature protein of the type-I CRISPR system and comprises of an N-terminal Histidine-Aspartate (HD) domain and a C-terminal DEXH helicase domain. Previously, *in vivo* experiments had shown that the HD domain is indispensable for a proper CRISPR response (Brouns et al., 2008; Cady et al., 2011). The HD domain was predicted to be a nuclease and was characterized to be a double-stranded DNAase in *Sulfolobus sulfotaricus* (Han and Kraus, 2009) but a single-stranded DNase in *Streptococcus thermophilus* (Sinkunas et al., 2011). We carried out structural and biochemical analysis of the HD domain from *T. thermophilus* to characterize its structure and its nuclease activity. The material presented in this chapter is reprinted (with minor changes) from Mulepati, S., and Bailey, S. (2011) Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J. Biol. Chem.* **286**, 31896-31903, with permission from the American Society for Biochemistry and Molecular Biology.

Results

Crystal Structure of Cas3^{HDdom}

Sequence analysis predicts that the Cas3^{HDdom} spans residues 5–260 of *T. thermophilus* HB8 cas3. The DNA sequence encoding this region was cloned into the pHAT2 expression vector (Peränen et al., 1996), expressed in *E. coli*, and purified to

homogeneity by affinity and size-exclusion chromatography. $Cas3^{HDdom}$ was crystallized by vapor diffusion with PEG 300 as the precipitant. Divalent metal ions inhibited crystallization and therefore were not included in crystallization experiments. Crystals belong to the space group $P4_322$ (a = b = 48.5 Å and c = 205.9 Å) and contain one $Cas3^{HDdom}$ molecule per asymmetric unit. The structure was determined by multiple isomorphous replacement and refined at 1.8 Å resolution to an R_{work} of 16.8% and an R_{free} of 19.5%. A representative section of unbiased electron density is shown in Fig. 3.1A. The final model displays good geometry with no Ramachandran outliers (Table 3.1) and contains all of the $Cas3^{HDdom}$ sequence except residues 81–101 and 183–188, for which electron density was not interpretable. No divalent metal ions were apparent in the electron density map. The tertiary structure of $Cas3^{HDdom}$, composed of 10 α-helices and two β-strands, is illustrated in Fig. 3.1B. Overall, $Cas3^{HDdom}$ adopts a globular structure with a concave surface formed by the five conserved motifs of the HD superfamily (Fig. 3.1C) (Aravind and Koonin, 1998).

To investigate the molecular basis of divalent metal ion binding by $Cas3^{HDdom}$, we measured diffraction data from crystals soaked in a stabilization solution containing 100 μM nickel sulfate. Soaks were performed at the crystallization pH of 4.2 because attempts to increase the pH severely reduced the quality of the x-ray diffraction data. Anomalous difference electron density maps, calculated from soaked data, contained a single strong peak (~25σ) positioned at the canonical metal-ion-binding site. The Ni^{2+} ion is coordinated, with octahedral geometry, by two water molecules and four conserved residues, within motifs I (His-24), II (His-69 and Asp-70), and V (Asp-205) (Fig. 3.1D). Superposition of the structures of $Cas3^{HDdom}$ with and without bound metal ion results in

a root mean square deviation of 0.28 Å over 235 Cα atoms, demonstrating that the structure of $Cas3^{HDdom}$ remains largely unchanged upon metal ion binding. The only significant difference is seen at the metal ion-binding site where the side chain of His-69 rotates to coordinate the metal ion (Fig. 3.1D).

Comparison with Other HD Domain Proteins

As expected, a search of the structural database by DALI (Holm and Sander, 1993) shows that $Cas3^{HDdom}$ is related to other HD domains, including many unpublished structures deposited by structural genomics initiatives. The most closely related is the unpublished structure of a Cas3" protein (MJ0384) from *Methanocaldococcus jannaschii* (PDB code 3M5F). These two proteins share only 14% amino acid identity, yet their structures align with a root mean square deviation of 3.2 Å over 142 Cα atoms (Z score of 7.9) (Fig. 3.2A). Although overall similar, there are some notable differences between the structures. First, two regions differ in topology. A 13-amino acid loop (residues 176–191) connects the seventh and eighth helices of $Cas3^{HDdom}$, whereas in Cas3", an insertion results in an additional α-helix connecting these helices (residues 168 –202) via an alternate path (Fig. 3.2B). The C terminus of $Cas3^{HDdom}$ also ends with a helix-strand-strand arrangement that is absent in the Cas3" structure (Fig. 3.2A). However, this difference may be the result of disorder in the electron density map of Cas3". The last 30 residues of Cas3" are not modeled, and the electron density describing the last helix (which contains motif V) lacks clear side chain features. The coordinates of many of the residues in this helix have been truncated in the model. The second difference between $Cas3^{HDdom}$ and Cas3" is in the configuration of the metal ion binding sites. Two Ca^{2+} ions

are modeled at the active site of *M. jannaschii* Cas3" (Fig. 3.3A). Strikingly, neither of these metal ions is found at the canonical metal binding site of HD domains (site A in Fig. 3.3), perhaps due to the disorder observed in the electron density in the region of motif V. Instead, four conserved residues within motifs II, III, and IV coordinate one of the Ca^{2+} ions (site B in Fig. 3.3A). The significance of the position of the second Ca^{2+} ion (site C in Fig. 3.3A) is unclear as it is located 4.0 Å from the nearest protein atom or water molecule.

The Cas3" residues interacting with the site B Ca^{2+} ion are conserved in $Cas3^{HDdom}$ (Fig. 3.3B), suggesting that $Cas3^{HDdom}$ could bind two metal ions. Yet, in the electron density maps generated from our Ni^{2+} -soaking experiments, we fail to observe a metal ion at site B. The motif IV histidine residues (His-137 and His-138), which coordinate the site B metal ion in Cas3", are oriented away from the binding site in the $Cas3^{HDdom}$ structure (Fig. 3.3B). However, these residues are in a loop that is constrained by crystal packing contacts and, as a result, may be unavailable for metal ion binding. We were unable to crystallize $Cas3^{HDdom}$ in the presence of divalent metal ions. In solution, $Cas3^{HDdom}$ may bind two metal ions, one at site A and the other at site B. At least five structures of HD domains have been determined (PDB codes 2PQ7, 3HC1, 3CCG, 2OGI, and 2O08) that have a metal ion bound at each of these sites (an example of which is shown in Fig. 3.3C). These five HD domains are of unknown function, but they all contain a conserved histidine residue in motif III and either one or two conserved histidine residues in motif IV.

The availability of the crystal structures of many HD superfamily members permits us to define the minimal fold of the HD domain. An inspection of the overlay of

HD domain structures reveals a common core of five α-helices that, along with their connecting loops, house the five motifs that define the HD superfamily (Fig. 3.2C) (Aravind and Koonin, 1998). Beyond this core structure, different members of the HD superfamily have unique structural elements that presumably help specify the individual functions of each HD domain family.

Nuclease Activity of Cas3^{HDdom}

S. thermophilus Cas3 has been shown to cleave ssDNA but not dsDNA in a Mg^{2+}-dependent manner. To investigate the activity of the *T. thermophilus* Cas3^{HDdom}, we incubated various concentrations of the protein in the presence of Mg^{2+} and ssDNA (M13mp18). The reactions were then analyzed by electrophoresis through agarose gels, and the DNA species was visualized with ethidium bromide (Fig. 3.4A). Under these conditions, no cleavage of ssDNA was detected. We also assayed cleavage of dsDNA (PvuII-linearized pUC19) and again observed no cleavage (Fig. 3.4B). As HD domains have been reported to utilize a variety of different divalent cations as cofactors (Proudfoot et al., 2004; Seto et al., 1988; An et al., 1979; Lo et al., 2004), we evaluated the ability of several other divalent cations (Ca^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, and Zn^{2+}) to activate the nuclease activity of Cas3^{HDdom}. Thus, cleavage reactions were repeated, substituting increasing concentrations of each of these metal ions for Mg^{2+}. In these experiments, reaction products were detected as a smear on the agarose gel in the presence of low concentrations (20 μM) of Ni^{2+}, Mn^{2+}, Co^{2+}, Cu^{2+}, and Zn^{2+} (Fig. 3.5A). No cleavage was observed in either the absence of divalent metal ions, the presence of EDTA, or the presence of up to 20 mM Ca^{2+}. We also tested and detected no cleavage of

dsDNA in the presence of any of these metal ions (Fig. 3.5B).

Metal Ion Binding Increases Thermal Stability of Cas3^{HDdom}

To further characterize the interaction between Cas3HDdom and divalent cations, we performed a Thermofluor assay (Vedadi et al., 2006; Lo et al., 2004; Pantoliano et al., 2001). This assay measures the change in the fluorescence signal of SYPRO orange dye as it interacts with a protein undergoing thermal unfolding. The fluorescence signal of the dye is quenched in an aqueous environment but becomes unquenched when exposed to the hydrophobic core of the protein upon unfolding. The midpoint of the unfolding transition is taken as an approximation of the melting temperature (T_m). This assay can assess ligand binding because ligands that bind more tightly to the folded form of the protein than to the unfolded form are likely to increase the apparent T_m of that protein (Matulis et al., 2005). In the presence of EDTA, the apparent T_m of Cas3^{HDdom} is 49.6 ± 0.2 °C (Fig. 3.6). In line with our activity data, the addition of 20 mM Mg^{2+} and 100 μM Ni^{2+} increases the apparent T_m of Cas3^{HDdom} by ~6 °C and ~15 °C, respectively (Fig. 3.6).

Mutational Analysis of Cas3^{HDdom}

A series of point mutant proteins were generated in which putative active site residues (His-24, His-69, Asp-70, Lys-73, His-105, His-138, His-139 Ser- 202, Ser-209, and Asp-205) or the surface-exposed aromatic residues that surround this site (Trp-102 and Phe-253) were replaced with alanine. The location of these mutations in Cas3^{HDdom} is highlighted in Fig. 3.7A. Alanine mutants were expressed and purified in the same manner as the wild-type protein (Fig. 3.7B). To ensure that any potential defects observed

in nuclease activity could not be attributed to global misfolding and to assess divalent metal ion binding, the apparent T_m of each alanine mutant was determined in the absence or presence of Ni^{2+} (Table 3.2). In the absence of Ni^{2+}, nine of the 12 mutant proteins have T_m values similar to or greater than that of wild-type protein, indicating that these mutations did not destabilize the fold of $Cas3^{HDdom}$. Three of the mutant proteins (K73A, H105A, and S209A) have a T_m lower than wild-type, suggesting that some destabilization did occur. For wild-type $Cas3^{HDdom}$, the addition of Ni^{2+} results in an increase in apparent T_m (ΔT_m) of 15°C. The mutation of residues not implicated in metal ion binding (Lys-73, Trp-102, Ser-202, Ser-209, and Phe-253) results in similar or larger ΔT_m values, supporting the evidence that these residues do not participate in metal ion binding. With the exception of His-105, the mutation of residues predicted to bind metal ions (His-24, His-69, Asp-70, His-138, His- 139, and Asp-210) results in a decreased ΔT_m (Table 3.2), suggesting that, in solution, these residues are involved in metal ion binding.

We next tested each of the alanine mutants for ssDNA endonuclease activity using the M13mp18 phage cleavage assay in the presence of Ni^{2+} (Fig. 3.7C). Under these conditions, wild- type $Cas3^{HDdom}$ generated cleavage products that migrated as a tight smear on an agarose gel, whereas the mutation of residues predicted to bind metal ions abolished this nuclease activity. These results confirm that the activity we observed is not the result of a contaminating protein. The other alanine mutations abolished (K73A), suppressed (W102A and S209A), or had little or no affect (S202A and F252A) on the nuclease activity of $Cas3^{HDdom}$.

Discussion

Cas3 is functionally essential (Brouns et al., 2008; Cady and Toole, 2011) and is the signature gene of the type I CRISPR/Cas system (Makarova et al., 2011). The HD nuclease domain of Cas3 is proposed to cleave the ssDNA revealed upon cascade binding to target DNA (Sinkunas et al., 2011). Consistent with this hypothesis, we show that *T. thermophilus* $Cas3^{HDdom}$ cleaves ssDNA but not dsDNA. This result also establishes that the helicase domain of Cas3 does not alter the substrate specificity of its HD domain. Mg^{2+} activated the endonuclease activities of *S. thermophilus* Cas3 (Sinkunas et al., 2011) and *S. sulfataracus* Cas3" (Han and Krauss, 2009). In contrast, the transition metal ions Mn^{2+}, Co^{2+}, Ni^{2+}, and Zn^{2+} activate the endonuclease activity of *T. thermophilus* $Cas3^{HDdom}$ but not Mg^{2+} or Ca^{2+} (Fig. 3.7C). It is also noteworthy that *T. thermophilus* $Cas3^{HDdom}$ appears much more active in the presence of transition metal ions, particularly Ni^{2+}, than *S. thermophilus* Cas3 is in the presence of Mg^{2+} (Sinkunas et al., 2011). More quantitative data will be needed to establish whether this is significant. Which metal ion, or ions, is used *in vivo* by *T. thermophilus* Cas3 remains to be determined. We cannot rule out the possibility that the *in vitro* requirement for transition metal ions could be a sign of a missing cofactor found in the cell. However, in *T. thermophilus*, the total intracellular concentration of Mn^{2+}, Ni^{2+}, and Zn^{2+} are ~160, 150, and 550 μM, respectively, whereas the concentration of Co^{2+} is undetectable (Kondo et al., 2008). Considering that only 20 μM of Mn^{2+} or Ni^{2+} is needed to activate $Cas3^{HDdom}$ *in vitro*, these two ions are the most likely *in vivo* candidates of the ions studied.

The crystal structure of *T. thermophilus* $Cas3^{HDdom}$ provides the first view of an HD domain with nuclease activity. $Cas3^{HDdom}$ adopts a globular structure with a concave

surface that contains the active site and presumably binds substrate DNA. Comparison of the structure of the Cas3HDdom with bound Ni^{2+} and other HD domains with bound divalent metal ions suggests that $Cas3^{HDdom}$ binds two metal ions at its active site. In line with this, the mutation of residues predicted to form the metal ion binding sites results in proteins with smaller ΔT_m values upon addition of Ni^{2+} compared with the wild-type protein. The D70A mutant has the smallest ΔT_m value, consistent with the observation that among two metal-ion dependent enzymes, the most critical residue for metal binding is often an aspartate (Yang et al., 2006). His-105 is the only residue predicted to bind metal ion that, when mutated, has a ΔT_m value comparable with that of wild-type protein (Table 3.2). However, this residue most likely is a metal ion ligand as it is highly conserved and coordinates metal binding in structures of other HD domains (PDB codes 2PQ7, 3HC1, 3CCG, 2OGI, and 2O08).

The metal-binding data and analysis presented here also imply that all HD domain proteins with histidine residues in motifs III and IV will have two metal ions at their active sites. Thus, unlike the HD domains characterized to date (Hogg et al. 2004; Kondo et al., 2007; Zimmerman et al., 2008), proteins in this subset of HD domains, which include Cas3 and Cas3", are likely to utilize a two metal-ion mechanism for catalysis (Freemong et al., 1988; Beese and Steitz, 1991). The fact that $Cas3^{HDdom}$ appears to bind two metal ions in the absence of substrate is also somewhat distinct, as it is generally observed that metal ion binding by two-metal-ion dependent enzymes requires the presence of cognate substrate (Yang et al., 2006).

We used structure-guided mutagenesis to confirm both the importance of metal ion-binding residues to the activity of $Cas3^{HDdom}$ and to examine the role of other residues

close to the metal ion-binding sites. The mutation of the residues predicted to bind metal ions, including His-105 (Fig. 3.3B), completely ablates the nuclease activity of $Cas3^{HDdom}$ under the conditions tested. These results, coupled with both our structural analysis and ΔT_m data (Table 3.2), are consistent with two metal ions bound at the $Cas3_{HDdom}$ active site and establish the importance of these ions for nuclease activity. Mutation of the invariant Lys-73 produced an inert enzyme (Fig. 3.7C). Because of the proximity of this residue to the metal ion binding sites (Fig. 7A) and its positive charge, it is likely that it helps correctly position a phosphate group of the substrate for catalysis. Mutation of Trp-102 or Ser-209 also resulted in a protein with a reduced activity (Fig. 3.7C). The position of these residues within the substrate-binding cleft (Fig. 3.7A) suggests that they may play a role in substrate recognition.

Studies of *S. solfataricus* Cas3" have shown that this enzyme has distinct substrate specificity compared with *S. thermophilus* and *T. thermophilus* Cas3, as it cleaves dsDNA but not ssDNA (Han and Krauss, 2009). Additionally, comparison of mutational studies between *T. thermophilus* cas3 (presented here) and *S. solfataricus* Cas3 (Han and Krauss, 2009) suggests that the active site geometry of *S. solfataricus* Cas3" may also be distinct. First, mutation of either His-69 or His-105 ablates $Cas3^{HDdom}$ nuclease activity (Fig. 3.7C). However, mutation of the corresponding residues in *S. solfataricus* Cas3" results in a protein with near wild-type activity (Han and Krauss, 2009). Secondly, mutation of Glu-92 in *S. solfataricus* Cas3", a motif III residue that is conserved in Cas3" but not Cas3, inactivated nuclease activity. Inspection of the structure of *M. jannaschii* Cas3" suggests that this glutamate may replace the second histidine residue of motif IV (His-143 in *T. thermophilus* and His-124 in *M. jannaschii*). This

histidine forms part of metal ion site B (Fig. 3.3) and is essential for the nuclease activity of *T. thermophilus* Cas3^{HDdom} (Fig. 3.7C). In the structure of *M. jannaschii* Cas3", Glu-92 is in close proximity to site B, but its side chain is oriented away from the metal ion (Fig. 3.3A). *S. solfataricus* Cas3" lacks the second histidine of motif IV. Thus, Glu-92 could substitute for this histidine in coordinating the metal ion, potentially explaining the importance of this residue for the nuclease activity of *S. solfataricus* Cas3". The significance of this mutational data and the difference in substrate specificity awaits further studies of the Cas3 and Cas3" proteins. However, these results may indicate that *S. solfataricus* Cas3" has a different functional role or mechanism of action within the CRISPR response.

Materials and methods

Cloning and Mutagenesis

The 780-bp sequence encoding Cas3^{HDdom} was amplified from the *Thermus thermophilus* HB8 cas3 gene (TTHB187) and cloned into the pHAT2 expression vector (nen et al., 1996) between its NcoI and EcoRI restriction sites. The resulting plasmid encodes the Cas3^{HDdom} polypeptide fused to an N-terminal His_6 tag. Alanine mutations were introduced into the *cas3^{HDdom}* gene by the QuikChange site-directed mutagenesis method (Stratagene). All mutations were verified by DNA sequencing.

Expression and Purification

Wild-type and mutant pHAT2-*cas3^{HDdom}* plasmids were transformed into the T7 EXPRESS strain of *E. coli* (New England Biolabs). The cells were grown at 37 °C in

Luria-Bertani medium to an A_{600} of 0.4. Expression was induced by the addition of 0.2 mM isopropyl 1-thio-β-D-galactopyranoside. Following overnight incubation at 20 °C, the cells were harvested by centrifugation. Cell pellets were lysed in lysis buffer (20 mM Tris-HCl, pH 8.0, 1 M NaCl, and 10% glycerol) and then clarified by centrifugation at 18,000 rpm for 30 min. The lysates were loaded on a 5-ml immobilized metal affinity column (Bio-Rad) charged with nickel sulfate. The column was washed with lysis buffer containing 20 mM imidazole, and the bound protein was then eluted with 250 mM imidazole. The elution was then loaded onto a HiLoad 26/60 S200 column (GE Healthcare) pre-equilibrated with gel-filtration buffer (20 mM Tris-HCl, pH 8.0, 1 mM EDTA, and 200 mM NaCl). Fractions containing $Cas3^{HDdom}$ were pooled, dialyzed against gel-filtration buffer lacking EDTA, and concentrated to 8 mg/ml using an Ultracel 10 K centrifugal filter unit (Millipore). Purified proteins were >95% pure as judged by SDS- PAGE and Coomassie staining.

Crystallization

Crystals of $Cas3^{HDdom}$ were obtained using the sitting-drop vapor-diffusion method by mixing 2 μl of $Cas3^{HDdom}$ at□8mg/ml with1 μl of solution containing 10 mM phosphate-citrate buffer at pH 4.2 and 10% PEG 300. Crystals were stabilized and cryo-protected in a solution containing 10 mM sodium acetate, pH 4.2, and 30% PEG 300, and then flash-frozen in liquid nitrogen. Platinum and osmium derivative crystals were obtained by soaking crystals for 90 min in a stabilization solution containing either 1 mM of K_2PtCl_4 or $(NH_4)_2OsCl_6$.

Data Collection and Structure Determination

X-ray diffraction data were collected at beamline 9.2 at the Stanford Synchrotron Radiation Light Source and processed with either XDS (Kabsch, 2010) or HKL2000 (Otwinowski and Minor, 1997). SHELX (Sheldrick, 2008) was used to find the position of osmium sites in the $(NH_4)_2OsCl_6$ derivative crystal first. The phases derived from these osmium sites were then used to calculate a difference Fourier map to find the heavy atom positions in the platinum derivative. All phases were calculated using SOLVE and improved by solvent-flattening in RESOLVE (Terwilliger, 2004). Model building was carried out in Coot (Emsley and Cowtan, 2004), and the model was refined with PHENIX (Afonine et al., 2010). Coordinates of the metal-free and metal-bound structures have been deposited with the Protein Data Bank (PDB) codes 3SK9 and 3SKD, respectively.

Analysis of Methanococcus jannaschii Cas3"

To access the structure of *M. jannaschii* Cas3", we calculated difference electron density maps (2.3 Å resolution) from the deposited coordinates and structure factors (PDB code 3M5F) using the phenix.model_vs_data script (Afonine et al., 2010). The Rwork/Rfree from this calculation is 22.9/26.3%, in good agreement with the published Rwork/Rfree of 23.1/26.0%.

Nuclease Assay

$Cas3^{HDdom}$ nuclease assays were performed as described previously (Sinkunas et al., 2011). Magnesium chloride (Mg^{2+}), manganese chloride (Mn^{2+}), nickel sulfate (Ni^{2+}), copper chloride (Cu^{2+}), cobalt chloride (Co^{2+}), calcium chloride (Ca^{2+}), and

zinc chloride (Zn^{2+}) were used in metal ion substitution reactions at the indicated concentrations. All reactions were terminated with 20 mM EDTA. The products of reactions were separated by electrophoresis through 1% agarose gels and visualized by ethidium bromide staining.

Thermofluor Assay

The apparent melting temperature values of wild-type and mutant $cas3^{HDdom}$ were determined as described previously (Vedadi et al., 2006). Experiments were performed in a buffer containing 20 mM Tris-HCl, pH 8.0, and 200 mM NaCl. The final concentration of protein was 10 μM ($\varepsilon = 55460\ M^{-1}\ cm^{-1}$). Reactions were heated from 20 to 80 °C, and the fluorescence intensity was recorded at 0.2 °C intervals. Fluorescence intensities were plotted as a function of temperature, and the midpoint of the unfolding transition taken as an estimation of the melting temperature.

Acknowledgements

We thank Gabriel Brandt and Jennifer Kavran for critical reading of the manuscript and Jürgen Bosch for use of a real-time PCR machine.

Table 3.1. Data collection and processing statistics

Data collection	**Native**	**Ni soak**	**Pt soak**	**Os soak**
Resolution (Å)[a]	1.8 (1.9-1.8)	2.00 (2.11-2.00)	2.23 (2.31-2.23)	2.29 (2.37-2.29)
R_{sym}[a, b]	6.1 (49.5)	15.4 (70.5)	14.9 (93.6)	6.5 (14.9)
I/σ[a]	22.7 (4.5)	9.5 (3.2)	8.3 (1.2)	18.1 (9.4)
Redundancy[a]	11.7 (10.7)	12.0 (11.5)	3.8 (3.8)	3.8 (3.8)
Completeness (%)[a]	99.4 (96.4)	99.9 (99.5)	100 (100)	99.6 (97.6)
Wilson B factor	25.56			
Mean figure of merit	0.51			
Heavy atom sites		1	4	1
Refinement				
R_{work}[c]	16.69	16.52		
R_{free}[c]	19.54	20.62		
r.m.s.d. bond (Å)[d]	0.016	0.018		
r.m.s.d. angle	1.425	1.518		
No. of atoms	1953	1957		
B-factors	32.7	37.3		

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

[b] R_{sym} is $\Sigma|I_o - I|/\Sigma I_o$, where I_o is the intensity of an individual reflection, and I is the mean intensity for multiple recorded reflections.

[c] R_{work} is $||F_o - F_c||/F_o$, where F_o is an observed amplitude, and F_c is the calculated amplitude; R_{free} is the same statistic calculated over a subset of the data that has not been used for refinement.

[d] r.m.s.d., root mean square deviation.

Table 3.2. Melting temperatures of wild-type and mutant Cas3^{HDdom}

	Melt temperature (°C)	
	0 μM Ni2+	**100 μM** Ni^{2+}
WT	49.6 ± 0.2	64.8 ± 0.6
H24A	43.6 ± 0.4	62.8 ± 0.4
H69A	49.8 ± 1.8	62.2 ± 0.8
D70A	52.6 ± 0.2	55.2 ± 0.4
K73A	45.0 ± 0.4	65.0 ± 0.6
W102A	49.0 ± 0.4	67.2 ± 0.4
H105A	45.8 ± 0.4	61.2 ± 0.4
H137A	50.8 ± 0.4	62.0 ± 0.2
H137A	50.8 ± 0.4	62.0 ± 0.2
H138A	49.2 ± 0.2	61.8 ± 0.4
S202A	51.2 ± 0.8	66.2 ± 0.4
D205A	50.6 ± 0.6	61.0 ± 0.4
S209A	47.2 ± 0.2	62.4 ± 0.4
F253A	49.6 ± 0.4	65.4 ± 0.6

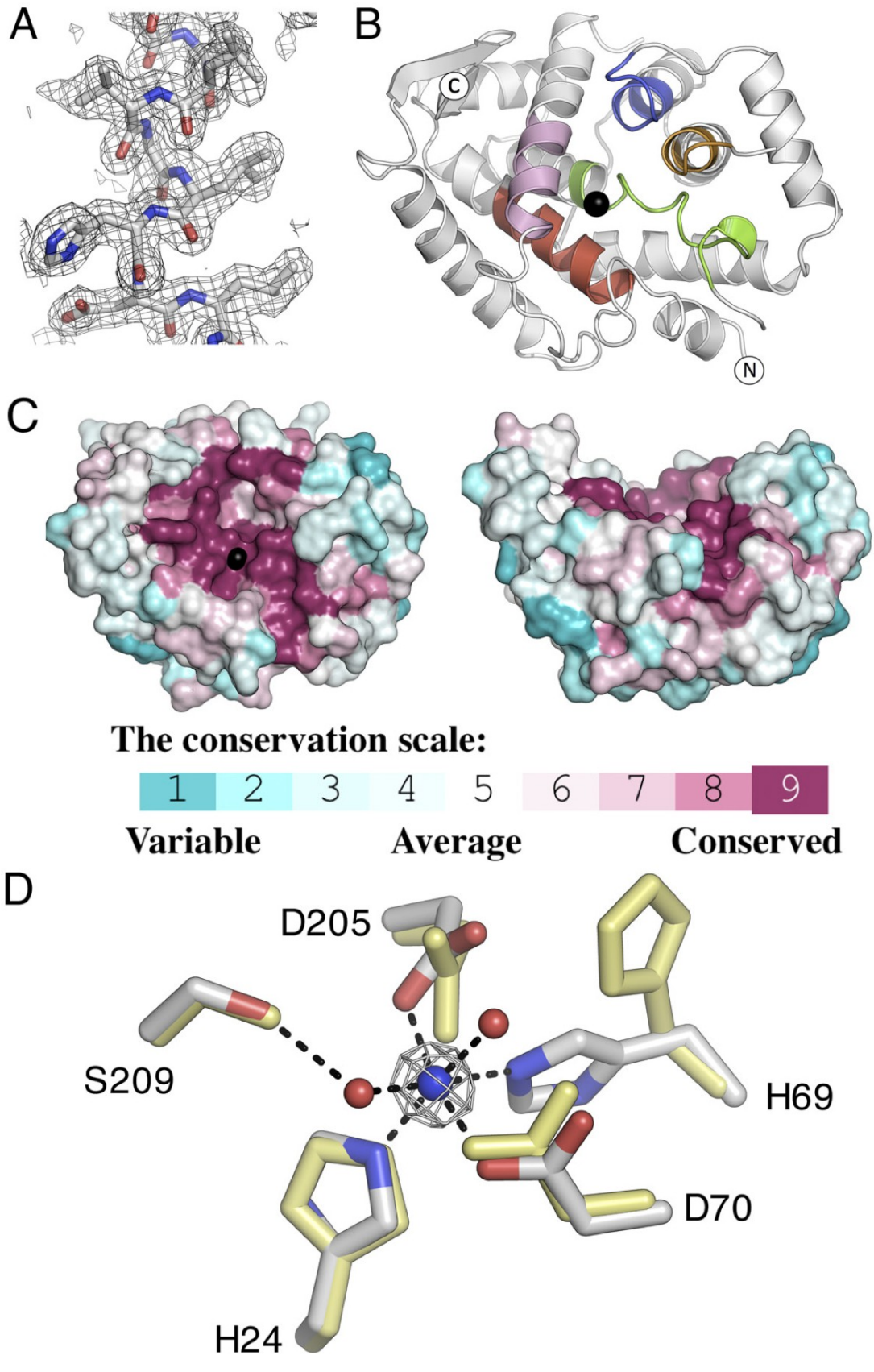

A
B
C
N
C
The conservation scale:
1 2 3 4 5 6 7 8 9
Variable
Average
Conserved
D
D205
S209
H69
D70
H24

Figure 3.1. Crystal structure of *T. thermophilus* Cas3^{HDdom}. *A*, unbiased F_o - F_c electron density map contoured at 3σ. The residues, which are represented as sticks, were omitted from the map calculation. *B*, ribbon trace of the Cas3^{HDdom} structure. The HD domain motifs are colored as follows: motif I (red), motif II (green), motif III (orange), motif IV (blue) and motif V (pink). The black sphere represents the cognate metal binding site. The N and C termini are labeled. *C*, Amino acid sequence conservation scores mapped onto the surface of two orthogonal views of the Cas3^{HDdom} using CONSURF (Armon et al., 2001). The structure to the left is in the same orientation as shown in *B*. The conservation scale is drawn below the two views of the structure. The black sphere represents the cognate metal-binding site. *D*, Cas3^{HDdom} metal ion-binding site. Residues colored by element are in the metal ion-bound configuration. Residues colored yellow are in the protein alone configuration. The Ni^{2+} (blue) and two water molecules (red) are represented as spheres. Anomalous difference electron density map contoured at 5σ (white mesh) reveals the site of the Ni^{2+} ion.

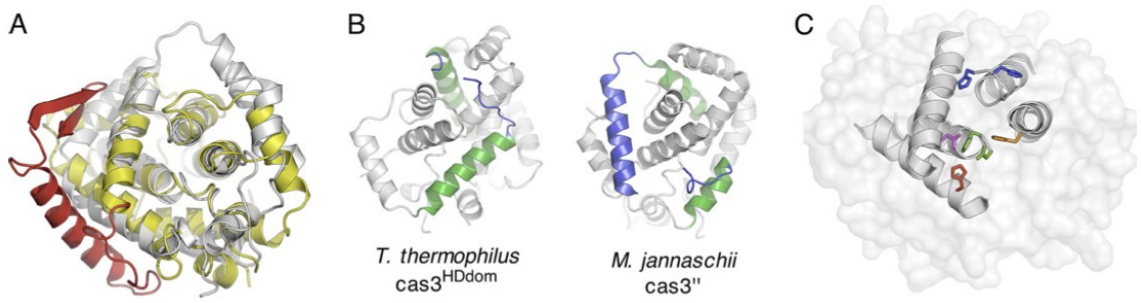


Figure 3.2. Comparison of Cas3^{HDdom} with other HD domains. *A*, structure of *T. thermophilus* Cas3^{HDdom} (yellow and red) superimposed on the structure of *M. jannaschii* Cas3" (white). The additional helix-strand-strand element in the structure of Cas3^{HDdom} is colored red. The molecules are oriented as shown in *B*. *B*, side-by-side view of the structures of *T. thermophilus* Cas3^{HDdom} (left) and *M. jannaschii* Cas3" (right). The helical insertion found in *M. jannaschii* Cas3" and its equivalent loop in *T. thermophilus* Cas3 are colored blue. The helices either side of this region are colored green. *C*, ribbon trace of the minimal core five α-helices of the HD domain taken from the structure of Cas3^{HDdom}. Residues from the five HD domain motifs are colored as shown in *B*. A surface representation of Cas3^{HDdom} is shown in the background.

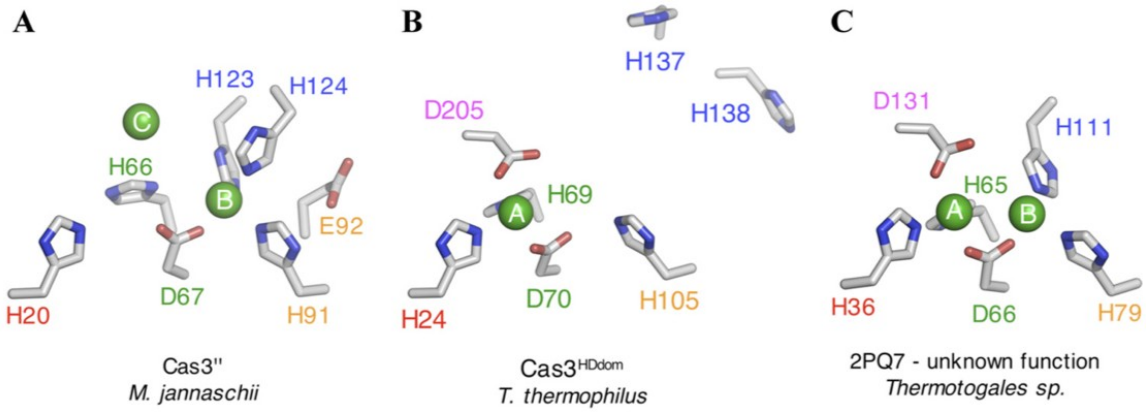


Figure 3.3. Metal ion-binding sites in HD domains. The residues that form the metal ion-binding sites of *M. jannaschii* Cas3" (*A*), *T. thermophilus* Cas3^{HDdom} (*B*), and a protein of unknown function (PDB 2PQ7) from a *Thermotogales* species (*C*). The text color of residue labels indicates the motif that the residue belongs to as shown in Fig. 1*B*. Labeled metal ions are shown as green spheres. His-138 of *T. thermophilus* Cas3^{HDdom} is modeled in two alternative conformers; for clarity, only one of these conformers is shown.

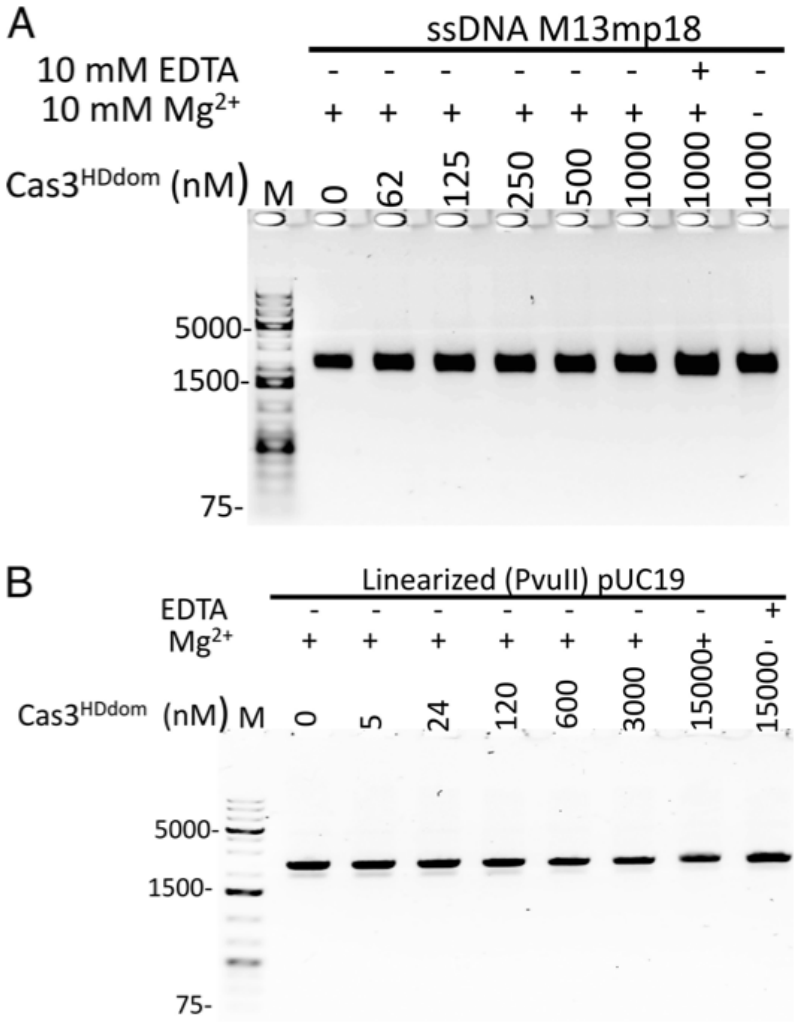


Figure 3.4. The nuclease activity of Cas3^{HDdom} is not activated by Mg^{2+}. *A*, activity of Cas3^{HDdom} on ssDNA in the presence of Mg^{2+}. Reaction mixtures containing 10 mM Tris-HCl, pH 7.5, 60 mM KCl, 10 mM $MgCl_2$, 10% glycerol, and 4 nM circular single-stranded M13mp18, and the indicated amounts of Cas3^{HDdom} were incubated for 2 h at 37 °C. Reactions with EDTA or no added metal served as controls. All reactions were quenched with 20 mM EDTA, and the products were then resolved by a 1% agarose gel and visualized by ethidium bromide staining. *B*, activity of Cas3^{HDdom} on dsDNA in the presence of Mg^{2+}. Assays were performed as above except that the reaction mixtures contained 4 nM pUC19 linearized with PvuII.

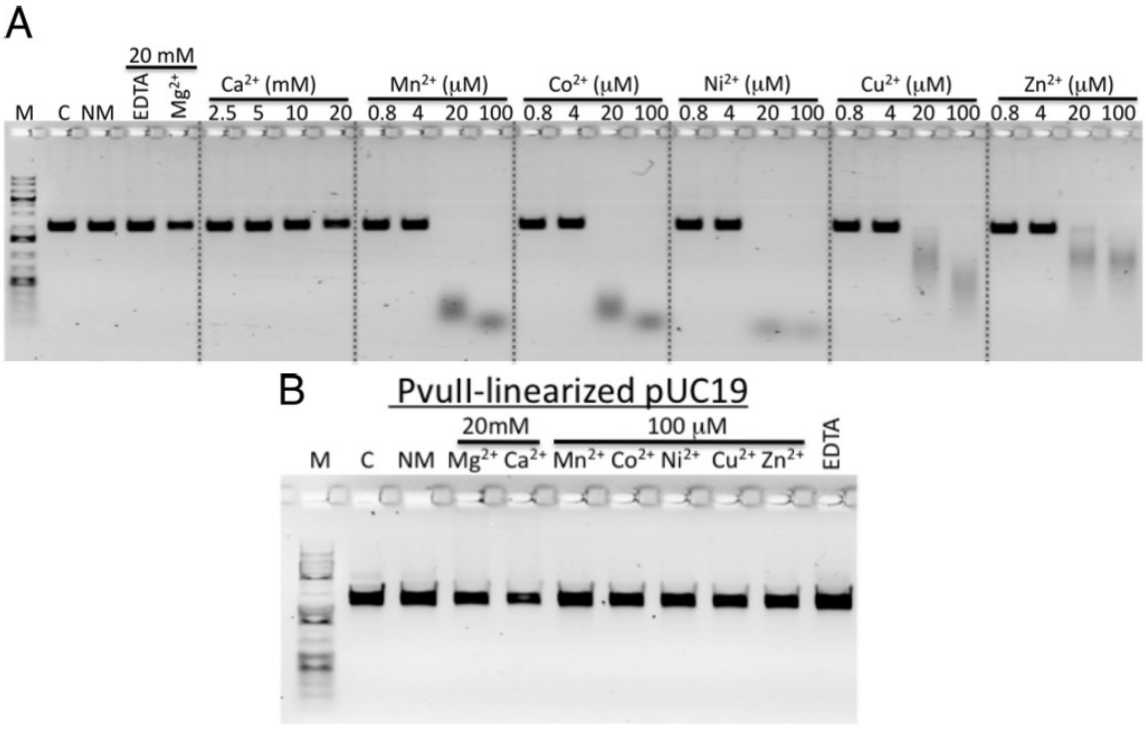


Figure 3.5. The ssDNA endonuclease activity of Cas3^{HDdom} is activated by transition metal ions. *A*, activity of Cas3^{HDdom} in the presence of difference divalent metal ions. Reaction mixtures containing 10 mM Tris-HCl, pH 7.5, 60 mM KCl, 10% glycerol, 1 μM Cas3^{HDdom}, and 4 nM circular single-stranded M13mp18 DNA, and the indicated amount of each divalent metal ion were incubated for 2 h at 37 °C. Reactions with EDTA, no added metal (NM) or no protein (C) served as controls. All reactions were quenched with 20 mM EDTA, and the products were then resolved by a 1% agarose gel and visualized by ethidium bromide staining. *B*, activity of Cas3^{HDdom} on dsDNA in the presence of different divalent metal ions. Assays were performed as above except that reaction contained 4 nM pUC19 linearized with PvuII.

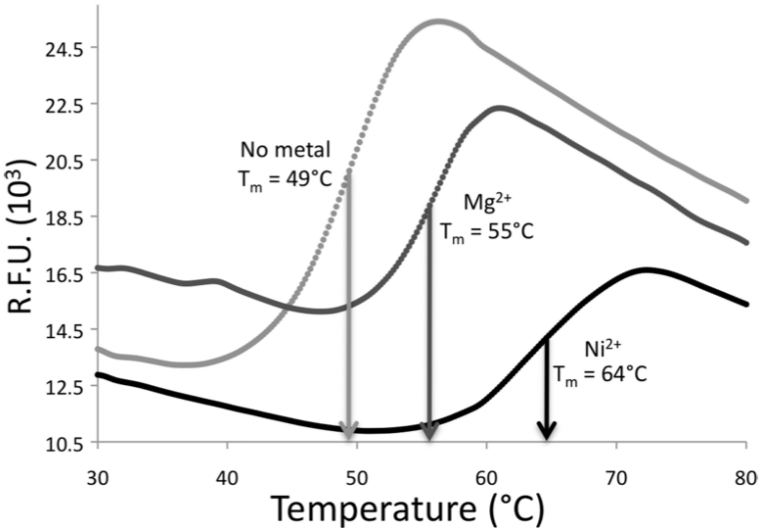


Figure 3.6. Effects of metal ions on the thermal stability of $Cas3^{HDdom}$. Thermofluor assays were performed in the absence of metal or in the presence of either 20 mM Mg^{2+} or 100 μM Ni^{2+}. Arrows indicate the apparent melting temperatures. R.F.U., relative fluorescence units.

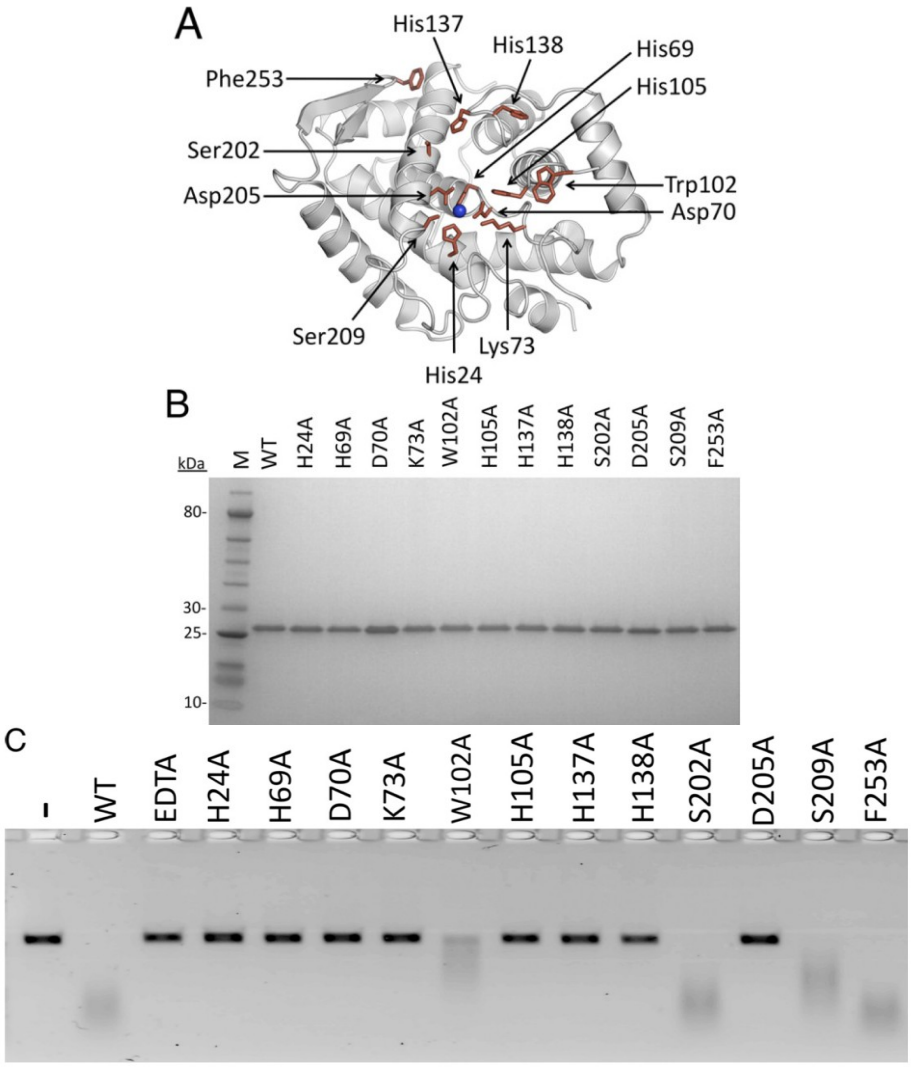


Figure 3.7. Mutational analysis of Cas3^{HDdom}. *A*, ribbon trace of Cas3^{HDdom} (white). Residues selected for mutation are represented as sticks (red), and the Ni^{2+} ion as a sphere (blue). *B*, an SDS-PAGE of the purified mutants, stained with Coomassie Blue. *C*, ssDNA endonuclease activity of the mutants. Reaction mixtures containing 10 mM Tris-HCl, pH 7.5, 60 mM KCl, 10% glycerol, 100 nM Cas3^{HDdom}, and 60 μM NiSO4, and 4 nM circular single-stranded M13mp18 DNA were incubated for 50 min at 37 °C. Reactions with EDTA or no Cas3^{HDdom} (-) served as controls. The reaction products were resolved by a 1% agarose gel and visualized by ethidium bromide staining.

Chapter 4

In vitro reconstitution of the *Escherichia coli* CRISPR system reveals unidirectional, ATP-dependent degradation of the target DNA

Introduction

We previously determined the structure of the Cas3 HD domain from *T. thermophilus* and demonstrated that its single-stranded DNase activity is stimulated by transition metal ions. However, we were not sure how Cas3 is able to degrade a double-stranded target DNA. Thus, to investigate the mechanism of foreign DNA degradation by Cascade, we reconstituted *in vitro* the interference stage of the type-IE CRISPR system from *E. coli.* The results in this chapter have been reprinted here with minor changes from Mulepati, S., and Bailey, S. (2013) *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.* **288**, 22184-22192, with permission from American Society for Biochemistry and Molecular Biology.

Results

Overexpression and Purification of Recombinant E. coli Cas3

To facilitate expression and purification, the gene encoding *E. coli* Cas3 was cloned with an N-terminal His_6 maltose-binding protein tag. The maximum yield of soluble protein (1 mg of pure protein/L of culture) was obtained when cultures were grown at 20 °C, and expression was induced in early log phase (A_{600} of 0.3). Cultures grown at higher temperatures or cultures that were induced above an A_{600} of 0.3 produced

little or no soluble Cas3. Tagged protein was purified from clarified cell lysate by nickel affinity and size exclusion chromatographies. Tobacco etch virus protease was added to remove the tag, and untagged protein was isolated by additional nickel affinity and size exclusion steps. Untagged Cas3 eluted from the size exclusion column at the volume expected for a Cas3 monomer and was over 90% pure, as judged by SDS-PAGE and Coomassie staining (Fig. 4.2A). Mutant variants of Cas3 were produced in a similar manner as wild-type protein, except for co-expression with the chaperone HtpG (Yosef et al., 2011), to compensate for their lower solubility.

Cascade-directed Cleavage of Plasmid DNA by Cas3

The HD domain of Cas3 specifically cleaves ssDNA (Mulepati and Bailey, 2011; Sinkunas et al., 2011; Beloglazova et al., 2011). Previously, we have shown that transition metal ions and not magnesium ions activate the nuclease activity of *Thermus thermophilus* Cas3 (Mulepati and Bailey, 2011). Therefore, we tested the nuclease activity of *E. coli* Cas3 on circular single-stranded DNA (M13 phage) with a selection of divalent metal ions before attempting to reconstitute the activity of the *E. coli* CRISPR system. Consistent with the results from *T. thermophilus*, nickel ions stimulated the nuclease activity of the *E. coli* protein (Fig. 4.2B). Because magnesium ions are necessary for the ATPase activity of Cas3 (Sinkunas et al., 2011), magnesium and transition metal ions were included in subsequent reconstitution assays.

To test whether Cascade can direct stand-alone Cas3 to degrade DNA target in a reconstituted system, we incubated Cas3 and Cascade with a plasmid target bearing functional PAM and complementary protospacer sequences. Following incubation,

proteins were removed by phenol extraction, and the DNA was analyzed by electrophoresis through agarose gels and ethidium bromide staining. We found that in the presence of ATP, Mg^{2+}, and transition metal ions, in particular Co^{2+}, Cascade directed Cas3 to degrade the plasmid target, as shown by a nonspecific smear of dsDNA products on the agarose gel (Fig. 4.3A). Reactions containing either Mg^{2+} or select transition metal ions, but not both, degraded plasmid target to a much lesser extent, consistent with the differing metal ion requirements of the two domains of Cas3 (Fig. 4.3, A–C) (Mulepati and Bailey, 2011; Sinkunas et al., 2011). Control plasmids lacking a protospacer sequence were not degraded. Target degradation was also ablated when a critical residue in the nuclease active site of Cas3 was mutated (D75A) (Fig. 4.3B) (Mulepati and Bailey, 2011; Sinkunas et al., 2011; Beloglazova et al., 2011; Westra et al., 2012).

Previous electrophoretic mobility shift assays have demonstrated that Cascade requires the CasA subunit when binding to dsDNA target (Sashital et al., 2012; Mulepati et al., 2012; Westra et al., 2012). Consistently, a subcomplex of Cascade lacking the CasA subunit (CasBCDE) was unable to direct degradation of the plasmid target. The addition of CasA to the reaction restored this activity (Fig. 4.3D).

ATP is required for DNA target degradation (Westra et al., 2012) (Fig. 4.3). In the absence of ATP, the Cascade-Cas3 fusion was shown to nick the DNA target, and an ATPase-deficient variant of the fusion nicked the target both in the presence or absence of ATP (Westra et al., 2012). With 50 nM stand-alone Cas3, only 13% of the plasmid target was nicked in the absence of ATP (Fig. 4.3A). However, when the concentration of Cas3 was varied, nicking activity increased in a concentration-dependent manner (Fig. 4.3E); 68% of the target was nicked at 300 nM Cas3. When ATP was included in the

reaction, target was completely degraded except at the lowest concentrations of Cas3. In addition, the nicking activity of an ATPase-deficient variant (D452A) of Cas3 was stimulated by the presence of ATP (Fig. 4.3E). In the absence of ATP, the variant Cas3 nicked 55% of the target DNA, but in the presence of ATP, close to 100% of the target DNA was nicked. These data suggest that Cas3 recruitment to target DNA is stimulated by the binding but not the hydrolysis of ATP. To further examine the effects of ATP on Cas3 activity, we monitored target degradation as a function of ATP concentration (Fig. 4.3F). At high ATP concentrations, we observed a smear on the agarose gel corresponding to degradation products with a wide range of sizes. At lower ATP concentrations, the average product size decreased and spanned a smaller range. These results suggest that the frequency of cutting by the nuclease domain is coupled to the rate of DNA unwinding by the helicase domain.

Cascade Bound to DNA Target Activates the ATPase Activity of Cas3

The helicase domain of S. thermophilus Cas3 harbors both ATP-dependent helicase and ssDNA-dependent ATPase activities (Sinkunas et al., 2011). Using an NADH-coupled assay (Kiianitsa et al., 2003), we investigated the ATPase activity of E. coli Cas3 by testing the effects of reaction components on the rate of ATP hydrolysis (Fig. 4.3G). ATPase activity was not stimulated by dsDNA and was stimulated only modestly by ssDNA (~3-fold). The addition of Cascade alone failed to stimulate the ATPase activity, but with the addition of plasmid target, the rate of ATP hydrolysis was stimulated 44-fold. This stimulation is dependent on base pairing between the crRNA and protospacer sequences because targets lacking a protospacer failed to stimulate the

ATPase activity. No ATPase activity was detected with an ATPase-deficient variant (D452A) of Cas3. These results suggest that the ATPase activity of Cas3 is tightly regulated and relies on the recruitment of Cas3 by Cascade to a protospacer.

Degradation of DNA Targets Requires both PAM and Seed Sequences

Mutations in the PAM or seed sequences of DNA targets render cells with an otherwise functional CRISPR system sensitive to phage infection (Semenova et al., 2011). Binding studies revealed that this is a result of the reduced affinity between Cascade and the mutant DNA targets (Semenova et al., 2011; Sashital et al., 2012). To determine if the activity of our reconstituted CRISPR system is also dependent on PAM and seed sequences, we monitored nicking activity on plasmid targets containing point mutations in either the PAM or the protospacer. These reactions were performed in the absence of ATP to avoid smearing of the DNA products on the agarose gels, allowing us to quantify the activity through the ratio of nicked product to negatively supercoiled substrate. Mutations in the PAM sequence abolished target nicking, mutations in the seed sequence reduced nicking activity (particularly at positions 1 and 4), and mutations outside the seed region generally had little to no effect (Fig. 4.4A). We also tested the ability of these variant targets to activate the ATPase activity of Cas3 (Fig. 4.4B) and found that the mutations had similar effects on ATPase activation as they had on nicking activity. Altogether, these results establish that the reconstituted assay recapitulates the observed in vivo dependence for target PAM and seed sequences.

Cascade Can Direct Cas3 to Degrade Linear DNA, and Degradation Is Unidirectional

To determine if Cascade and stand-alone Cas3 can degrade linear DNA and if degradation proceeds from the protospacer in one or both directions, plasmid targets were linearized using either of two restriction enzymes, KpnI or ScaI. The protospacer is positioned 3 kb from the 5'-end of the target strand in the KpnI-treated plasmid and 2 kb away in the ScaI-treated plasmid. After reaction with the reconstituted CRISPR system, the linear KpnI- and ScaI- treated targets were clearly degraded, yielding products that were resistant to degradation of 3 and 2 kb, respectively (Fig. 4.5A). This pattern of resistance suggests that degradation is unidirectional, initiating in or near the protospacer and proceeding upstream, leaving the downstream DNA intact (Fig. 4.5A). As observed with negatively supercoiled targets, degradation of linear DNA was also found to be ATP- and Cascade-dependent, and mutation of either the nuclease (D75A) or helicase domain (D452A) of Cas3 ablated this degradation (Fig. 4.5B).

To investigate if negative supercoiling affects the rate of target degradation by Cascade and stand-alone Cas3, we compared the rates of degradation of negatively supercoiled with linearized plasmid targets (Fig. 4.5C). Fitting the data to a single-exponential decay yielded observed rate constants (k_{obs}) of 2.92 and 0.66 min^{-1} for negatively supercoiled and linear target, respectively (Fig. 4.5C). This suggests that the *E. coli* CRISPR system prefers negatively supercoiled target to linear target by 4.5-fold. Consistent with the nuclease assay, both substrates stimulated ATPase activity, but activity with supercoiled target was greater than that of the linearized target by 2-fold (Fig. 4.3G).

Mapping Degradation of Target DNA by Cas3

When Cascade binds to foreign DNA, the crRNA base-pairs to the target strand and displaces the non-target strand. DNA footprinting experiments show that the majority of the protospacer DNA is protected when bound to Cascade except for a 19-base region of the non-target strand (Fig. 4.1) (Jore et al., 2011). To determine if Cas3 nicks this accessible region, we performed a reconstitution assay in the absence of ATP, purified the nicked product from an agarose gel, and sequenced it using primers that flanked the protospacer region. A clear interruption in the sequence of the non-target strand was observed, whereas the sequence of the target strand was uninterrupted (Fig. 4.6A), indicating that nicking occurs in the accessible region of the non-target strand 11 bases from the 3'-end of the PAM. Next, we performed similar experiments sequencing the linear product, enriched in assays containing low concentrations of ATP (Fig. 4.6A). Again, a clear interruption in the sequence of the non-target strand was observed, 11 bases from the 3'-end of the PAM (Fig. 4.6A). However, sequence information from the target strand was unread- able in the region of the protospacer, consistent with the presence of multiple cuts in this strand (Fig. 4.6A).

To map the degradation of target DNA in more detail, we repeated reconstitution assays on synthetic dsDNA targets, one labeled with ^{32}P at the 5'-end of the target strand and the other at the 5'-end of the non-target strand. In the absence of ATP (or in the presence of ATP but using the ATPase-deficient mutant of Cas3, D452A), the target strand was not cleaved, whereas the non-target strand was cut weakly within the protospacer, 7 and 11 bases from the PAM sequence (Fig. 4.6, B–D). When ATP was included in the reactions, multiple cuts were observed in both strands. In the target strand,

cleavage occurred in the region 3' of the protospacer and in the flanking upstream DNA (Fig. 4.6, B and D). A similar cleavage pattern was observed for the non- target strand (Fig. 4.6, C and D). These results reaffirm that degradation of target DNA is unidirectional because we observe no cleavage downstream of the protospacer sequence. The nuclease-deficient mutant (D75A) of Cas3 did not cut the synthetic DNA target. Targets lacking a PAM sequence also failed to be cut by wild-type Cas3.

Discussion

During the interference stage, the *E. coli* Type I-E system proceeds through the identification and degradation of foreign DNA. Cascade recognizes foreign DNA and then recruits Cas3 for the ATP-dependent degradation of the target. Studies of the *E. coli* system have greatly increased our understanding of target recognition (Jore et al., 2011; Wiedenheft et al., 2011; Semenova et al., 2011; Sashista et al., 2012; Mulepati et al., 2012; Westra et al., 2012). However, the mechanisms underlying Cas3 recruitment and subsequent target degradation are poorly understood. This could be a result of an inability to produce a recombinant form of stand-alone *E. coli* Cas3 suitable for bio- chemical analysis. Here, we report the production of stand-alone *E. coli* Cas3 with which we could reconstitute the *E. coli* Type I-E system in vitro. Using this in vitro system, we investigate the mechanism of Cas3 recruitment and subsequent target degradation.

Cascade binding to target DNA is a prerequisite for recruitment of Cas3. For Cascade to bind, DNA targets require a protospacer complementary to the crRNA and a PAM (Semenova et al., 2011). Cascade binding generates an R-loop structure in the target DNA that exposes part of the non-target strand (Figs. 4.1 and 4.7) (Jore et al.,

2011). Our results indicate that this exposed ssDNA serves as the binding platform for Cas3 and is also the site for the initial nicking of the DNA target (Figs. 4.6 and 4.7). Thus, complex formation between Cascade and target DNA provides Cas3 with the ssDNA required both for loading the helicase domain and as the substrate for nicking by the nuclease domain. Additional protein- protein interactions with Cascade, in particular the CasA subunit, may also play a role in recruitment (Westra et al., 2012). Nicking of target does not require ATP hydrolysis (Westra et al., 2012) but is stimulated by the presence of ATP (Fig. 4.3, A and E), probably because ATP binding stimulates recruitment of Cas3. Mutations in the PAM and seed sequence, which reduce the binding affinity of Cascade (Semenova et al., 2011), inhibit the cleavage of DNA target (Fig. 4.4A). Thus, the nuclease activity of Cas3 is tightly regulated. Only DNA that has been correctly engaged by Cascade and formed an R-loop will be degraded. Similarly, we also find that the ATPase activity is tightly regulated, being significantly activated only in situations where Cascade can form an R-loop with DNA target (Fig. 4.3G).Tight regulation is presumably necessary to control the deleterious effects Cas3 could have on the host chromosome or other beneficial DNA within the cell.

Following nicking, further DNA cleavage requires ATP hydrolysis by Cas3 (Fig. 4.3, A and E), presumably to provide the energy for DNA unwinding, which generates the ssDNA substrate for the nuclease domain (Fig. 4.7). The coupling of dsDNA unwinding to ssDNA degradation is reminiscent of the mechanism employed by the RecBCD family of enzymes in homologous recombination (Wigley, 2013). To further investigate DNA target degradation, we mapped the sites of this ATP-dependent cleavage using labeled synthetic DNA. We found that Cas3 extensively cuts both strands within

the protospacer and upstream of the PAM (Fig. 4.6). This, as well as results from monitoring degradation of linear plasmids (Fig. 4.5), shows that the progression of target degradation is unidirectional, proceeding only upstream of the protospacer (Fig. 4.7). Cas3 may also have an active role in recycling Cascade (Sinkunas et al., 2011) because we also observe cuts in the target strand of the protospacer (Fig. 4.6), suggesting that the target strand has been unwound from the crRNA (Fig. 4.7). Consistently, *E. coli* Cas3 has been shown to harbor ATP-dependent R-loop unwinding activity (Howard et al., 2011).

The activities of the two domains of Cas3 are coupled because the helicase domain generates the substrate for the nuclease domain. We monitored degradation of plasmid target as a function of ATP concentration to gain further insight into this coupling (Fig. 4.3F). The unwinding activity of the helicase domain should increase with ATP concentration. Our results suggest that, under these conditions, the nuclease domain cuts the DNA less frequently, giving rise to products with a wide range of sizes. When the helicase rate is low, as observed with lower ATP concentrations, the nuclease domain makes cuts more frequently, which generates smaller sized products.

Genetic screening and expression experiments have shown that the chaperone HtpG positively modulates E. coli Type I-E resistance by maintaining functional levels of Cas3 (Yosef et al., 2011). Consistent with this, we have shown that R-loop formation by Cascade is sufficient to recruit Cas3 to DNA targets, suggesting that additional factors, such as HtpG, are not essential at this step.

The Cascade-Cas3 fusion has been shown to degrade negatively supercoiled but not relaxed (i.e. nicked or linear) DNA (Westra et al., 2012). In our reconstituted system, stand-alone Cas3 can degrade both negatively supercoiled and linear DNA (Fig. 4.5).

However, the rate of degradation of negatively supercoiled DNA is greater than that of linear DNA by 4.5-fold. Negatively supercoiled DNA is probably a better substrate because of the increased energy required to melt the DNA strands over the length of the protospacer in relaxed versus negatively supercoiled DNA (Westra et al., 2012). Indeed, negative supercoiling stimulates other processes that rely on strand separation, such as RecA-mediated homologous recombination (Cai, 2001). Because the most likely substrate for the Type I CRISPR systems *in vivo* is negatively supercoiled DNA (Westra et al., 2012), further analysis of Type I systems, with both fused and stand-alone Cas3, will be needed to understand if there is functional significance to targeting relaxed DNA.

While this manuscript was in preparation, Sinkunas et al. (Sinkunas et al., 2013) reported the *in vitro* reconstitution of the Type I-E CRISPR system from *S. thermophilus*. Like E. coli, this system contains stand-alone Cas3 and Cascade. In agreement with the results reported here, they show that the reconstituted *S. thermophilus* system is able to cleave linear DNA and that target degradation is unidirectional. They also go on to map the cleavage sites, revealing a pattern similar to that observed in the *E. coli* CRISPR system. Thus, the molecular mechanisms of the Type I-E CRISPR systems appear conserved.

Materials and methods

Cloning and Mutagenesis

The genes encoding *E. coli* Cas3 and high temperature protein G (HtpG) were amplified from genomic DNA (American Type Culture Collection) and directionally cloned into pMAT and pRSFDuet-1 (Novagen), respectively. pMAT was engineered by

inserting DNA encoding maltose-binding protein into the SpeI site of pHAT4 (Peränen, 1996). QuikChange site-directed mutagenesis (Stratagene) was used to create point mutants. Plasmid targets were prepared by cloning synthetic oligonucleotides carrying the appropriate sequence into pBAT4 (Peränen, 1998). Primers and oligonucleotides are listed in Table 4.1. All clones were verified by DNA sequencing.

Protein Expression and Purification

E. coli Cascade, CasA, and a subcomplex of Cascade lacking the CasA subunit (CasBCDE) were expressed and purified as described previously (Mulepati et al., 2012). *E. coli* Cas3 was overexpressed in the T7Express strain of *E. coli* (New England Biolabs). Cells were grown at 20 °C to an A_{600} of 0.3, at which point protein expression was induced with 0.2 mM isopropyl-β-D-1-thiogalactopyranoside. After overnight growth, the cells were harvested, lysed in buffer L (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, and 10% glycerol), clarified by centrifugation, and loaded onto a 5-ml immobilized metal affinity chromatography column (Bio-Rad). The column was washed consecutively with buffer L supplemented with 5 mM imidazole and then 1 M NaCl. The remaining bound proteins were eluted with buffer L supplemented with 250 mM imidazole. The sample was directly loaded onto a HiLoad 26/60 S200 size exclusion column (GE Healthcare) pre-equilibrated in buffer A (20 mM Tris-HCl, pH 8.0, 200 mM NaCl, and 1 mM dithiothreitol). Fractions containing Cas3 were pooled and desalted into buffer B (20 mM Tris-HCl, pH 8.0, and 200 mM NaCl). The N-terminal His_6-maltose-binding protein tag was removed by overnight treatment with tobacco etch virus protease at 4 °C. The cleaved sample was then flowed through an immobilized metal affinity chromatography

column, concentrated, and loaded onto a HiLoad 26/60 S200 size exclusion column pre-equilibrated with buffer B. Purified Cas3 was concentrated to 5 μM, flash-frozen, and stored at -80 °C. The D75A and D452A Cas3 mutants were co-expressed with HtpG in T7Express cells and purified like the wild-type protein.

Preparation of Synthetic DNA Targets

PAGE-purified oligonucleotides (Table 4.1) were 5'-labeled with γ-[^{32}P] ATP (PerkinElmer Life Sciences) using T4 polynucleotide kinase (New England Biolabs). Duplexes were formed by mixing the target and non-target strands, heating at 95 °C for 2 min, and then cooling to room temperature over 2 h. DNA ladders were prepared using a Sanger sequencing kit (Asymmetrix).

Reconstitution Assay

Reactions were performed in buffer containing 5 mM HEPES, pH 7.5, and 60 mM KCl. The indicated amounts of divalent metal ions, target DNA, Cascade, Cas3, and ATP were assembled together and incubated at 37 °C for 30 min or the indicated duration. All reactions were terminated by the addition of 20 mM EDTA. The range in divalent metal ion concentrations was chosen based on their estimated cellular concentrations (Graham et al., 2009; Macomber et al., 2011; Anjem et al., 2009). Proteins were removed by phenol extraction. Plasmid DNA was analyzed by electrophoresis through 1% agarose gels and ethidium bromide staining. Labeled synthetic DNA was analyzed by electrophoresis through 10% polyacrylamide gels and autoradiography.

ATPase Assay

ATP hydrolysis by Cas3 was monitored using an NADH-coupled ATPase assay as described previously (Kiianitsa et al., 2003). Two reaction mixtures were prepared, each in 10 mM HEPES, pH 7.5, 60 mM KCl, and 10% glycerol. Mixture A contained 0.5 mM NADH, 4 mM ATP, and 20 mM $MgCl_2$ as well as 4 nM DNA where indicated. Mixture B contained 6 mM phosphoenol pyruvate (Sigma-Aldrich) and 0.4 units/l pyruvate kinase/lactate dehydrogenase (Sigma-Aldrich) as well as 40 nM Cascade and/or 200 nM Cas3 where indicated. Mixtures A and B were incubated separately at 37 °C for 10 min before equal volumes of both were mixed to initiate a 100-µl reaction. Absorbance at 340 nM was measured every 30 s for 10 min. The rate of NADH oxidation was calculated from the linear decrease in A_{340}. All reactions were performed at 37 °C.

Acknowledgements

We thank Brian A. Learn for helpful discussions and Jennifer M. Kavran for critical reading of the manuscript.

Table 4.1. Primers and oligonucleotides used in these studies

	Sequences (5'-3')
Primers for gene amplification	
Cas3 forward	GTGTGTGAATTCATGGAACCTTTTAAATATATATGCC
Cas3 reverse	GTGTGTCTCGAGTTATTTGGGATTTGCAGGGATG
HtpG forward	GTGTGTCCATGGCGAAAGGACAAGAAACTCGTGG
HtpG reverse	GTGTGTGAATTCTCAGGAAACCAGCAGCTGGTTC
Primers for site directed mutagenesis	
D75A forward	GTTATTTTTCATTGCTCTTCATGCTATTGGAAAGTTTGATATACG
D75A reverse	CGTATATCAAACTTTCCAATAGCATGAAGACCAATGAAAAATAAC
D452A forward	GTCGAAGTGTTTTAATTGTTGCTGAAGTTCATGCTTACGACAC
D452A reverse	GTGTCGTAAGCATGAACTTCAGCAACAATTAAAACACTTCGAC
Oligonucleotides used to construct the plasmid target[a]	
Target strand	CATGGACAGCCCACATGGCATTCCACTTATCACTGGCATG
Non-target strand	AATTCATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTC
Oligonucleotides used to construct synthetic DNA targets	
Target strand	TTAAAGGCCGCTTTTTCAATCTACACAATTGAGCAAATCAGACAGCCCA CATGGCATTCCACTTATCACTGGCATTGATTTGCTCAATTTTGTAGATTG ACGGAACGAGGGTAGA
Non-target strand	TCTACCCTCGTTCCGTCAATCTACAAAATTGAGCAAATCAATGCCAGTG ATAAGTGGAATGCCATGTGGGCTGTCTGATTTGCTCAATTGTGTAGATTG AAAAAGCGGCCTTTAA

[a] Oligonucleotides used to construct plasmid targets with variant Protospacer and PAM sequences contained mutations as indicated throughout.

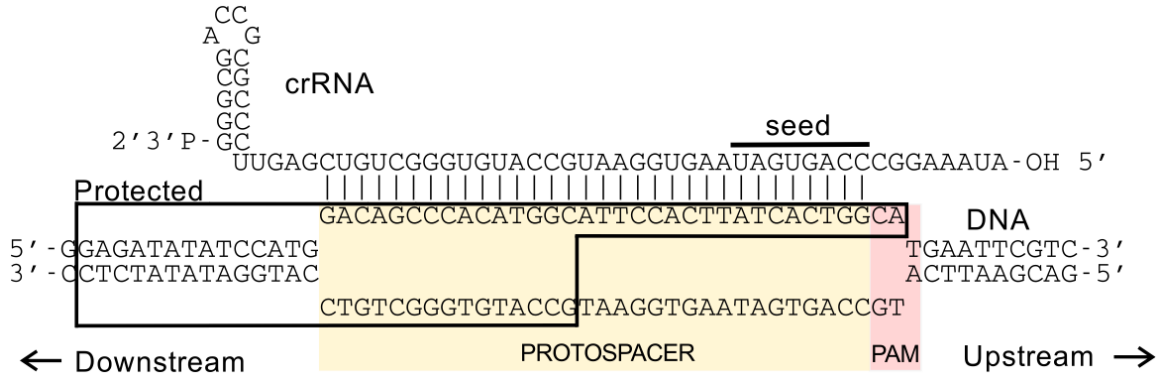


Figure 4.1. Schematic representation of the R-loop formed between Cascade and DNA target. The positions of the PAM and protospacer are shaded yellow and red, respectively. The location of the seed sequence is also indicated. Outlining delineates the region of the DNA protected in footprinting experiments (Jore et al., 2011).

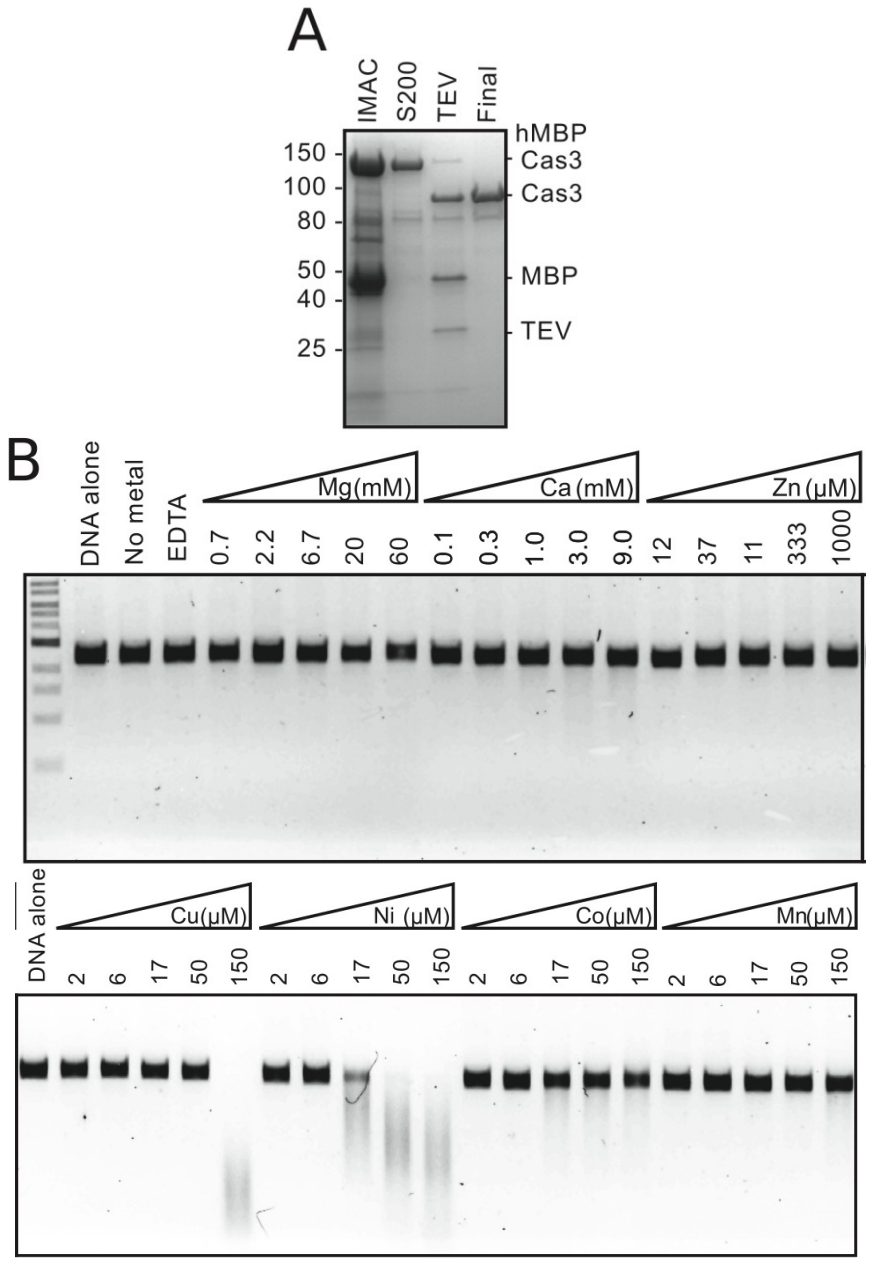
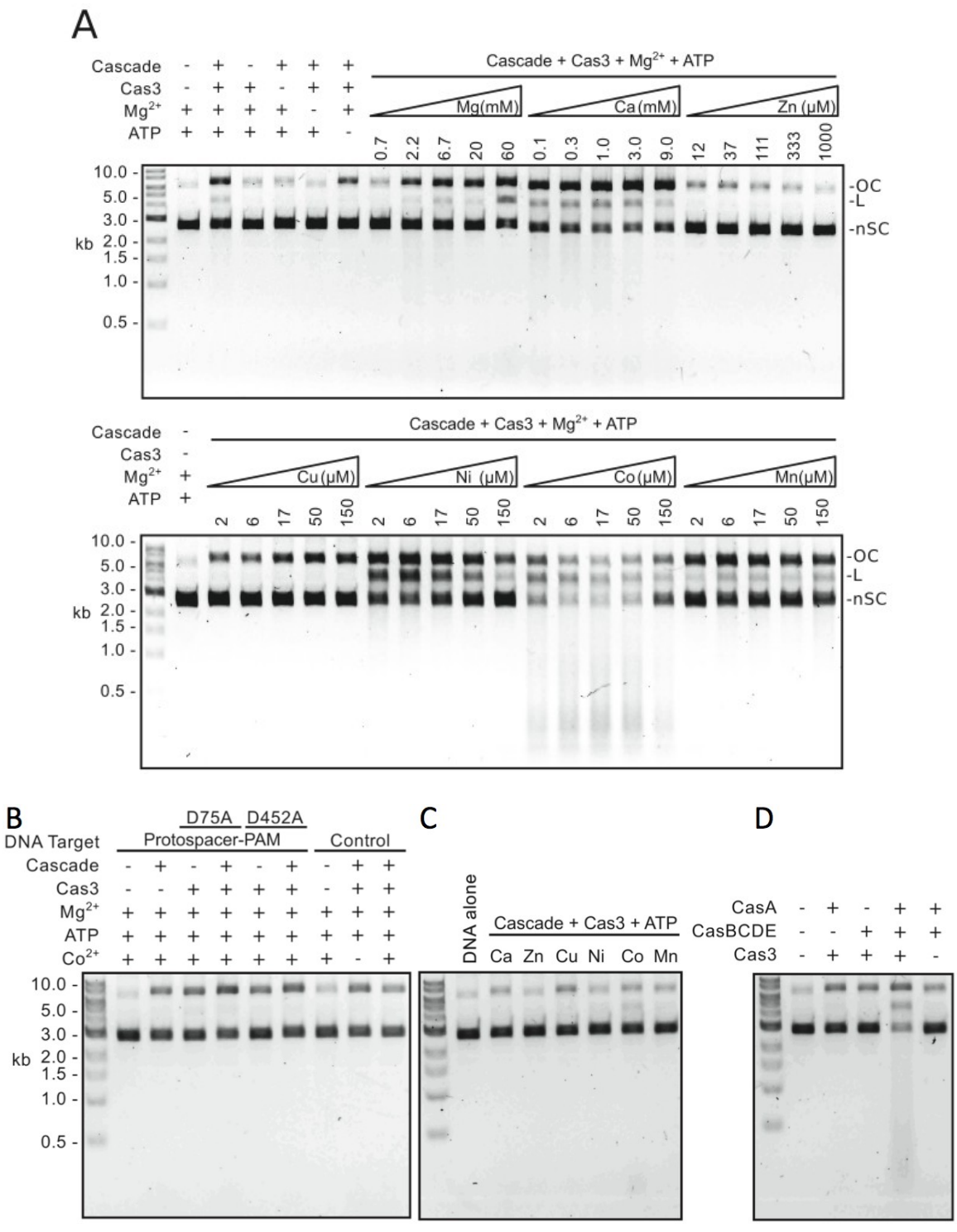


Figure 4.2. Purification and single-stranded DNA nuclease activity of *E. coli* Cas3. *A*, Coomassie-stained SDS-polyacrylamide gel of samples taken during purification of Cas3. *B*, Cas3 nuclease activity is stimulated by transition metals. Reaction mixtures containing 500 nM Cas3, 4 nM circular, single-stranded M13mp18 DNA, and different metal ions, as indicated, were incubated for 1 h at 37 °C.


A
Cascade
Cas3
Mg2+
ATP
Cascade + Cas3 + Mg2+ + ATP
Mg(mM)
Ca(mM)
Zn (μM)
0.7 2.2 6.7 20 60 0.1 0.3 1.0 3.0 9.0 12 37 111 333 1000
10.0
5.0
3.0
kb 2.0
1.5
1.0
0.5
-OC
-L
-nSC
Cascade + Cas3 + Mg2+ + ATP
Cu (μM)
Ni (μM)
Co (μM)
Mn(μM)
2 6 17 50 150 2 6 17 50 150 2 6 17 50 150 2 6 17 50 150
B
DNA Target
D75A D452A
Protospacer-PAM
Control
Co2+
C
DNA alone
Cascade + Cas3 + ATP
Ca Zn Cu Ni Co Mn
D
CasA
CasBCDE
Cas3

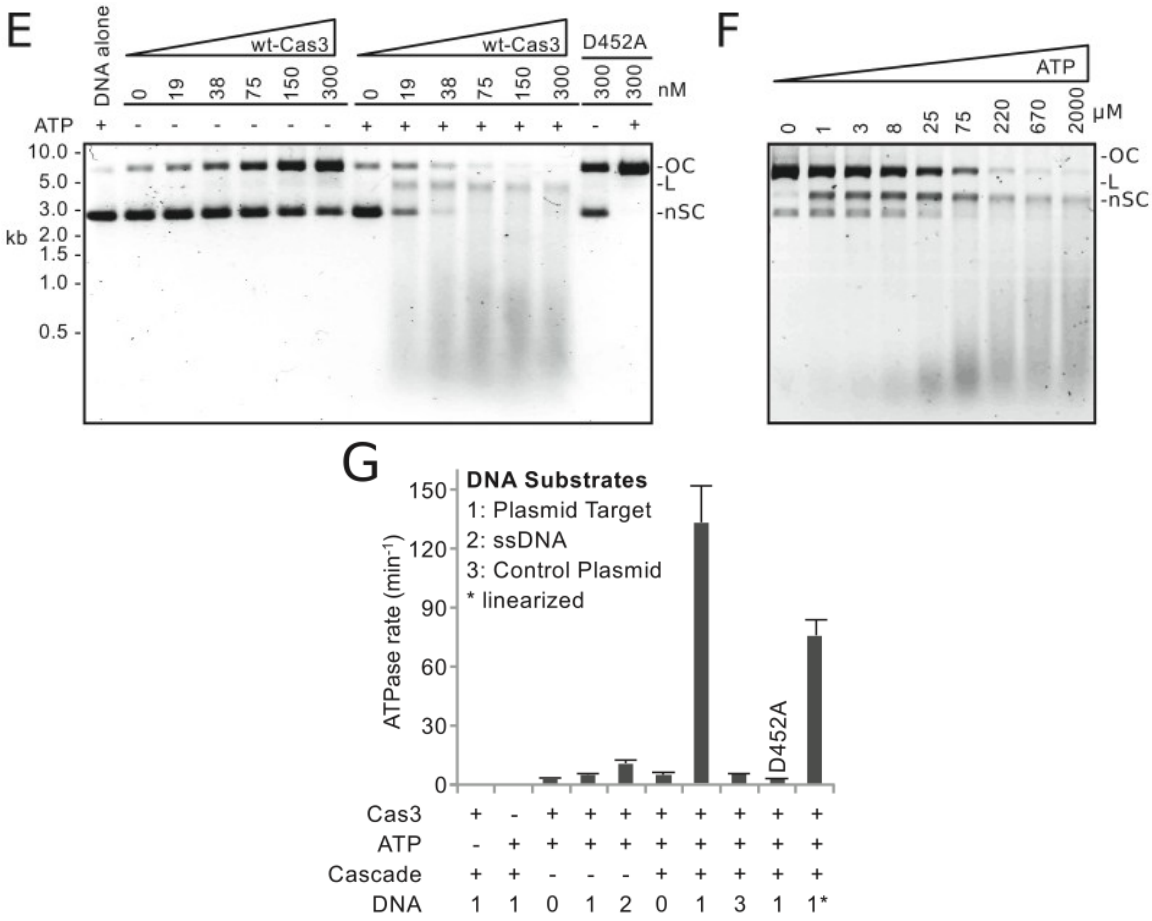


Figure 4.3. Cascade-mediated nuclease and ATPase activities of Cas3. *A*, nuclease activity of Cas3. Reaction mixtures containing 20 nM Cascade, 2 mM ATP, 2 nM plasmid DNA, and 50 nM Cas3 were incubated for 30 min at 37 °C. Metal ions, when included, were at the concentrations indicated. If not indicated, Mg^{2+} was included at 10 mM. *B*, reactions performed as in A with 10 mM Mg^{2+} and, when present, 10 μM Co^{2+}. The control plasmid lacks a protospacer. *C*, reactions performed as in A with either 9 mM Ca^{2+}, 1 mM Zn^{2+}, 150 μM Cu^{2+}, 150 μM Ni^{2+}, 150 μM Co^{2+}, or 150 μM Mn^{2+}. *D*, the CasA subunit is necessary for plasmid degradation. Reaction mixtures containing 20 nM CasA and/or 20 nM CasBCDE with 10 mM Mg^{2+}, 10 μM Co^{2+}, 2 mM ATP, 2 nM plasmid DNA, and 50 nM Cas3 were incubated for 30 min at 37 °C. E, target cleavage

was monitored at increasing concentrations (as indicated) of Cas3 in the absence or the presence of 2 mM ATP. All reactions contained 10 mM Mg^{2+} and 10 μM Co^{2+}. *F*, target cleavage was monitored as a function of ATP concentration (as indicated). All reactions contained 50 nM Cas3, 10 mM Mg^{2+}, and 10 μM Co^{2+}. In *A–D*, the position of negatively supercoiled (nSC), linear (L), and nicked or open circle (OC) DNA is indicated. *G*, rates of ATPase hydrolysis. Error bars, S.D. of the rate constant, taken from at least three independent measurements. DNA substrates are plasmid target containing PAM and protospacer (1), single-stranded M13 phage DNA (2), and control plasmid, lacking a protospacer (3). *, plasmid DNA was linearized before the reaction.

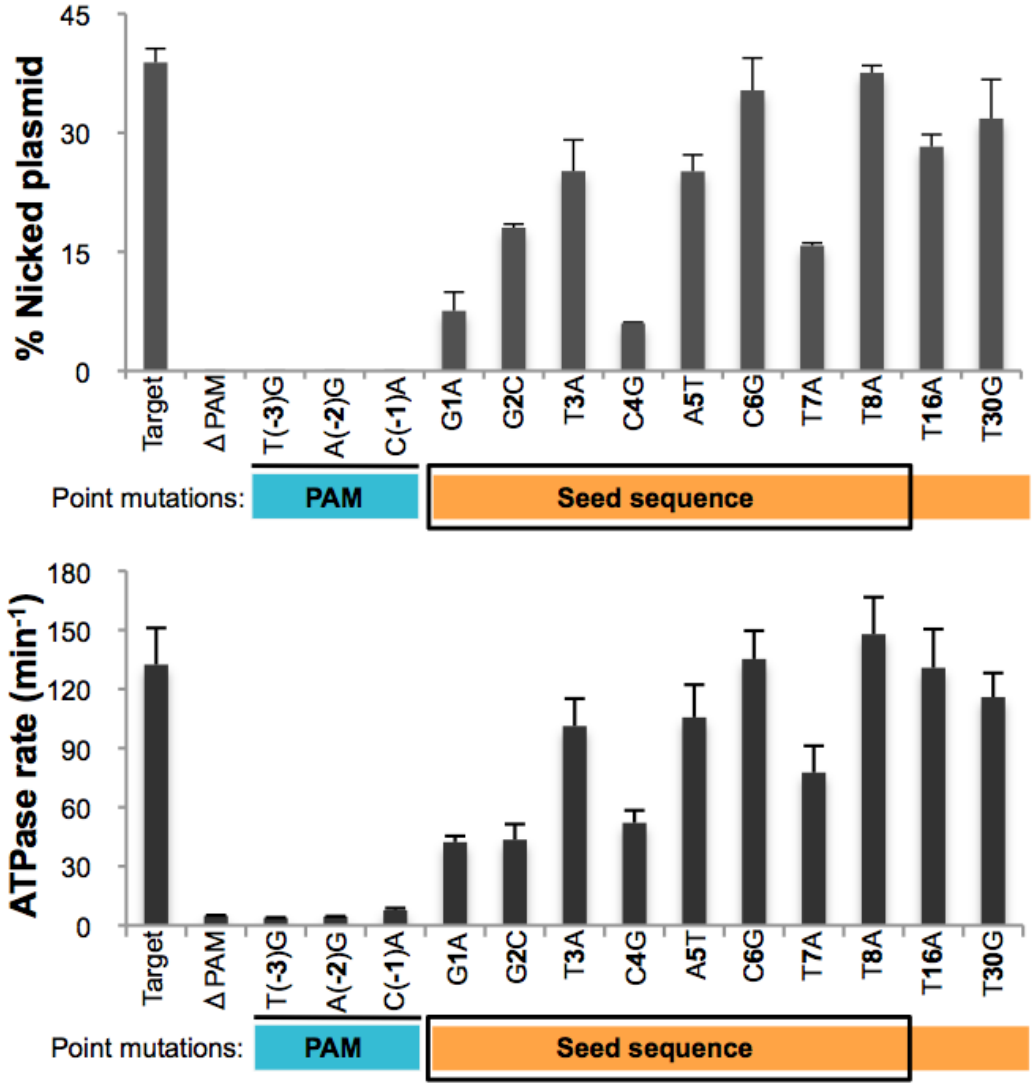


Figure 4.4. Degradation of target DNA requires both the PAM and seed sequences. *A*, extent of nicking by Cas3 (100 nM) using negatively supercoiled target (2 nM) containing the indicated point mutations in the PAM and protospacer. All reactions contained 10 mM Mg^{2+} and 10 μM Co^{2+}. *B*, rate of ATP hydrolysis by Cas3 (100 nM) in the presence of the mutant DNA targets as in *A*. In both panels, error bars indicate S.D., taken from at least three independent measurements. PAM, DNA target containing a protospacer sequence but not a PAM.

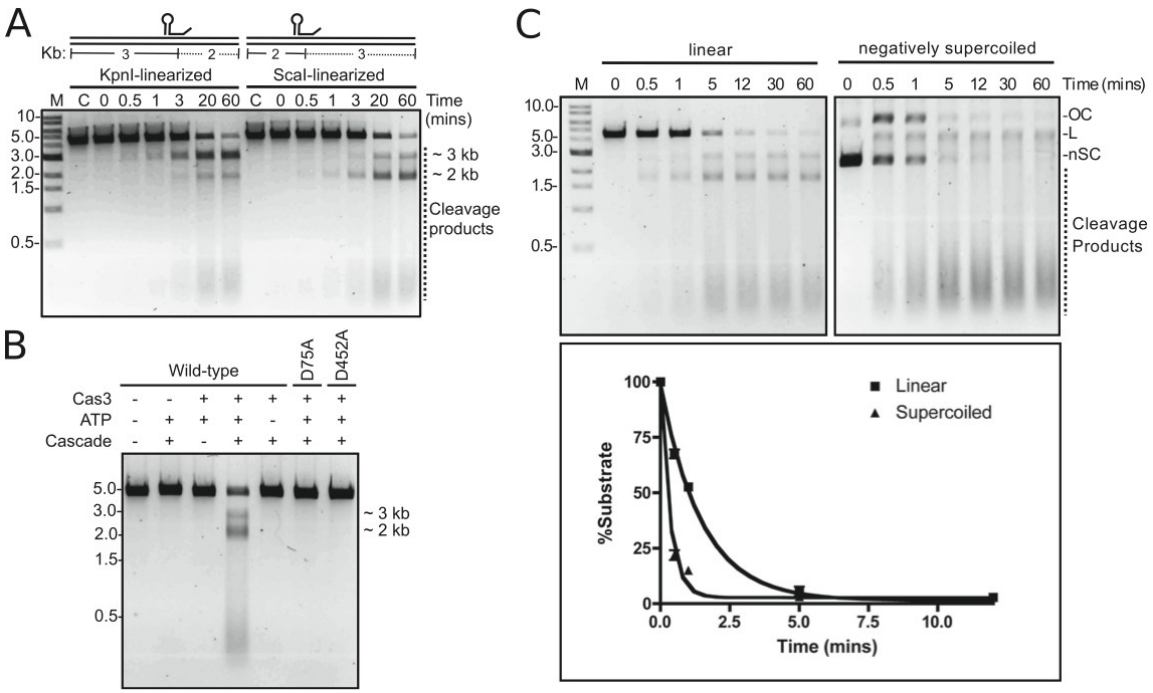
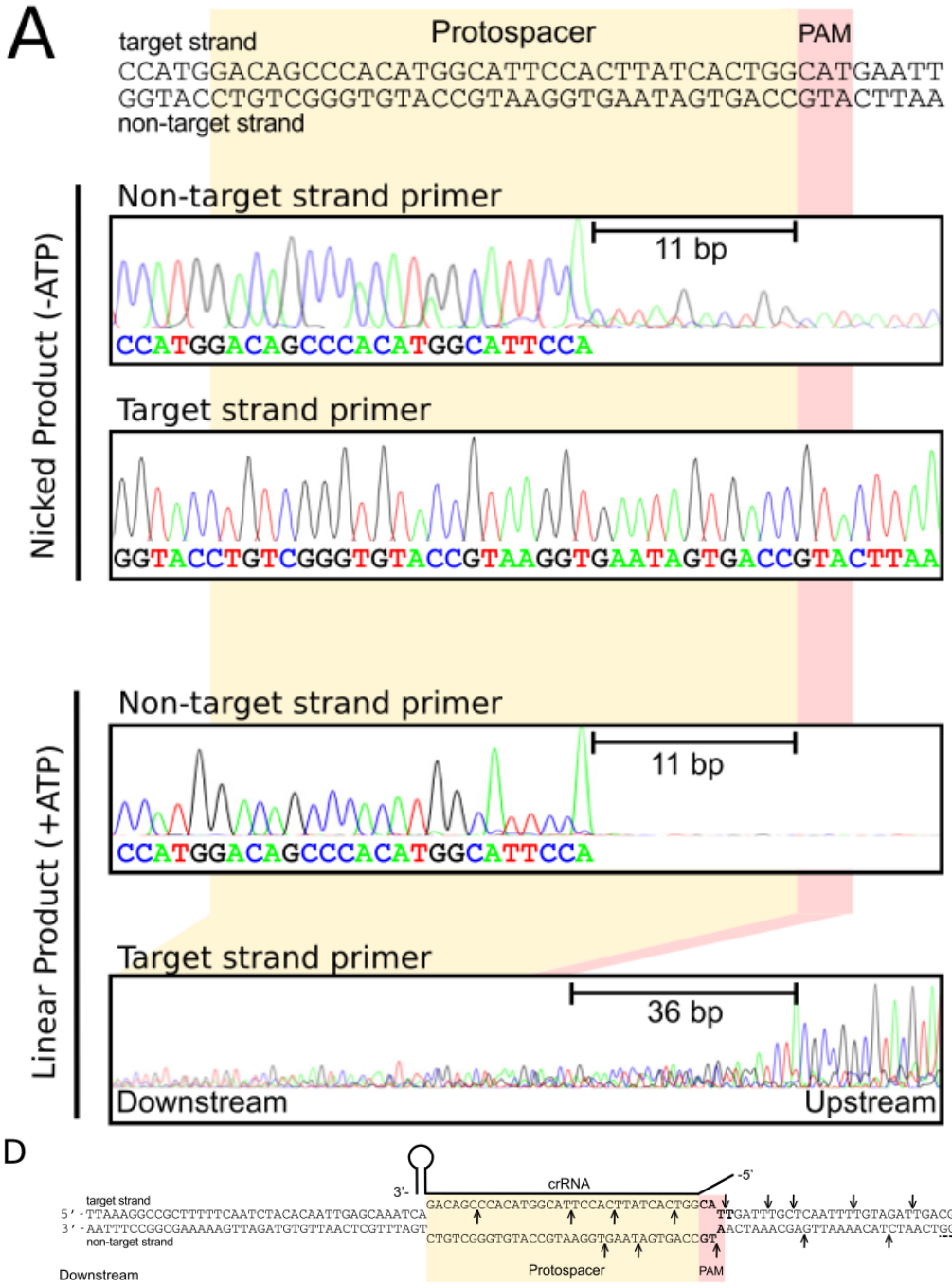


Figure 4.5. Cas3 cleaves linear DNA and proceeds unidirectionally. *A*, nuclease activity of Cas3 on plasmid target linearized with KpnI or ScaI. Reaction mixtures containing 20 nM Cascade, 2 nM linear plasmid, 10 mM Mg^{2+}, 10 μM Co^{2+}, and 2 mM ATP were incubated for the indicated time at 37 °C. *B*, nuclease activity of Cas3 on plasmid target linearized with ScaI. Except where indicated, the reaction conditions were as in *A*. Reactions were incubated at 37 °C for 60 min. *C*, comparison of the rate of cleavage of linear (ScaI) and negatively supercoiled plasmid. Reaction conditions were as in *A*, except that incubation times were as indicated. The position of negatively supercoiled (nSC), linear (L), and nicked or open circle (OC) DNA is indicated. In all panels, *M* denotes the marker lane, and *C* denotes a control reaction without Cas3.


A
target strand
Protospacer
PAM
CCATGGACAGCCCACATGGCATTCCACTTATCACTGGCATGAATT
GGTACCTGTCGGGTGTACCGTAAGGTGAATAGTGACCGTACTTAA
non-target strand
Nicked Product (-ATP)
Non-target strand primer
11 bp
CCATGGACAGCCCACATGGCATTCCA
Target strand primer
GGTACCTGTCGGGTGTACCGTAAGGTGAATAGTGACCGTACTTAA
Linear Product (+ATP)
Non-target strand primer
11 bp
CCATGGACAGCCCACATGGCATTCCA
Target strand primer
36 bp
Downstream
Upstream
D
3'-
crRNA
-5'
target strand
5'-TTAAAGGCCGCTTTTTCAATCTACACAATTGAGCAAATCA
GACAGCCCACATGGCATTCCACTTATCACTGGCA
TTGATTTGCTCAATTTTGTAGATTGACGGAACGAGGGTAGA-3'
3'-AATTTCCGGCGAAAAAGTTAGATGTGTTAACTCGTTTAGT
AACTAAACGAGTTAAAACATCTAACTGCCTTGCTCCCATCT-5'
non-target strand
CTGTCGGGTGTACCGTAAGGTGAATAGTGACCGT
Protospacer
PAM
Downstream
Upstream

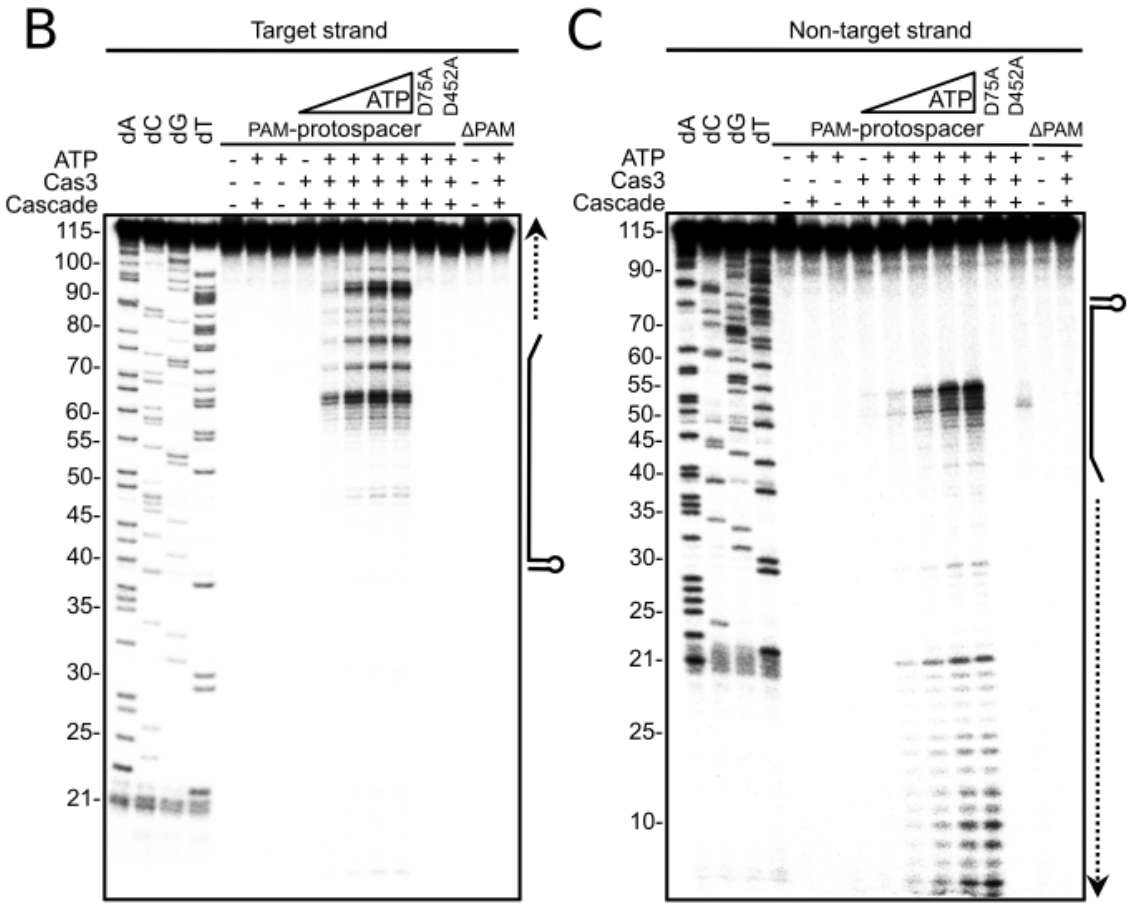


Figure 4.6. Mapping of the Cas3 cleavage sites. *A*, sequencing of the target and non-target strand of the nicked plasmid (no ATP) or the linearized plasmid (10 μM ATP). *B*, cleavage of a synthetic dsDNA labeled at the 5'-end of the target strand. The position of the crRNA is marked at the side of the gel. Sequencing lanes are marked *dA*, *dC*, *dG*, and *dT*. Δ*PAM*, DNA target containing a protospacer sequence but not a PAM. *C*, same as *B* but labeled at the 5'-end of the non-target strand. For both *B* and *C*, reactions containing 40 nM Cascade, 100 nM Cas3, 10 mM Mg^{2+}, 10 μM Co^{2+}, and 2 mM ATP were incubated for 30 min at 37 °C, unless otherwise indicated. *D*, schematic of the R-loop formed between the DNA target and the crRNA. Arrows or a dotted line indicate the sites of cleavage by Cas3.

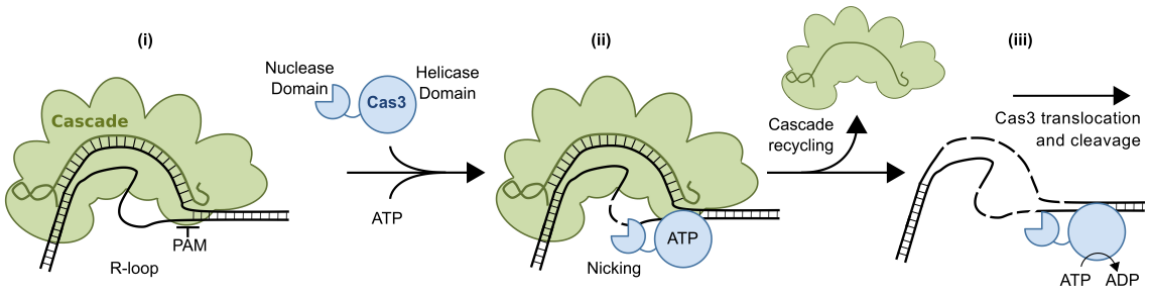


Figure 4.7. Schematic representation of Cascade-mediated DNA target degradation by Cas3. Binding of Cascade (*green*) to DNA target displaces the non-target strand of the protospacer. The displaced strand then serves as a binding platform for the recruitment of Cas3 (*blue*) (*i*). Once bound, Cas3 nicks the non-target strand in a reaction that is stimulated by the presence of ATP but does not require ATP hydrolysis (*ii*). ssDNA binding stimulates the ATPase and helicase activity of Cas3, which subsequently translocates in the 3'–5' direction on the non-target strand, unwinding the DNA target. Unwinding provides the ssDNA substrate for the nuclease domain and probably releases Cascade from the protospacer. Thus, the combined actions of the helicase and nuclease domains of Cas3 degrade the DNA target in a unidirectional manner (*iii*).

Chapter 5

Crystal structure of the type I Cascade complex from *Escherichia coli* bound to its target DNA.

Introduction

We previously determined the crystal structure of the *T. thermophilus* CasA subunit of Cascade. We next wanted to investigate how Cascade, as a whole, is able to efficiently bind to target DNA. Several other groups succeeded in determining individual high-resolution structures of the CasB, CasC, and CasE homologs (Agari et al., 2008; Sashital et al., 2011; Lintner et al., 2011), and cryo-electron microscopy structures (~9 Å) of Cascade from *E. coli* showed the overall organization of the Cascade subunits (Wiedenheft et al., 2011). Still, how these subunits interact with the crRNA and the target DNA was unclear.

Several Cascade-like ribonucleoprotein complexes have been reported in all of the other type I subtypes, as well as in type III CRISPR system (Lintner et al., 2011; Nam et al, 2012; Rouillon et al, 2013; Staals et al, 2013). As such, we pursued the crystal structure of one of these complexes given its broad application. To investigate the mechanism of target binding in the CRISPR interference step by such large complexes, we continued our structural studies on the *E. coli* Cascade complex by means of X-ray crystallography. We had a short but successful collaboration with Irimpan I. Mathews at the Stanford Synchrotron Radiation Lightsource (SSRL), who collected all of the X-ray

diffraction data from the SeMet Cascade crystals. The results presented here in this chapter are in the process of being submitted for publication.

Results

Structure determination

Cascade is a large 405-kDa-ribonucleoprotein complex that is able to scan for foreign DNA sequences based on complementarity to its crRNA. To determine the structural basis of target DNA recognition in the interference step, we crystallized the whole type-IE Cascade complex from *E. coli* bound to its target DNA. Crystals were grown by vapor diffusion with PEG 8,000 as the precipitant (Figure 5.1). The complex crystallized in space group $P3_121$ ($a = b = 225.321$ Å, $c = 293.208$ Å, $\alpha = \beta = 90°$, and $\gamma = 120°$), and consisted of stoichiometric amounts of the CasA, CasB, CasC, CasD, and CasE protein subunits as seen in the SDS-PAGE gel (Figure 5.1B). The nucleic acid extracted from the crystals consisted of both the crRNA and target DNA (Figure 5.1C). Although a partially complementary non-target strand was also used during crystallization, the strand did not crystallize with the complex, and was most likely unwound by Cascade during Cascade-DNA complex preparation (Figure 5.1C).

Initial crystals of the complex only diffracted to about ~9 Å, but the resolution was significantly improved to ~ 3.6 Å (Table 5.1) through step-wise stabilization of the crystals as described in the methods section and elaborated in Appendix B. The diffraction limit was further extended by soaking the wild-type crystals with a W_3-cluster, making it possible for the structure to be determined to a resolution of 3.03 Å (Table 5.1).

The structure was determined by single-wavelength anomalous dispersion (SAD) method, using thiomerosal-soaked crystals and SeMet-labeled protein crystals. The crystallographic asymmetric unit contained one Cascade-DNA complex. The structure was refined at 3.03 Å resolution to an R_{work} of 22.71% and an R_{free} of 27.37 (Table 5.1). A representative section of the unbiased F_o-F_c electron density around a section of crRNA-DNA hybrid is shown in Figure 5.2A.

Overall Structure of Cascade-DNA complex

The general organization of the Cascade subunits was previously revealed in two cryo-electron microscopy (cryo-EM) structures, with (~9 Å)- and without (~8 Å)- a complementary RNA strand (Wiedenheft et al. 2011). While this manuscript was in preparation, cryo-EM structure of Cascade bound to a 75-nucleotide dsDNA to ~ 9 Å resolution was reported (Hochstrasser et al, 2014). The crystal structure presented here shows the structure of the *E. coli* Cascade complex bound to a complementary DNA strand (Figures 5.2B and 5.2C). The structure consists of all the subunits of Cascade and the target DNA. It has the expected seahorse shape with CasE and CasA at its head and tail respectively. The crRNA-DNA hybrid forms the core of the structure, and the six subunits of CasC form a filament around the heteroduplex in a right-handed helical axis (Figure 5.2C). The Cascade-DNA structure presented here aligned best to this cryo-EM structure of the Cascade-dsDNA complex upon rigid-body docking, suggesting that the crystal structure closely represents in-solution conformation of the Cascade-dsDNA complex (Figure 5.3B). Unlike in the cryo-EM structures, portions of the $CasC_6$, CasE, and crRNA subunits are disordered in the Cascade-DNA crystal structure. The head

region of the Cascade complex in general has higher B-factors, suggesting greater structural flexibility in this region.

The crRNA-DNA hybrid forms a distorted, arched-ladder

The crRNA-DNA hybrid is at the core of the presented structure. As expected, the 5'-end of the crRNA forms a hook, and the repeat sequence at the 3'-end forms a stem-loop (Figure 5.4A). In the crystal structure, the entirety of the crRNA in the spacer region is present, with some nucleotides in the stem-loop being disordered. The target strand used for crystallization has a protospacer sequence flanked by the PAM on its 5' end (5'-CAT-3') and 6 random nucleotides on its 3' end. However, only the sequence between positions 1 and 33 could be modeled, as most of the flanking sequences, including the PAM, are disordered in the structure. Additional non-continuous density extends beyond position 33 at the 5' end between the CasE and CasB1 subunits, and is likely the phosphate backbone of the disordered nucleotides. The 5'- and 3'- ends of the target strand (~102 Å apart end-to-end) roughly span the length of a B-DNA with an identical sequence (~ 107 Å).

RNA-DNA hybrids most often form A-form-like helices (Horton and Finzel, 1996). However, the heteroduplex presented here, between its 5'-handle and its 3'-repeat, is reminiscent of a discontinuous arched-ladder spiraling through half a turn of an extended right-handed helical axis (Figures 5.4A and 5.4B). At the 5'-end, the handle is bent such that it forms two loops (I and II) (Figure 5.4D). Closer examination reveals that the discontinuity in the ladder is due to disruptions in the Watson-Crick base-pairing at specific positions along the length of the protospacer (Figure 5.4E). The nucleotides at

positions 6, 12, 18, 24, and 30 of both the crRNA and DNA are flipped out of the helix in opposite directions (Figure 5.4C). This observation suggests that not all of the bases in the spacer of the crRNA are used during target binding. In addition, the -1 position of the crRNA is flipped out as well.

Between the flipped-out bases, the spacer and protospacer regions make five distinct right-handed, semi-helical, duplex regions (duplex I-V), each consisting of five Watson-Crick base-pairs (Figures 5.4C and 5.4E). DNA-RNA hybrids form A-form-like helices, with some B-form-like features often appearing in the DNA strand (Horton and Finzel, 1996). The duplexes in the structure are A-form-like, but further compacted due to numerous interactions with the protein subunits. The five duplex regions are structurally similar as they superimpose on each other with identical distortion at similar positions of the duplex. The phosphate backbones of the two strands in the duplexes are generally ~20-21 Å apart. However, in the case of the nucleotides at (positions 5, 12, 18, 24, and 30) or directly in front (positions 5, 11, 17, 23, and 29) of the flipped positions, the hybrid backbones are more constricted, and are only ~18.5 Å apart. Besides the five duplexes, Watson-Crick base-pairing also exists between bases at position 31 and 32.

CasC subunits distort the crRNA

CasC, with six copies, is the most prevalent subunit of the complex. The monomers form a filament around the nucleic acid core with $CasC_1$ (next to CasE) at the head and $CasC_6$ (next to CasA) at the tail of the complex (Figure 5.4C). In doing so, CasC subunits extend from the -5 to 32 positions of the crRNA. Such an organization of the CasC subunits is reminiscent of filaments formed by RecA around DNA during

homologous recombination (Chen et al., 2008). Although RecA can form extended filaments, the length of the crRNA defines CasC oligomerization in the case of Cascade. $CasC_1$ to $CasC_5$ interact with the spacer of the crRNA. It was previously thought that $CasC_6$ completely encapsulates the 5'-hook of the crRNA (Wiedenheft et al., 2011). However, in the structure presented, it only interacts with loop II of the 5' handle (Figure 5.4D).

The structure of the individual CasC subunits is reminiscent of a right hand with a distinct finger (59-180), palm (1-58, 181-189, and 224-363), and thumb (190-223) domain (Figure 5.5A). All the CasC subunits have the three domains. However, the finger domain of $CasC_6$ is highly disordered and could not be modeled owing to a lack of electron density for most of this domain (Figure 5.3). Investigation of crystal packing shows that the finger domains of the other CasC subunits are stable both in the presence ($CasC_3$-C_5) and absence ($CasC_1$-C_2) of crystal contacts. The lack of electron density in the crystal structure and poor features in the cryo-EM maps for the finger domain of only the $CasC_6$ subunit implies greater flexibility in this domain. Rigid-body fitting of the crystal structure into the Cascade-dsDNA cryo-EM map suggests that the helix corresponding to residues 136-145 in $CasC_5$ interacts with the dsDNA outside the protospacer. This domain consists of multiple Lys residues (K136, K137, K141, K144, and K145) that could interact with the phosphate backbone of the dsDNA and assist in DNA bending directly upstream of the protospacer. Corresponding residues in the $CasC_6$ finger domain are also likely to interact with the dsDNA based on its positioning in the Cascade-dsDNA cryo-EM structure.

Unlike the finger domains, the palm domains of all six CasC subunits (C_1 to C_6) follow a helical pattern around the central axis of Cascade. Each of the palm domains is comprised of a conserved RNA-recognition motif (RRM), and has a positively charged concave surface immediately above the thumb domain (Figure 5.5B). This is the main site of CasC-crRNA interaction, and is lined with conserved polar residues (N19, R20, K27, K45, R49, and Q42) that hydrogen bond extensively with the phosphate backbone, as well as the flipped-out bases (R46 and K50 of CasC) of the crRNA (Figure 5.6D). Each of the CasC subunits makes similar interactions. These observations suggest that the spacing of the CasC subunits dictates the position of the flipped bases of the crRNA.

An additional residue that is not part of the basic patch, but which also interacts with the crRNA, is M166. In the case of each of the CasC subunits, this residue partially intercalates between the 3rd and 4th crRNA bases of each of the five duplexes (Figures 5.6D and 5.4E), causing the bases to be further apart. As a result, their complementary DNA bases also have greater spacing between them. The phosphate backbone of the target DNA is slightly bent as well after the 4th base. Such Met-stacking interactions are often used to bend nucleic acid strands or stabilize unwound strands (Churchill et al., 2010; Firczuk et al., 2011; Chen et al., 2008). It is thus likely that M166 indirectly stabilizes the target DNA strand during unwinding by modulating the separation between the crRNA bases that are involved in base pairing with the target DNA. Although present, M166 in $CasC_6$ is not involved in such an interaction, possibly because the $CasC_6$ palm does not interact with the crRNA spacer. Similar Met stacking interactions are used by RecA to destabilize base-stacking interactions every 4th base to create

discrete 3-nucleotide segments to assist in 'conformational proofreading' during homology search (Chen et al., 2008).

The CasC thumb domains protrude through the crRNA-DNA hybrid

In addition to the crRNA-binding motif, almost all of the residues in the thumb domain and the concave side of the palm domain (below the basic patch) are highly conserved (Figure 5.5C). The importance of these residues is evident in the context of a CasC filament as they are involved in interacting with the neighboring CasC subunits (Figure 5.6A). Starting from $CasC_6$ (next to the 5'-handle), the thumb domain of each CasC protrudes towards the head of Cascade (CasE), and interacts with the palm domain of the neighboring CasC subunit. In the process, each of the thumb domains also passes through the distorted regions in the crRNA-DNA hybrid (Figure 5.6A). For example, the $CasC_5$ thumb goes through bend II, and the $CasC_4$ thumb goes through bend III. Both then interact with the palm domains of $CasC_4$ and $CasC_3$ respectively.

The thumb domain can extend through the distorted regions of the crRNA-DNA hybrid due to the many conserved aromatic residues (H213, W199, and F200), making stacking interactions with the bases at the end of the duplex. Additionally, L214 of each CasC thumb makes van der Waals interaction with the flipped-out DNA bases. This pattern only continues until $CasC_2$ since a similar path of the $CasC_1$ thumb would clash with the stem-loop of the crRNA. The D194 and S220 residues in the $CasC_1$ thumb are bent such that the residues in between form an extended loop that is rotated by ~100°, and interact with CasE instead (Figure 5.6B). Such a structural rearrangement of the

thumb domain at only one end of Cascade could serve as a start or stop signal for CasC polymerization along the crRNA during Cascade assembly.

In the process, CasC surrounds each of the flipped bases of the crRNA, and renders it incapable of target binding (Figure 5.6C). These extensive interactions result in a stable structure where CasC wraps around the crRNA, and likely prevents dissociation of the latter. This also suggests that only 5-base segments of the crRNA are available for target binding (Figure 5.6C). Consistent with this observation, the target-bound structure shows that the DNA bases opposite the flipped crRNA bases are also distorted and not involved in base pairing (Figure 5.4E). H213 in the CasC thumb domain replaces the position of the flipped-out bases instead to stabilize the duplexes (Figure 5.8B). The same duplex is capped by L214 on its opposite end.

We previously reconstituted Cascade-mediated target degradation by Cas3 *in vitro* (Mulepati et al, 2013). A similar assay was carried out with a target DNA with non-complementarity at positions of the flipped-out bases. Almost wild-type-like phenotype with the mutant plasmid confirms that the flipped-out bases in the crystal structure are not as important for base pairing during target binding (Figure 5.11). While this manuscript was in preparation, Fineran *et al.*, using *in vivo* targeting assay, showed higher tolerance for mutations at the flipped positions of the protospacer during CRISPR targeting. Hence, the bases at the flipped positions are not as important for specificity during DNA binding (Fineran et al., 2014).

CasA stabilizes the first crRNA-DNA duplex

CasA sits at the tail of the Cascade-DNA complex, and its structure closely resembles known structures of its homolog from *T. thermophilus* (Mulepati et al., 2012; Sashital et al., 2012). *E. coli* CasA also adopts a chair-like conformation where its N-terminal domain (NTD) forms the seat and its C-terminal domain (CTD) forms the backrest (Figure 5.8A). Its NTD sits below the expected position of the PAM, and its CTD roughly spans the seed region of the target DNA (Figure 5.2B).

CasA alone does not bind nucleic acid, but is required for the binding of Cascade to dsDNA (Mulepati et al., 2012; Sashital et al., 2012). A conserved loop (L1, residues 130-143) in CasA is essential for this binding, as F129 in this loop is thought to intercalate with the PAM and cause the initial destabilization of the target dsDNA (Sashital et al., 2012). Although present in the crystal, both the PAM sequence and the L1 loop are disordered. The L1 loop is also disordered in every CasA crystal structure currently available in the Protein Data Bank (4AN8, 4EJ3, 4F3E, 4H3T). Bases in the crRNA-DNA hybrid are flipped at positions 6, 12, 18, 24, and 30. In addition, the crRNA base at position -1 is flipped as well. Based on the spacing of the flipped bases in the hybrid and the ability of Phe to make stacking interactions, it is likely that, during PAM interrogation, the -1 DNA base is also flipped and stabilized by base-stacking interactions made by the F129 in L1 (Figure 5.4E). The surprising lack of the expected PAM-L1 electron density might be the result of at least two possibilities: the PAM-L1 interaction (i) is transient and not essential after spacer-protospacer annealing, or (ii) occurs only in the context of a dsDNA target (the target in the crystal structure is single-stranded). Differentiation between these possibilities will require additional investigation. Although, the recent demonstration that base pairing of the PAM region is required for target

degradation by Cas3 suggests that a transient PAM-L1 interaction is more likely (Hochstrasser et al., 2014).

The CasA CTD consists of a four-helix bundle (Fig. 7A), and cryo-EM and docking studies suggest that loop L2 (residues 405-411) (Figure 5.8B) becomes ordered upon target binding (*11, 23, 24*). Consistent with these observations, the L2 residues are ordered in the complex structure, and K409 in this loop makes direct hydrogen-bond interactions with the phosphate backbone (Figure 5.8B). The L2 loop in the crystal structure of CasA from *Acidimicrobium ferrooxidans* (4H3T) is ordered even in the absence of DNA, but its superposition on *E. coli* CasA (within Cascade) suggests that its L2 loop faces away from the crRNA (and the target DNA). Furthermore, the L3 loop (residues 469-474) between helix III and IV makes hydrogen bonds with the flipped base at position 6 through H472 and K474 (Fig. 5.8B). Despite being within the seed region (1-8), base pairing at position 6 was previously shown to be irrelevant for target binding, and hence, CRISPR interference (Jore et al., 2011; Semenova et al., 2011; Mulepati and Bailey, 2013). The fact that the base at this position is flipped out of the helix, making it unavailable for base pairing to render specificity, explains these earlier observations.

The cryo-EM structures suggested that CasA goes through a rotation movement upon binding to the target sequence. Modeling of the CasA crystal structure into the apo-Cascade cryo-EM map suggests that the NTD and CTD of CasA twist in opposite directions with respect to CasD (Figure 5.8A). The relative movement is such that the position of CasA in the apo-Cascade complex creates space for the loading of dsDNA near L1 of CasA. Furthermore, the CasA-CTD moves towards the NTD upon target binding. This suggests that the CTD loops stabilize the flipped-out base at position 6

during target DNA unwinding, likely enabling base-pairing of the target strand with the crRNA to form duplex I (duplex I in Figure 5.5E). Since five nucleotides of the target strand base pair with the crRNA in duplex I, it also suggests that CasA stabilizes half a turn of the dsDNA while going through its conformational change.

CasB propagates melting of target DNA

CasB has been shown to be essential for target interference (Brouns et al., 2008; Wiedenheft et al., 2011). CasB dimerizes on its own and can bind to both DNA and RNA (Agari et al., 2008; Nam et al., 2012). In the structure presented, $CasB_1$ and $CasB_2$ form a dimer as expected, which spans between the head (CasE) and tail (CasA) of the Cascade complex (Fig. 5.9A). In the conformation trapped in the crystal structure, besides the target DNA, CasB interacts extensively with CasA, and the palm and thumb domains of $CasC_2$-C_5. Surprisingly, CasB subunits make no significant contact with CasE, $CasC_1$, $CasC_6$, CasD, and the crRNA (Figure 5.9B). Hence, the head (CasE) and tail (CasA) of Cascade are not connected in the DNA-bound crystal structure. Each of the CasB subunits spans about 12 nucleotides of the protospacer, and CasB as a whole covers positions 9 through 30.

The CasB subunits make specific contacts with the flipped-out bases of the target DNA (Figure 5.9A). The flipped bases at positions 12 and 24 are held between H123 of CasB and L214 of CasC. Similarly, the flipped bases at positions 18 and 32 are held between G20 of CasB and L214 of CasC. $CasB_2$ interacts with the flipped bases at positions 12 and 18, and $CasB_1$ interacts with the flipped bases at positions 24 and 32. At position 24 (and 12), conserved residues in $CasB_1$ hydrogen bond with the phosphate

backbone (N98) and the base (R119), and base-stack with the flipped-out base (H123) (Figure 5.9C). At position 30 (and position 18), conserved charged residues base pair with the phosphate backbone (R26) and the flipped-out base (R27) (Fig. 5.9D).

Comparison of the CasB subunits in the apo- and DNA-bound structures suggest that the subunits move closer to the PAM and the crRNA, constricting the space next to the crRNA. In doing so, CasB likely uses its contacts at positions 12, 18, 24, and 32 to move the target strand closer to the crRNA, and stabilize the melted DNA during unwinding. Furthermore, the CasB subunits (like in CasA) are arranged such that each of the subunits can stabilize half a turn of a B-form DNA.

Implications on Cascade assembly

One of the major questions that arises from the intricate interactions revealed by the crystal structure is how a large complex like Cascade assembles. Is there an order in which the Cascade subunits are added onto the complex? The structure revealed that CasD caps the 5' handle of the crRNA (Figures 5.7A and 5.7B). Along with CasC, CasD is the most conserved subunit within the Cascade complex. It has an N-terminal RRM motif, a C-terminal beta-sheet domain, and can also be classified into a palm and a thumb domain (Figure 5.7A). The anterior part of the thumb, corresponding to residues 89-104, is disordered in the structure. Still, this truncated thumb extends to the crRNA-DNA hybrid close to the expected position of the PAM (Figure 5.7B), and like $CasC_2$-C_6, protrudes between the distorted -1 position between the crRNA and the target DNA. In doing so, the CasD thumb assumes a similar position to the one occupied by the thumb domains of $CasC_6$ to $CasC_2$ (Figure 5.6C). This suggests that CasD thumb limits CasC

polymerization at $CasC_6$ (starting from $CasC_1$). Alternatively, the CasD thumb could also signal for CasC polymerization to proceed from $CasC_6$ to $CasC_1$.

The crystal structure shows that CasD interacts exclusively with the 5'-hook (loop II) of the crRNA in a target-bound conformation through the conserved basic patch in its palm domain (Figure 5.7B). This explains previous observations made in the case of CasD homolog (Cas5d) in the type I-C CRISPR system of *Bacillus Halodurans,* where deletion and point mutations of nucleotides in loop II are most detrimental to assembly of its Cascade-like complex (Nam et al., 2012). Together, these observations suggest that CasD caps the 5' end of the crRNA with its palm domain, and that the position of its thumb domain either allows for or stops CasC polymerization on the crRNA.

The crystal structure of a type-IA CasC homolog (Csa2 with ~11% identity) from *S. solfataricus* aligns to the *E. coli* CasC structure with an RMSD of ~3.7 (Lintner et al., 2011). Csa2 exists as a monomer in the absence of crRNA, but polymerizes in the presence of crRNA and its CasD homolog. Csa2 also has a palm and a shorter finger domain, but its thumb domain (residues 163-177) is disordered in all four Csa2 monomers present in its crystallographic asymmetric unit. This suggests that the thumb domain is structurally flexible and becomes ordered in specific orientations only in the presence of the crRNA template and the other Cascade subunits. This observation is consistent with the two different orientations of the thumb domains in $CasC_6$ and $CasC_1$-$CasC_5$ in the structure presented here. Residues corresponding to the CasC1 thumb domain form a bent loop (which is partially disordered). The similarities in the structure of CasC and Csa2 also reinforce the idea that type-I Cascade-like complexes have similar

architectures and likely similar orders of subunit assembly. Additional experiments need to be carried out to delineate the order of subunit addition during Cascade assembly.

Implications on DNA unwinding by Cascade

The crystal structure shows that the crRNA-DNA hybrid forms a discontinuous, arched, ladder-like structure. CasD and CasE cap the crRNA on its 5'- and 3'- ends respectively. Bases in both of the strands are flipped at regular 6-nucleotide intervals. The CasC subunits stabilize the flipped bases of the crRNA. The thumb domains of these subunits weave around the crRNA at the flipped position, suggesting that the flipped bases do not base pair during target binding. The flipped bases on the target strand are stabilized by specific contacts made by the CasA and CasB subunits. The short duplex segments between the flipped bases are narrow in diameter as a result of compaction by the protein subunits, but have A-form-like features at the nucleotide level.

Although the crystal structure shows only the target strand, Cascade melts dsDNA, and based on protection assays, interacts extensively with both the complementary and the non-complementary strands (Jore et al., 2011). Positions 1-14 are exposed as ssDNA while the rest is protected by Cascade. This raises the question of where in Cascade the non-complementary strand would bind. The crystal structure suggests that Cascade has a deep, electropositive groove between the CasB and CasC subunits, on the opposite side of CasB with respect to the complementary strand (Figure 5.12). This groove is likely involved in the stabilization of the non-complementary strand as it is lined with conserved Lys residues of CasB and CasC subunits, and based on the

length, would extend up to the end of $CasB_2$ before being exposed as a single-stranded region close to CasA.

Crystal structures of type I-C CasD homologs from *B. halodurans, Xanthomonas oryzae,* and *Streptococcus pyogens* have been previously reported (Nam et al., 2012; Koo et al., 2013). The structures of the first two were overlaid on top of *E. coli* CasD in Fig. 5.7C. Of these homologs, *E. coli* CasD is most closely related to Cas5d from *X. oryzae.* Structural alignment suggests that the disordered region of the CasD thumb forms a flexible helix-loop-helix motif in *X. oryzae* (Koo et al., 2012). Furthermore, the two monomers in the asymmetric unit of Cas5d have their thumb domains in different orientations, suggesting that this domain is very flexible. This motif is ordered in the *X. oryzae* structure due to crystal contacts that are not present in the case of *E. coli* Cascade crystals. Thus, we would predict that *E. coli* CasD to have a flexible thumb as well. Since the *X. oryzae* Cas5d has been reported to have non-specific DNA-binding activity, it is likely that CasD might also be involved during target DNA binding/unwinding.

Based on previous reports and our crystal structure, we propose a model as shown in Figure 5.10. Upon PAM recognition by the L1 loop of CasA-NTD (Sashital et al., 2012), its F129 residue likely intercalates and stabilizes the -1 PAM base to initiate DNA melting. This interaction is coupled to a conformational change in CasA (Wiedenheft et al., 2011) that brings the L2 and L3 loops in its CTD closer to position 6 of the target DNA. The conserved residues in these loops stabilize the flipped base at position 6 of the target, and the movement of the CTD domain (towards the PAM) likely facilitates melting of the target DNA duplex.

CasA conformational change also results in the concerted movement of the CasB subunits in the same direction. The conserved residues in the CasB subunits make specific interactions at positions 12, 18, 24, and 30. These contacts are likely used by CasB to further stabilize the melted dsDNA and in turn, position the bases next to the crRNA for Watson-Crick interactions. During all this conformational change, CasC-crRNA remains relatively unchanged, based on the superposition of CasC subunits before and after target binding (Figure 5.3). Hence, the CasC-crRNA acts as a template that is used by the other Cascade subunits during the unwinding of double-stranded target DNA. Based on the recent biochemical observations made in the case of the type-III Cascade-like Cmr complex in *Thermus thermophilus*, it is possible that Cascade and other Cascade-like complexes melt double-stranded DNA one half-turn of a B-form DNA at a time (Staals et al., 2013). Unlike Cascade, the *T. thermophilus* Cmr complex also exhibits nuclease activity, and upon binding, cleaves its target at six-nucleotide intervals starting at its 3' end (next to PAM in the protospacer). Given the structural similarity between Cascade-like complexes in the Type-I and Type-III systems, it is likely that these complexes use similar mechanisms for DNA/RNA targeting in the interference step. However, the possibility that target melting and cleavage in the type-III Cmr complex are separate events cannot be ignored.

Overall, the Cascade-DNA crystal structure presents a detailed architecture of the crRNA-target DNA hybrid, the structure of the individual subunits and their interactions in the complex, and how Cascade-like complexes bind to their targets during the CRISPR interference stage. The structure also provides invaluable insight for future experiments.

Materials and Methods

Expression and Purification of Wild-type and SeMet-labeled Cascade

Wild-type Cascade was expressed and purified as previously described (Mulepati et al., 2012). For seleno-methionine (SeMet) incorporation into Cascade, plasmids were transformed into the Rosetta 2 (DE3) strain (Novagen) of *E. coli* and and expressed in EZ-rich defined media (Neidhardt et al., 1974), where Met was replaced with SeMet. Cells were grown at 37 °C to OD_{600} of ~0.3 before adding 50 mg/L of SeMet and lowering the temperature to 20 °C. Cells were further allowed to grow to OD_{600} of ~0.4 before protein expression was induced with 0.5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG). SeMet-Cascade was purified with an identical procedure to wild-type Cascade. Cascade complexes were concentrated in a buffer consisting of 20 mM Tris-HCl at pH 8.0, 200 mM NaCl and 1 mM TCEP (*tris*(2-carboxyethyl)phosphine) to about 30 μM, flash frozen in liquid N_2 and stored at -80 °C until further use.

Cascade-DNA Complex Formation

To make target dsDNA, two complementary DNA strands with the sequences 5'-AATCAGACAGCCCACATGGCATTCCACTTATCACTGGCAT-3' or 5'-AATTGAG CAAATCAGACAGCCCACATGGCATTCCACTTATCACTGGCAT-3' were separately annealed to a partially non-complementary strand with the sequence 5'-GCCA TGTGGGCTGTCTTAACTC GTTTAGT-3' (Sigma). Double-stranded target DNA was prepared by slow-annealing 230 μM of the non-complementary strand with 200 μM of

the complementary strand in a buffer consisting of 20 mM Tris-HCl pH 7.5, 100 mM NaCl and 0.5 mM of EDTA.

Cascade and target dsDNA were incubated at 37 °C for 30 min at concentrations of 20 μM and 30 μM respectively, in a buffer containing 20 mM Tris-HCl at pH 8.0, 200 mM NaCl, and 1 mM TCEP. The complex solution was spun down and placed on ice for approximately 5 min.

Crystallization of Cascade-DNA Complex

Cascade-DNA crystals were obtained by sitting-drop vapor diffusion method. Initial crystals were obtained by mixing 2 μL of Cascade-DNA complex with 1 μL of a reservoir solution consisting of 0.1 M sodium cacodylate at pH 5.0, 0.1 M calcium acetate, and 9-11% PEG 8,000. Crystals of different sizes grew over 1-7 days at 20 °C. These crystals were subsequently stabilized in a reservoir solution consisting of 8.5 % PEG 8,000, and then used to prepare a seed solution. Larger crystals that grew to a size of approximately 500 μm x 300 μm x 300 μm were obtained by mixing 2 μL of cascade-DNA complex with 1 μL of freshly prepared seed solution. Crystals were allowed to grow for 7-10 days before being harvested. Crystals were gradually cryo-protected in 5% steps into the reservoir solution supplemented with 4 mM TCEP and 5 % each of glycerol, sucrose, PEG 400, and ethylene glycol. To obtain heavy-atom derivates, a few crumbs of different heavy atoms (W3-cluster: $\{[(W_3O_2(CH_3COO)_6(H_2O)_3]^{2+}(CF_3SO_3)_2\}$ or thimerosal) were transferred into the cryo solution and allowed to soak into the crystal over 24 h at 20 °C. Stabilized crystals were flash-frozen in liquid N_2.

Data collection and Structure Determination

X-ray diffraction data were collected at the Stanford Synchrotron Radiation Lightsource (SSRL) on beamlines 7-1, 11-1 and 12-2. Data were processed with XDS (Kabasch, 2010). The Cascade-DNA complex crystallized in space group $P3_121$ with unit cell dimensions as listed in table 5.1. Data from thiomerosal-soaked crystals were used for single-wavelength anomalous dispersion (SAD) phasing. SHELX (Sheldrick, 2008) was used to find the positions of Hg sites in the thiomerosal-soaked crystals. Phases were calculated from the thiomerosal-soaked crystals using SOLVE, and improved by solvent flattening in RESOLVE (Terwilliger, 2004). During solvent flattening in RESOLVE, non-crystallographic symmetry in between the 6 CasC subunits and 2 CasB subunits was averaged to get improved maps. This initial map was used to calculate a difference Fourier map to find the Se positions in the SeMet derivative crystals. Of the 113 possible Se sites, 106 ordered Se sites were located. While the best native crystals diffracted to only ~3.6 Å, native crystals soaked with the W-cluster extended diffraction to 3.03 Å (Table 5.1). Phases from the Hg and Se derivates were combined using Sigma (Read, 1986) and used along with the structure factor (F) from the W-cluster-soaked crystals to calculate the final experimental map. The unit cell has a high solvent content of ~70%. The asymmetric unit consists of a single cascade complex bound to the complementary DNA strand.

An atomic model of the cascade-DNA complex was built in COOT (Emsley and Cowtan, 2004). Models of *E. coli* CasA, CasB, and CasE-RNA were generated using the I-TASSER server (Roy et al., 2010) from homologs in *T. thermophilus* with known structures. These models were used as initial coordinates with changes made in COOT as

necessary. Once most of the complex was built, the model was refined with PHENIX (Afonine et al., 2010). The figure panels were made with either PyMOL (Delano, 2010) or Chimera (Goddard et al., 2007).

Cryo-electron Microscopy Map Fitting

The complete structure of Cascade-DNA complex was aligned to Cascade-RNA cryo-electron microscopy structure by means of rigid-body docking in Chimera. The individual subunits from the crystal structure were also docked into the Cascade-only cryo-electron microscopy structure.

In vitro Reconstitution Assay

Cleavage assays were performed as described previously (Mulepati et al., 2013). The mutant plasmid (mismatches underlined) consists of a protospacer with the sequence: 5'- GA<u>G</u>AGCCC<u>T</u>CATGG<u>G</u>ATTCC<u>T</u>CTTAT<u>G</u>ACTGG**CAT** – 3'.

Table 5.1: *Escherichia coli* Cascade-DNA crystal structure—Data collection and processing statistics

Data collection	**Native**	**W3 soak**	**Thimerosal**	**Selenium**
X-ray beamline	SSRL 7.1	SSRL 11.1	SSRL 11.1	SSRL 12.2
Wavelength (Å)	0.9999	1.2131	0.980112	0.97938
Unit cell				
$a = b$, c (Å)	225.32, 293.20	223.79, 290.61	224.00, 290.44	225.515, 291.295
$\alpha = \beta$, γ (°)	90, 120	90, 120	90, 120	90, 120
Resolution (Å)[a]	3.62 (3.68-3.62)	3.03 (3.05-3.03)	3.90 (3.98-3.90)	3.51 (3.57-3.51)
R_{sym}[a, b]	0.218 (2.362)	0.181 (2.959)	0.296 (3.237)	0.339 (1.838)
R_{pim}[a, c]	0.074 (0.921)	0.058 (0.948)	0.092 (1.033)	0.138 (0.754)
I/σ[a]	10.4 (1.3)	12.3 (1.0)	9.7 (1.0)	6.7 (1.1)
Redundancy[a]	9.8 (7.2)	11.0 (10.4)	11.2 (10.4)	6.6 (6.7)
Completeness (%)[a]	99.5 (91.8)	99.8 (96.4)	99.7 (96.3)	96.7 (92.9)
Wilson B factor		84.33		
Mean figure of merit		0.57		
Heavy atom sites		1	14	106
Refinement				
Resolution (Å)		39.46-3.03		
R_{work}[d]		22.71		
R_{free}[d]		27.37		
r.m.s.d. bond (Å)[e]		0.012		
r.m.s.d. angle		1.774		
No. of atoms		26670		
Protein		24745		
Nucleic acid		1925		
B-factors		46.7		
Ramachandran plot				
Most favored (%)		85.36		
Allowed (%)		8.92		
Outliers (%)		5.72		

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

[b] R_{sym} is $\Sigma|I_o - I|/\Sigma I_o$, where I_o is the intensity of an individual reflection, and I is the mean intensity for multiple recorded reflections.

[c] R_{pim} is $\Sigma(1/(N-1))^{\wedge 1/2}(|I_o - I|))/\Sigma I_o$

[d] R_{work} is $||F_o - F_c||/F_o$, where F_o is an observed amplitude, and F_c is the calculated amplitude; R_{free} is the same statistic calculated over a subset of the data that has not been used for refinement.

[e] r.m.s.d., root mean square deviation.

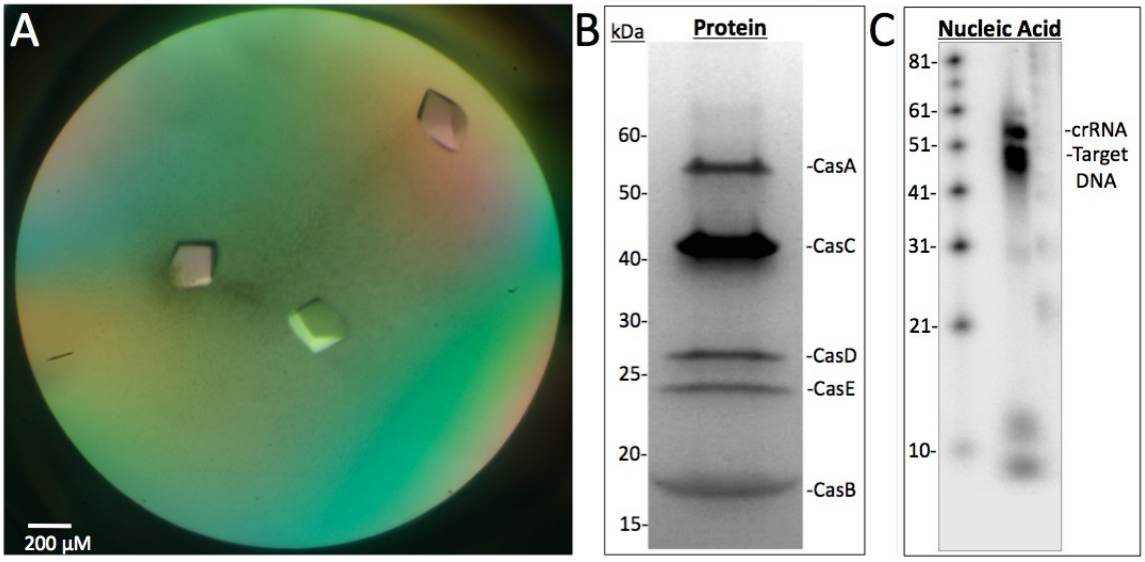


Figure 5.1. Crystals of the Cascade-DNA complex and its content. *A,* Crystals of the Cascade-DNA complex obtained from microseeding. *B,* Protein content of the Cascade-DNA complex. Large crystals (4-5) were looped and dissolved in H_2O. The content was run on an SDS-PAGE gel and then stained with Coomassie blue stain. *C,* Nucleic acid content of the Cascade-DNA complex. Crystals were dissolved in H_2O and deproteinated with phenol extraction. The nucleic acid extracted was ethanol precipitated and 5'-labeled with γ-[^{32}P] ATP (PerkinElmer Life Sciences) using T4 polynucleotide kinase (New England Biolabs). The sample was separated on a denaturing-urea gel and visualized by autoradiography.

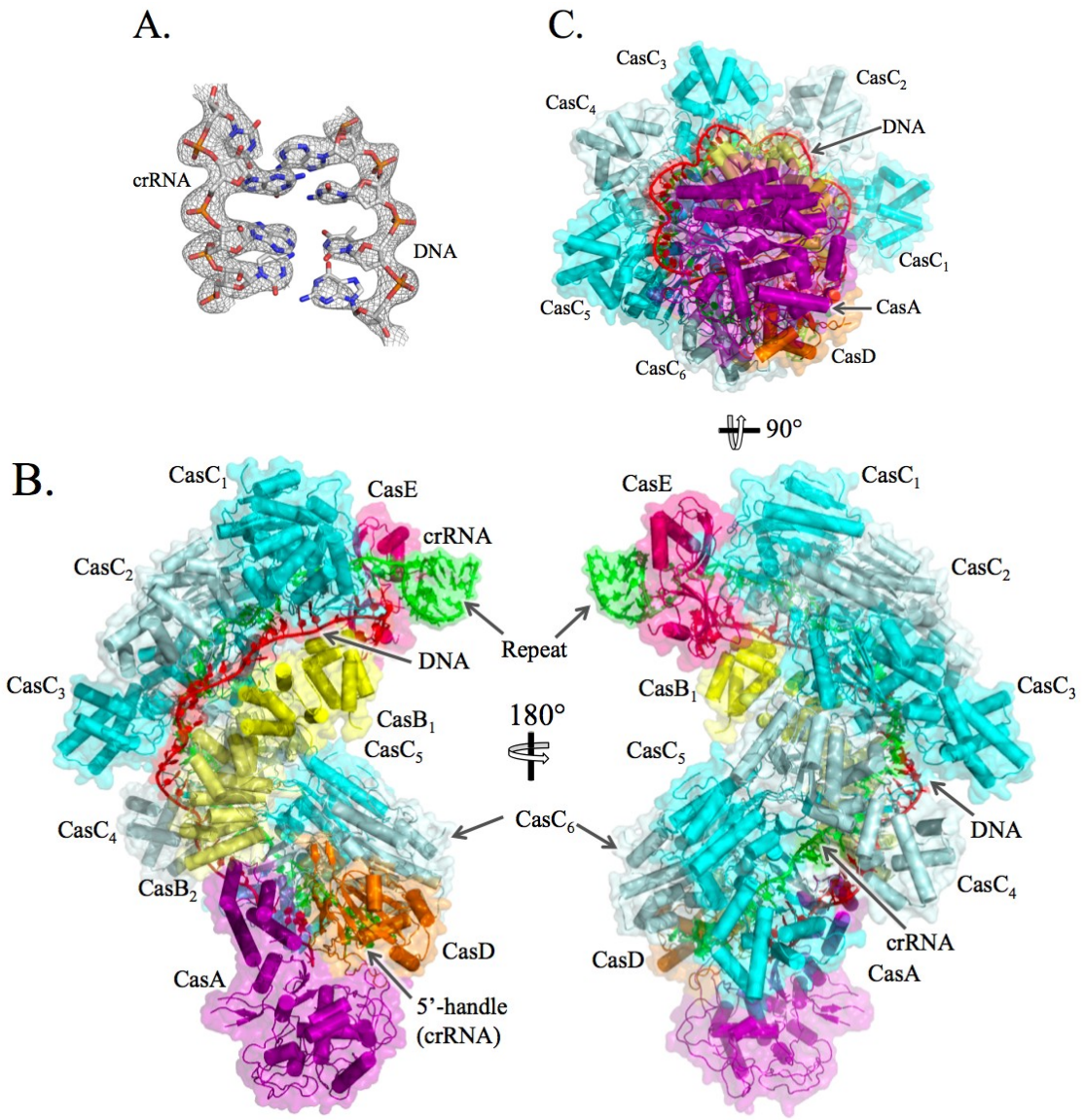


Figure 5.2. Crystal structure of the *E. coli* Cascade bound to target DNA. *A,* Unbiased F_o – F_c electron density map contoured at 3σ. The nucleotides, which are represented as sticks, were omitted from the map calculation. *B,* Overall structure of Cascade-DNA complex in two orthogonal orientations. Semi-transparent surface representation is superimposed on the model of the complex. *C,* Right-handed helical arrangement of the CasC hexamer around the crRNA-DNA heteroduplex. The structure is flipped such that the tail (CasA) of the complex is on the surface of the plane and its head (CasE) pointing into the plane.

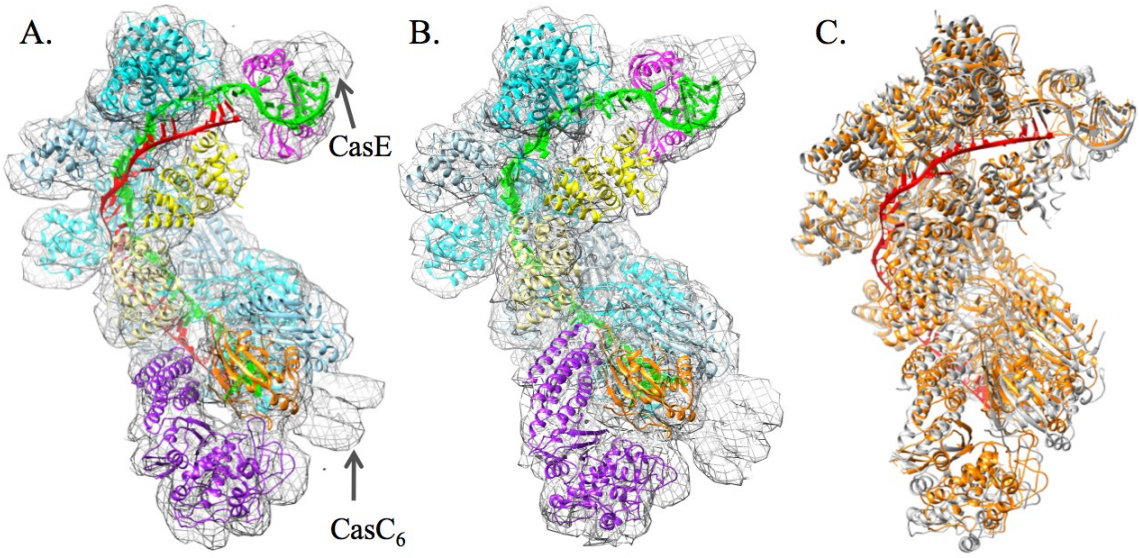


Figure 5.3. Comparison of the Cascade-DNA and apo-Cascade structures. *A,* Crystal structure of Cascade-DNA complex rigid-body docked into the cryo-electron microscopy map. *B,* Model of the apo-Cascade with crystal structures of the individual subunits from *A* rigid-body docked individually into the apo-Cascade cryo-electron microscopy map. *C,* Superimposed structures of the Cascade-DNA complex and the apo-Cascade. In *A* and *B*, the cryo-electron microscopy maps are represented as grey mesh and the subunits are colored as in Figure 5.2. In *C*, all the apo-Cascade subunits are colored in grey and the Cascade-DNA subunits are colored in orange, except for the target DNA colored in red.

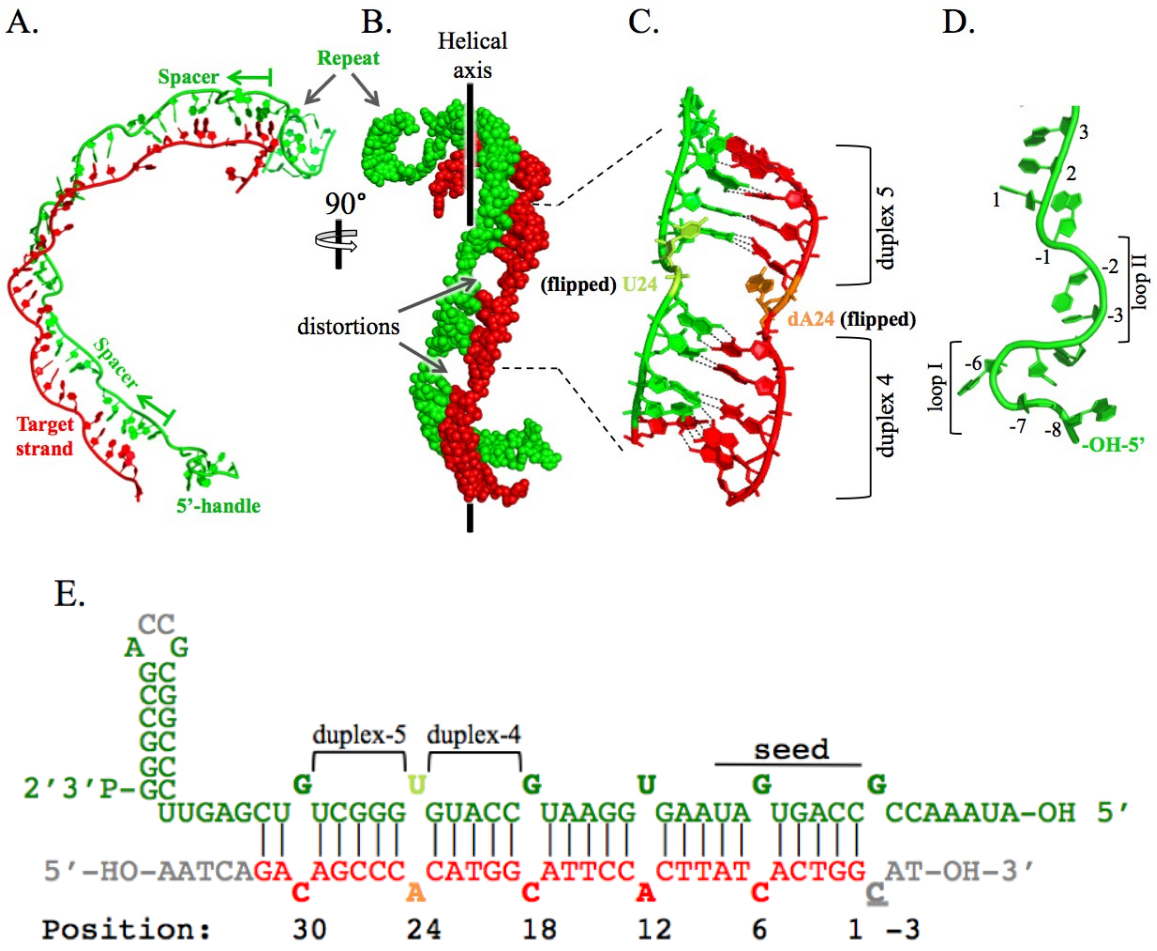


Figure 5.4. The crRNA-DNA hybrid forms an unusual arched-ladder structure. *A,* Cartoon representation of the interaction of the crRNA-DNA interaction over the length of the protospacer. *B,* View of the crRNA-DNA hybrid represented as spheres, rotated by ~90° around the helical axis of Cascade with respect to *A*. *C,* Structure of the semi-helical duplex regions interspaced by flipped-out bases in both the crRNA (U24 in limegreen) and the target DNA (dA24 in orange). *D*, Ordered loops formed by the 5' handle of the crRNA. *E*, Schematic representation of the Watson-Crick base pairs between the crRNA and the target DNA. crRNA is colored green and target DNA is colored red unless stated otherwise. Disordered nucleotides are colored grey and the Watson-Crick base pairs are represented as vertical lines.

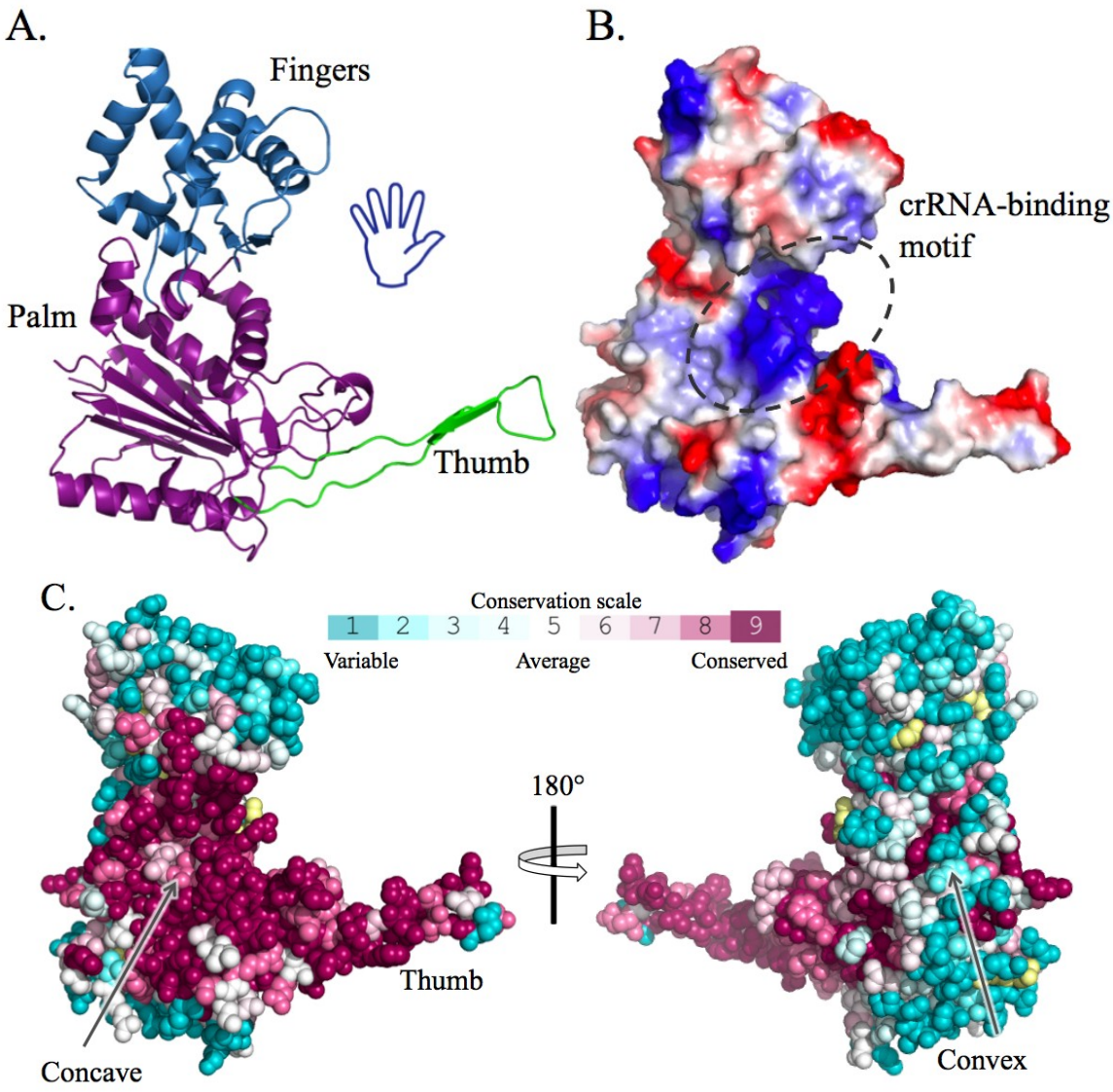
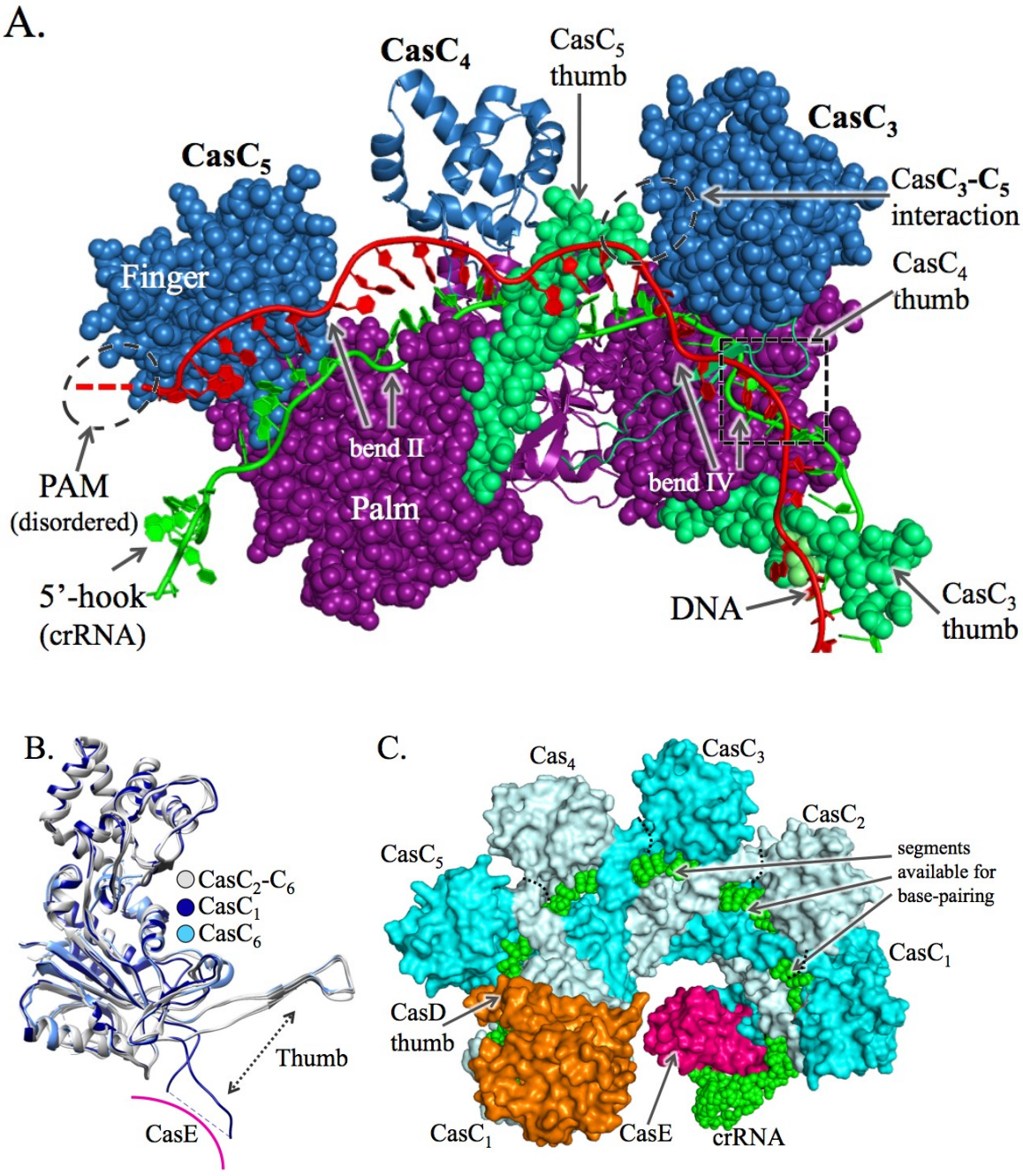


Figure 5.5. The CasC subunits of Cascade adopts a right-hand structure. *A,* Ribbon trace of the $CasC_3$ subunit of Cascade showing its Finger (blue), Palm (purple), and Thumb domains (green). *B,* An electrostatic potential surface of CasC shown in the same orientation as in *A.* The surface electrostatic potential is colored from positive (blue) to negative (red). *C,* Amino acid sequence conservation scores mapped onto the surface of $CasC_3$ using CONSURF.


A.
CasC4
CasC5 thumb
CasC3
CasC5
CasC3-C5 interaction
Finger
CasC4 thumb
bend II
bend IV
PAM (disordered)
Palm
5'-hook (crRNA)
DNA
CasC3 thumb
B.
CasC2-C6
CasC1
CasC6
Thumb
CasE
C.
Cas4
CasC3
CasC2
CasC5
segments available for base-pairing
CasC1
CasD thumb
CasC1
CasE
crRNA

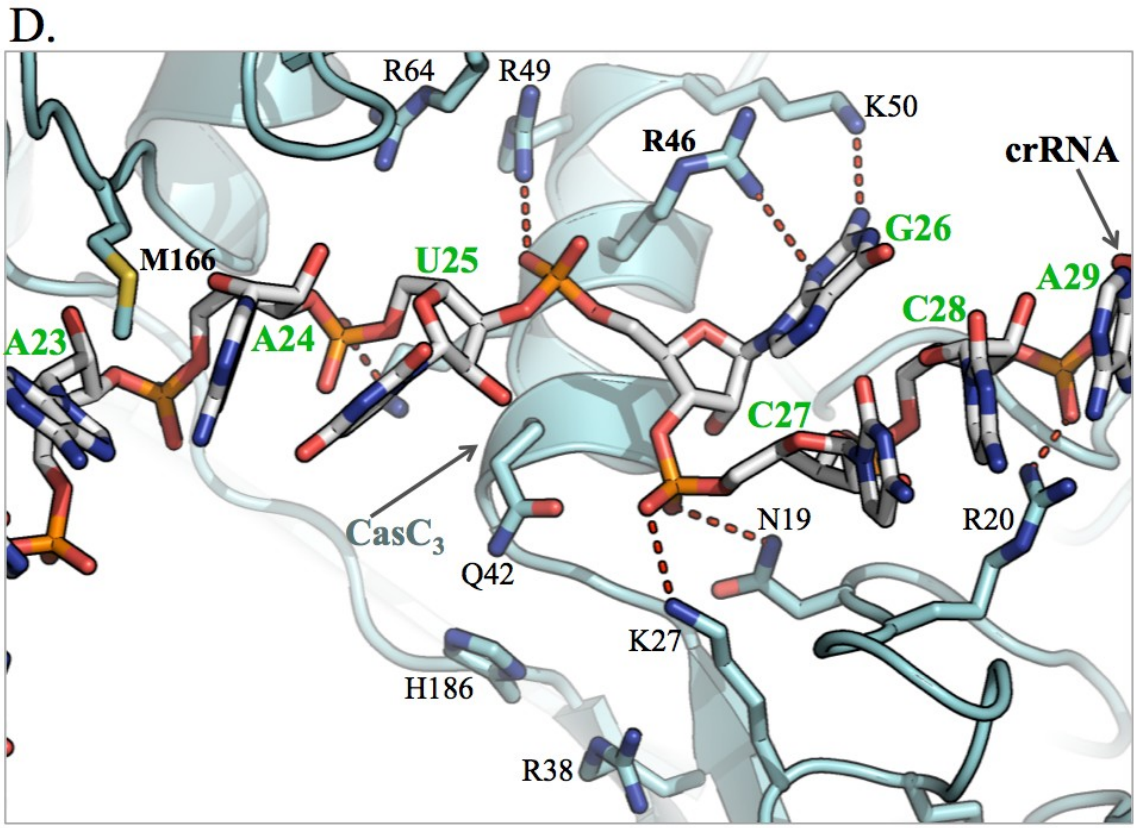


Figure 5.6. Implications of the helical arrangement of the CasC subunits on the crRNA-DNA hybrid structure. *A,* Interactions made by the palm and thumb domain result in evenly spaced out distortions in the crRNA-DNA hybrid. Only $CasC_3$ to $CasC_4$ are depicted and the other subunits have been removed for clarity. Interactions within the rectangular box are shown in *D. B,* Structural alignment of the six CasC subunits. $CasC_2$ to $CasC_6$ are colored in grey. $CasC_1$ and $CasC_6$ are colored in blue and cyan respectively. The position of CasE has been depicted with respect to $CasC_1$. *C,* Selective exposure of crRNA bases by CasC. CasA, CasB and the target DNA have been removed for clarity. *D,* Conserved residues in the basic patch of $CasC_3$ that interact with crRNA. $CasC_3$ is colored cyan. Individual elements are colored in case of the crRNA.

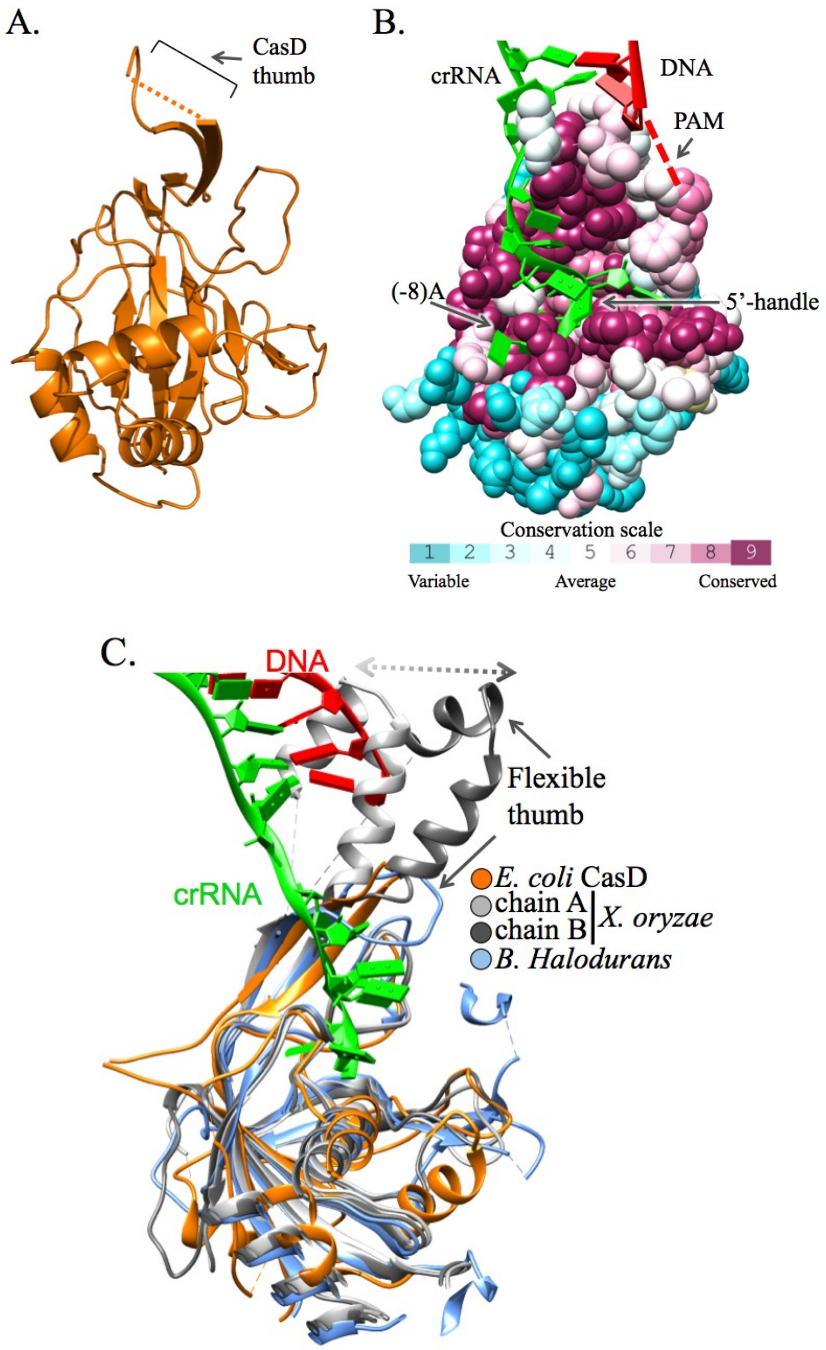


Figure 5.7. CasD caps the 5'-handle of the crRNA. *A,* Ribbon representation of CasD. *B,* The 5'-handle of crRNA interacts with the conserved groove in CasD. Amino acid sequence conservation scored mapped onto the surface of CasD. *C.* Structural alignment of CasD and its homologs from type I-C CRISPR systems.

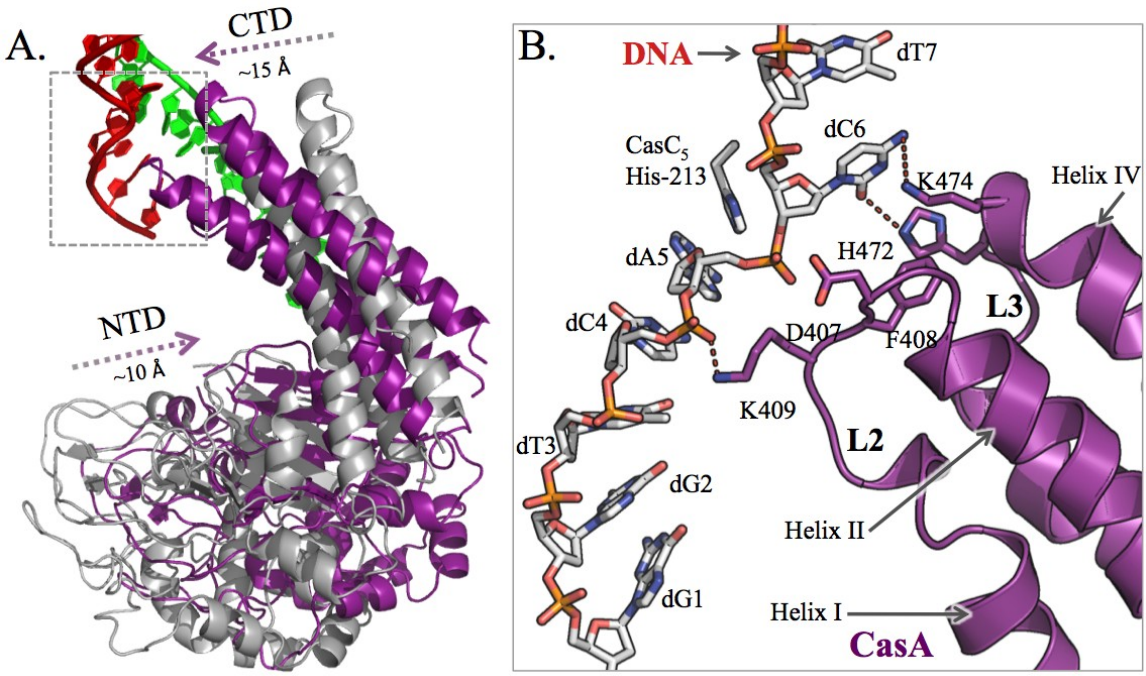


Figure 5.8. CasA conformational change results in specific contacts with the target strand. *A,* Ribbon diagram of CasA positions pre- (grey) and post- (purple) target binding. DNA and RNA are colored red and green respectively. *B,* The L2 and L3 loops in CasA results in the first bending of target DNA at position 6. Only the DNA strand is shown, and residues in the L2 and L3 loops close to the target DNA making specific interactions are shown as sticks. Hydrogen bonds are represented as red dotted lines. Also shown is His-213 from CasC ($CasC_2$ in this case) that makes aromatic stacking interaction at the sites of DNA distortions.

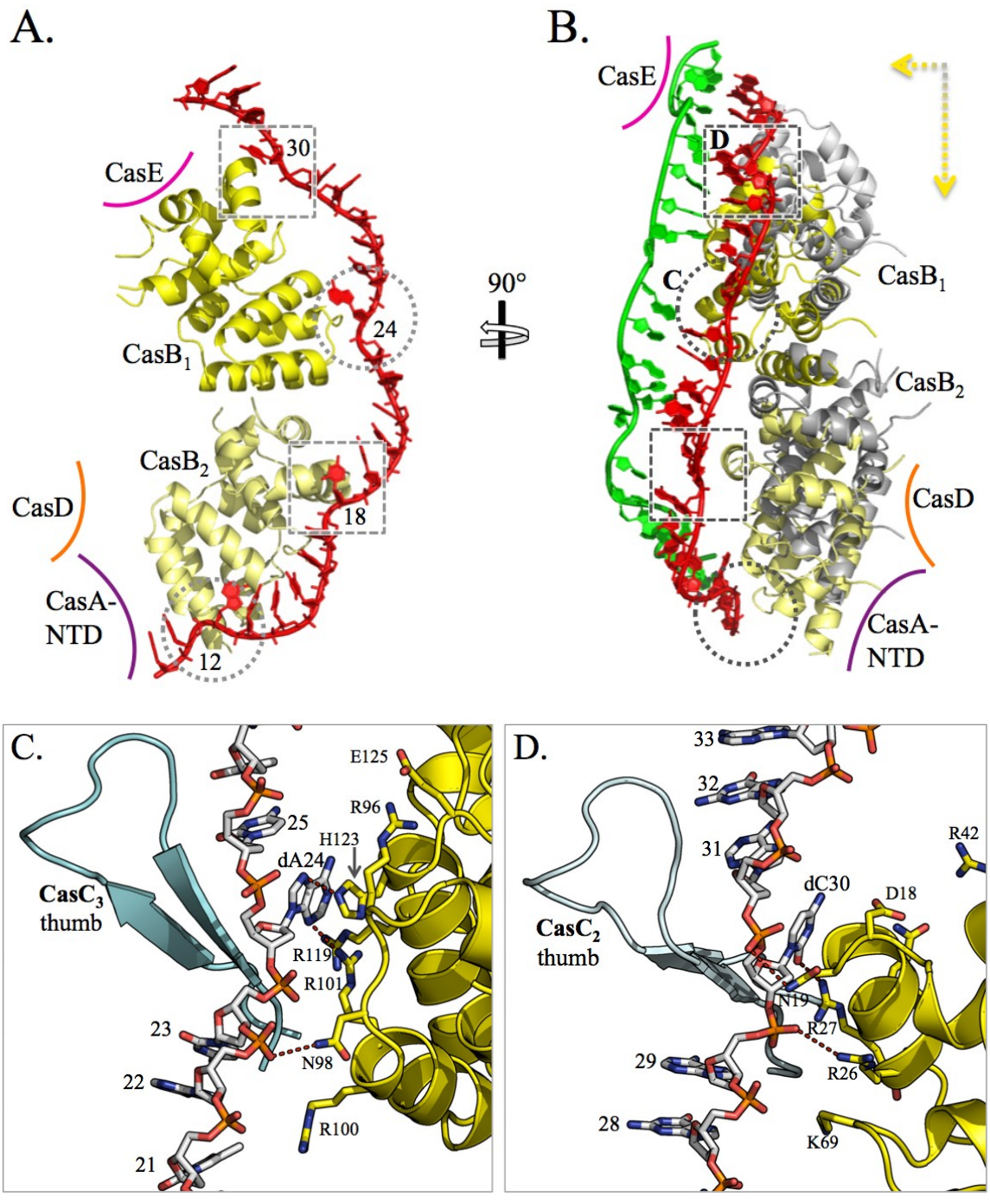

A.
CasE
30
24
CasB1
CasB2
18
CasD
CasA-
NTD
12
90°
B.
CasE
D
C
CasB1
CasB2
CasD
CasA-
NTD
C.
E125
R96
25
H123
dA24
CasC3
thumb
R119
R101
23
N98
22
R100
21
D.
33
32
31
R42
dC30
D18
CasC2
thumb
N19
R27
29
R26
28
K69

Figure 5.9. CasB subunits make specific interactions with the target DNA. *A,* Ribbon diagram of the two CasB (yellow) subunits next to the target DNA (red). The relative position of the CasA, CasB, and CasE subunits are marked. The dotted box or circles represent the sites of distorted bases on the DNA strand. *B,* Relative position of CasB subunits before (grey) and after (yellow) target binding to the DNA (red). The crRNA strand is colored green. The dotted arrows point to the relative movement of the CasB subunits upon target binding. The relative positions of the other subunits are marked as in *A.* C and D correspond zoomed-in positions shown in *C* and *D.C,* $CasB_1$-DNA interaction around dA24. *D,* $CasB_1$-DNA interaction around dC30. Similar interactions, as shown in *C* and *D*, also exist between $CasB_2$ and DNA at positions dA12 and dC18.

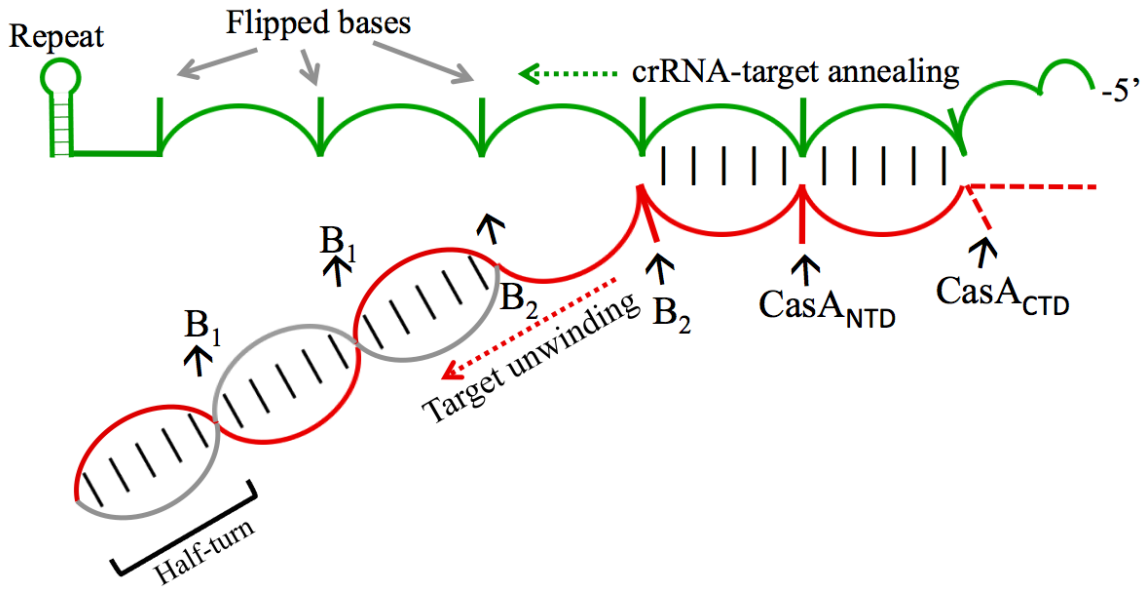


Figure 5.10. Schematic model of DNA unwinding by Cascade. The complementary and non-complementary strands of the target DNA are colored red and grey respectively. The crRNA is colored green. Watson-Crick base pairs are shown as vertical lines between the duplex strands. Cascade protein subunits are not shown for clarity.

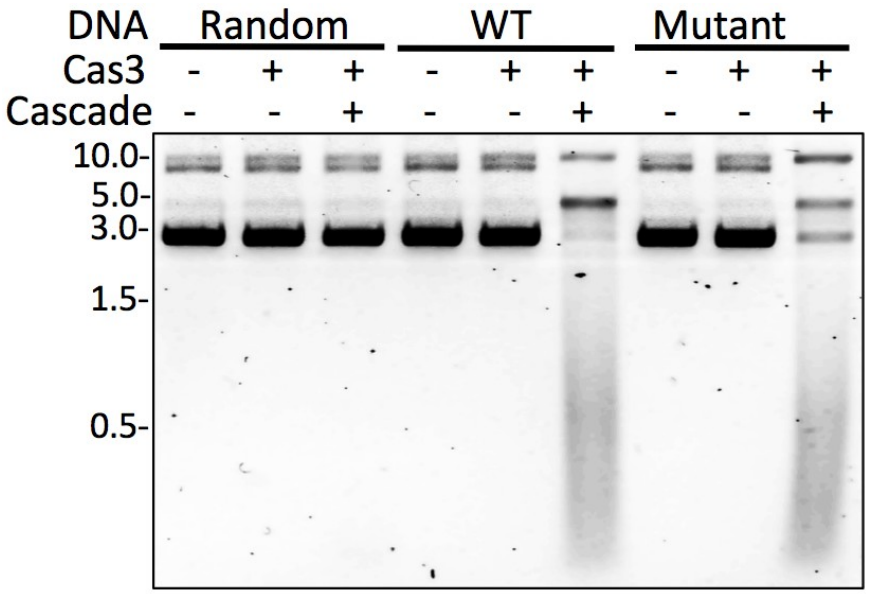


Figure 5.11. Reconstitution assay using wild-type and mutant protospacer. The latter consists of mismatches in each of the 6 flipped positions. Plasmid DNA (2 nM) consisting of complementary protospacer sequences were incubated with 20 nM Cascade, 50 nM Cas3, 2 mM ATP, 10 mM $MgCl_2$, and 10 μM $CoCl_2$ at 37 °C for 30 minutes before quenching with 20 mM EDTA. Mixtures were deproteinated before loading the samples on a 1% agarose gel. DNA bands were visualized with ethidium bromide staining. Random: pBAT4 plasmid, Mutant: Protospacer with mismatches at the flipped positions of the target DNA.

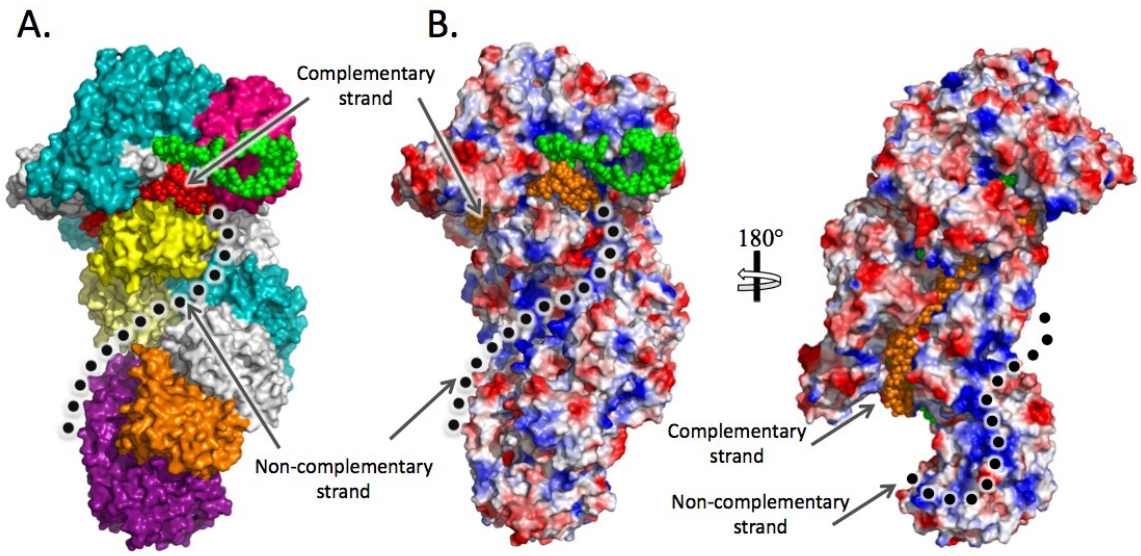


Figure 5.12. Potential path of the non-complementary strand of the target DNA lined with basic residues. *A,* Coloring scheme as the same as in Figure 5.2. *B.* An electrostatic potential surface of Cascade shown in two different orientations. The surface electrostatic potential is colored from positive (blue) to negative (red).

Chapter 6

Conclusion and Future Directions

In this chapter, I summarize the results of my thesis work, and present the questions that need to be addressed in the future. As mentioned earlier, the two main questions that I concentrated on during my thesis work are:

3. How are the different components of the CRISPR system able to recognize foreign DNA?
4. What is the fate of the foreign DNA, once recognized by the CRISPR system?

To answer these questions, I worked on Cascade and Cas3, the two main CRISPR components required in the interference stage. I carried out structural and biochemical experiments to investigate how Cascade binds to target DNA, and how both Cascade and Cas3 assemble to degrade DNA in the interference stage.

To investigate how the Cascade complex interacts with target DNA, we crystallized the type-I Cascade complex from *E. coli* bound to a complementary DNA strand, and solved its structure to ~3 Å by X-ray crystallography. The structure revealed an unusual arched-ladder structure formed by the crRNA-DNA core. This crRNA-DNA duplex is discontinuous as a result of regularly flipped-out bases, which in turn form five short duplex regions over the length of the spacer/protospacer. The spacing of the flipped bases is determined by the thumb domain of the CasC subunits, which polymerize on and then protect the crRNA. The CasC subunits also prepare the crRNA for target binding. Despite these findings, some expected interactions were not evident from the crystal structure.

The PAM motif is essential for target binding since the Cascade-PAM interaction precedes spacer-protospacer binding as shown to be the case in the type II CRISPR system (Sternberg et al., 2014). Although the PAM sequence (5'-CAT-3') is present in the target strand used during crystallization, it is disordered in the crystal structure. It is possible that a double-stranded PAM is required for the (CasA) L1- PAM interaction, and future crystallization experiments need to be designed using double-stranded target DNA. A single-stranded DNA strand that extends beyond the PAM might also stabilize the PAM-L1 interaction.

Although the crystal structure shows only the target strand, Cascade melts double-stranded target DNA, and based on protection assays, interacts extensively with both the complementary and the non-complementary strands (Jore et al., 2011). This raises the question of where in Cascade the non-complementary strand would bind. The crystal structure suggests that Cascade has a deep, positively-charged groove between the CasB and CasC subunits, on the opposite side of CasB with respect to the complementary strand. This groove is likely involved in the stabilization of the non-complementary strand. One of the surprising observations during Cascade structure determination was the high solvent content (~70%) of the Cascade-DNA complex crystals. The resulting large solvent channels in the crystal lattice would possibly allow for soaking experiments with the non-complementary strand into the Cascade-DNA crystal and may reveal its interaction with Cascade.

One of the major questions resulting from the intricate interactions revealed by the crystal structure is how a large complex like Cascade assembles. Is there an order in which the Cascade subunits are added onto the complex? The structure revealed that

CasD caps the 5' handle of the crRNA. Also, its thumb domain protrudes close to the -1 flipped base of the crRNA, blocking CasC interaction at this position. Also, towards the head of the complex, residues corresponding to the $CasC_1$ thumb domain are bent compared to that of the other CasC subunits. These observations are not sufficient to comment on the directionality of CasC polymerization. One way this could be tested is by expressing the CasC hexamers as a single protein linked by flexible linkers. Mutations can then be made in specific thumb domains, and their binding to crRNA probed by means of nuclease protection assays. Point mutations of conserved residues making interactions at the subunit-interfaces will also provide additional information on subunit assembly.

Based on the structural observations, it is possible that the target double-stranded DNA is unwound either half-turn or a turn at a time. Apart from the PAM, the crystal structure revealed all of the key interactions between the complementary strand of the DNA and the Cascade subunits. Alanine mutation analysis of residues making specific interactions with the DNA could likely trap the Cascade complex in different states and provide evidence on the possible mechanism of DNA unwinding by Cascade. DNA binding experiments using Cascade mutants are currently underway. Furthermore, using target DNA of varying lengths (shorter than the protospacer) during crystallization may also trap the complex at different conformations.

Besides Cascade, Cas3 is also essential in the interference stage as the latter consists of the nuclease-helicase required for target DNA melting and degradation. As such, to further investigate the mechanism of CRISPR interference, we crystallized the HD nuclease domain of *T. thermophilus* Cas3 and solved its crystal structure. It has a

conserved HD fold as predicted (Aravind and Koonin, 1998). We showed that this domain is a single-stranded endonuclease, and that its activity is stimulated by transition metal-ions. Even though we could only see one metal-ion at its active site, we showed by means of additional studies that a second metal-ion exists at the nuclease active site in solution, and that the nuclease likely uses a two-metal-ion dependent reaction mechanism. The absence of the second metal-ion was due to the acidic pH (~ 4.5) of the buffer used during protein crystallization.

The above results were directly applicable to our attempts to reconstitute the *E. coli* CRISPR interference stage *in vitro*. Cellular levels of transition metal ions like Co^{2+} and Ni^{2+} stimulate the endonuclease activity of *E. coli* Cas3. Cas3 can cleave single-stranded DNA but requires the R-loop-forming activity of Cascade (Jore et al., 2011) to target double-stranded DNA. Using the reconstituted system, we showed that Cascade recruits Cas3 to the sites of R-loops for DNA degradation. This recruitment is stimulated by the presence of ATP, and once recruited, Cas3 nicks the non-complementary strand at a specific site within the protospacer, and cleaves the target strand only upon ATP hydrolysis. Concerted helicase-nuclease activities of Cas3 powered by additional ATP hydrolysis results in unidirectional degradation of the target DNA.

A crystal structure of Cas3 would further add to our understanding of the mechanism of target degradation. Although, we do not have a crystal structure of the full Cas3 protein, we purified a Cas3 homolog from *Thermotoga maritima*, and identified initial crystal conditions. The structure of this protein is currently being pursued in the lab with initial models of the protein already built. The availability of both the Cascade complex and Cas3 presents us with the opportunity to investigate how Cascade is able to

recruit Cas3 during CRISPR interference. Based on DNA degradation patterns, Cas3 initially binds at the DNA fork created by Cascade, but Cas3 likely recognizes specific elements in Cascade as well. Since the DNA fork is expected to exist next to CasA, it is likely that Cas3 binds to one of the CasA domains. This hypothesis is supported by the observation that in several species, CasA and Cas3 exist as a fusion protein (Westra et al., 2012). The crystal structure revealed two disordered, solvent-exposed loops in the CasA-NTD (289-293 and 319-322) that could serve as the interface for CasA-Cas3 binding. The importance of these residues could be tested for Cas3 recruitment using assays established in Chapter 4. As mentioned earlier, the presence of large solvent channels in the Cascade-DNA crystals could allow for the Cas3 HD nuclease domain to be soaked into the crystals as well.

Several observations that we made with the reconstitution assay need to be investigated further. It is not clear how many Cas3 molecules bind to the R-loop before DNA degradation is initiated, and whether the same or different Cas3 molecules cleave the two strands of the target DNA. The possibility of Cas3 polymerization on the target DNA cannot be ruled out, however, it seems unlikely considering Cas3 has processive helicase activity (Sinkunas et al., 2011).

We also do not know how the nuclease and the helicase activities of Cas3 are coupled for efficient DNA degradation. The Cas3 helicase has a 3'-5' polarity (Sinkunas et al., 2011) but once Cas3 makes the initial cuts within the protospacer, it is not clear why it degrades the target unidirectionally with respect to the PAM. Increasing evidence suggests that CRISPR adaptation is synchronized to the interference stage (Datsenko et

al., 2012; Swarts et al., 2012), and as a result, target degradation will need to be investigated in the context of adaptation (Cas1 and Cas2 proteins) and vice versa.

Mapping of target DNA cleavage in the reconstitution assay suggests that Cas3 cleaves within the protospacer of the target strand at multiple sites. This region is tightly bound to the crRNA in the crystal structure presented in Chapter 5. These results suggest that Cascade is removed during target degradation. It is not clear whether the crRNA is cleaved, or if Cascade dissociates as a whole or as a smaller subcomplex. We showed in Chapter 2 that CasA is loosely bound to the Cascade complex at low concentrations, and that without CasA, the Cascade subcomplex is unable to bind double-stranded DNA. With the observation that CasA-Cas3 fusions exist in other species (Westra et al., 2012), it would be intriguing to investigate the interaction of the different subunits with Cas3. Förster Resonance Energy Transfer (FRET) experiments may be useful in delineating the conformational changes that might be at play at the Cascade-Cas3-DNA interface.

Overall, the structural and biochemical insights on the CRISPR interference stage presented as part of this thesis greatly increase our understanding of the CRISPR immune system. More importantly, it also raises critical questions that need to be addressed in the future, and would make an exciting endeavor for others to follow.

Appendix A

Methods

Preparation of Mineral Competent *E. coli* cells

1. Autoclave all the necessary equipments and solutions to avoid contamination.
2. Streak the *E. coli* strain of interest onto an LB plate with appropriate antibiotics and let the colonies grow overnight at 37 °C overnight.
3. Inoculate 1 L of sterile LB (with antibiotics, if needed) with a single colony from the LB plate and let the cells grow at 37 °C until an OD_{600} of ~0.5 is reached. Overnight starter cultures can be used as well to inoculate the 1 L culture.
4. Pellet cells at 2,500 rpm for 20 minutes at 4 °C and discard the supernatant.
5. Resuspend the cells in 30 ml of Tfb1 with a sterile 10 ml pipette.
6. Let the cells sit on ice for 10 minutes (At this point, the cells are very fragile)
7. Pellet the cells at 2,000 rpm for 20 minutes at 4 °C.
8. Discard the supernatant and carefully suspend the cells in TfbII.
9. Pipet aliquots into sterile 1.5 ml centrifuge tubes and flash freeze in liquid N_2 and store at -80 °C.

Tfb1: 30 mM CH_3COOK, 50 mM $MnCl_2$, 100 mM KCl, 10 mM $CaCl_2$, 15% glycerol

TfbII: 10 mM MOPS-NaOH at pH 7.0, 75 mM $CaCl_2$, 10 mM KCl, 15% glycerol

Mineral competent transformation protocol

1. Thaw ~ 100 µl of mineral competent cells on ice.
2. Add DNA (1 µl of plasmid or 10 µl of ligation mixture) to the cells in the 1.5 ml tube and mix gently. Incubate on ice for 15 min.
3. Heat shock the transformation mixture at 42 °C for 30-45 sec and then immediately place on ice. Incubate on ice for an additional 10-15 min.
4. Add ~300 µl of autoclaved LB media to the same tube and incubate at 37 °C with shaking for 30-60 min. This step is not necessary for Amp^r vectors.

5. Plate half of the mixture on LB plates with appropriate antibiotics and incubate the transformation plates at 37 °C overnight.

Standard Cloning Protocol- PCR reaction

Genes of interest can be amplified from genomic DNA, plasmids, or other sources using Polymerase Chain Reaction (PCR) as follows:

1. Dissolve the primers in dH_2O and make a 500 µM stock. Add an equivalent amount of water (in µl) to the nmols of primers in the tube to make this stock solution. Make a 10 µM working solution for subsequent reactions. Store primer solutions at -20 °C.

2. Prepare the following mixture in the order listed.

Component	[Stock]	Volume (µL)	[Final]
dH_2O		33.0	
HF Buffer*	5 x	10.0	1 x
Primer A†	10 µM	2.5	0.5 µM
Primer B†	10 µM	2.5	0.5 µM
dNTPs	10 mM	1.0	200 µM (each)
DNA‡		0.5	
Phusion	2.0 U/µl	0.5	0.02 U/µl
Total		50	

* HF Buffer will be the generic choice, but you can substitute the GC Buffer if your template DNA has a high GC content.

† Working solutions of primers are made by diluting 500 µM stock 1 in 50 to get working concentration of 10 µM.

‡ Concentration of DNA will vary based on its source. In general, whether the source is a plasmid or genomic DNA, you should use 0.5 µL.

Phusion: Phusion High-Fidelity DNA Polymerase (NEB F-530)

3. Setup a PCR cycle as follows:

Cycle	Temp. (°C)	Time	No. of Cycles
Pre-cycle	98	30 sec	1
Denaturation*	98	10 sec	35
Annealing†	?	20 sec	35
Extension	72	15-30 sec/kb	35
Post-cycle	72	4 min	1
	4	∞	1

* Denaturation is done at 98°C for Phusion DNA Polymerase, and you can vary the time beyond this suggestion, though it is not necessary.

† Annealing is done at temperature that is equal to the lower T_m of your two primers plus 3 °C (e.g. Primer A T_m=61 °C and Primer B T_m=52 °C, the Annealing temperature should be 52 °C + 3 °C = 55 °C).

4. Run a small amount of PCR product (~ 5 µL) on an ethidium bromide-stained agarose gel to verify amplification of DNA of appropriate length.

5. Clean the PCR product with GeneJET PCR purification kit (Thermo Scientific).

Digestion of PCR products and Vectors

1. Setup the following reaction mixture in a 1.5 ml tube.

Component	[Stock]	Volume (µl)	[Final]
DNA		28	
Buffer†	10 x	4	1 x
H_2O		6	
Enzyme 1	~10-20 U/µl	1	0.25 - 0.50 U/µl
Enzyme 2	~10-20 U/µl	1	0.25 - 0.50 U/µl
Total		40	

† Appropriate buffer should be chosen that supports activities of both restriction enzymes.

2. Mix thoroughly and incubate at 37 °C for 2-3 hours. The mixture can be incubated for much shorter time-periods (10-30 min) if using *Fast-Digest* restriction enzymes. Some of the enzymes have higher non-specific (star) activities and should not be incubated for longer than suggested.

3. Run the digested DNA on an ethidium bromide-strained agarose gel, excise the appropriate DNA bands and purify using GeneJET gel extraction kit (Thermo Scientific). Store the purified DNA/vector at -20 °C.

DNA ligation

1. Assemble the following reaction and incubate at ~20-25 °C for 60 min. This mixture can also be incubated overnight at 16 °C.

Component	Volume- Control (µl)	Volume- Insert (µl)
Vector	2.0	2.0
Insert	0	5.0
10 x Buffer	2.0	2.0
H_2O	15.0	10.0
T4 DNA Ligase	1.0	1.0
Total	20.0	20.0

2. Transform 10 µl of the ligation mixture into competent cells

Standard site-directed mutagenesis

1. Design primers such that 20-22 bases are included on both the 5' and 3' side of the mutation site with the last base being either a guanine or a cytosine.
2. Assemble the following reaction in the order listed:

Component	[Stock]	Control (µl)	Mutation (µl)	[Final]
dH_2O		44.5	37.5	
Pfu Ultra buffer	10 x	5.0	5.0	1 x
Primer A	10 µM	0	2.5	0.5 µM
Primer B	10 µM	0	2.5	0.5 µM
dNTPs	10 mM	0	1.0	200 µM each
DNA		0.5	0.5	
Pfu Ultra	2.5 U/µl	0	1.0	0.05 U/µl
Total		50.0	50.0	

3. Mix thoroughly and setup a PCR program as follows:

Cycle	Temp. (°C)	Time	No. of cycles
Pre-cycle	95	1 min	1
Denaturation	95	50 sec	18
Annealing	60	50 sec	18
Extension	68	2 min/kb	18
Post-cycle	68	7 min	1
	4	∞	1

4. Digest the PCR product from step 3 as follows:

Component	[Stock]	Volume (µl)	[Final]
H_2O		24.0	
DNA (step 3)		20.0	
DpnI buffer	10 x	5.0	1 x
DpnI enzyme	20 U/µl	1.0	0.40 U/µl
Total		50.0	

5. Incubate the mixture in step 4 at 37 °C for 4 – 6 hours. If convenient, the reaction can be incubated overnight at 37 ° C.

6. Transform 10 µl of the digested product into competent cells using standard transformation protocol. The control plate ideally should have no colonies. The mutation plate should have at least 10-fold more colonies than the control plate. If no colonies are observed in the mutation plate, different primers (of different length and sequence) can be tested next.

Multi-site mutagenesis

1. Assemble mixture in the order below:

Component	Control	Mutagenesis
H_2O	to 25 µl	to 25 µl
10x *Taq* ligase buffer	1x	1x
ATP	-	2 mM
dNTPs	-	1 mM each
Gel purified primers (1-5)	-	0.2 µM each
Template DNA	~100 ng	~100 ng
Pfu Ultra DNA polymerase	-	1 µl
NEB Taq ligase	-	1 µl
NEB T4 polynucleotide kinase	-	1 µl
Total	**25 µl**	**25 µl**

2. PCR cycle setup as follows:

   i. 37 °C for 30 min
   ii. 95 °C for 3 min
   iii. 95 °C for 1 min
   iv. 55 °C for 1 min
   v. 65 °C for 16 min (2 min/kb)
   Repeat 30 cycles for steps (iii), (iv), and (v)
   vi. 65 °C for 10 min
   vii. 4 °C until ready for next step

3. Add 1 µL of NEB DpnI directly to the mixture and incubate the reaction for 4-6 hours @ 37 °C (overnight if convenient).

4. Transform 5 µL of the digested DNA into DH5α strain of *E. coli* and let the colonies grow overnight on the LB plate.

5. Grow 2-3 colonies and test digest with appropriate restriction enzymes to verify the entire gene is intact. Because this protocol uses primers that anneal on a single (same) template strand, I have found that in case of some of the colonies, the plasmids are truncated (and usually your gene is truncated). So I make sure the digestion products from my test digests look identical to results I expect in case of the template DNA.

6. If the test digestions make sense, send 2 samples for sequencing.

Notes

1. I get better results (higher number of colonies in my mutagenesis plate compared to my control plate) when I gel purify each of the primers (Denaturing UREA gel).

2. When I use 5 primers (for 5 mutations), I usually find 4-5 of the sites mutated (different combinations of sites mutated in DNA from different colonies).

3. I get more colonies when I do the mutagenesis with shorter plasmids (overall length of the plasmid does have an effect).

Large Scale Plasmid Purification

1. Prepare the following buffers:

   Buffer A (500 ml): 40 mM glucose, 25 mM Tris-HCl at pH 8.0, 10 mM EDTA
   Buffer B (200 ml): 0.2 N NaOH, 1% sodium dodecyl sulfate
   Buffer C (500 ml): 7.5 M Ammonium acetate at pH 7.6
   Buffer D (200 ml): 2 M Ammonium acetate at pH 7.4
   Buffer E (200 ml): 10 mM Tris-HCl at pH 8.0, 0.1 mM EDTA

2. Transform the plasmid of choice into DH5α cells and plate on LB plate with appropriate antibiotics.

3. Pick a single colony and use it to grow 3 x 1L of LB culture overnight at 37 °C.

4. Pellet the overnight culture. Freeze and store the cell pellets at -20 °C until if not being used immediately.

5. Resuspend the cells in 10 ml of Buffer A and add lysozyme to a concentration of 5.2 mg/ml. Immediately place the cells on ice.

6. Add 100 ml of buffer B, mix with a stir rod and place on ice for an additional 5 mins.

7. Add 75 ml of buffer C, mix well, and let the solution sit on ice for 1 h.

8. Centrifuge the solution at 10,000 rpm for 30 min and recover the supernatant in a 500 ml bottle after passing it through a cheese cloth.

9. To the supernatant, add 120 ml of isopropanol, mix gently, and centrifuge at 12,000 rpm for 30 min. Discard the supernatant and remove as much of the liquid as possible.

10. Resuspend the pellets in 20 ml of Buffer D, add 10 µl of 10 mg/ml RNase A, and incubate the mixture at 37 °C for 30 min. Centrifuge the solution at 5,000 rpm for 30 min to pellet any precipitations.

11. Treat the supernatant twice with equal volumes of phenol:chloroform:isoamyl alcohol (25:24:1) solution and once with chloroform:isoamyl alcohol.

12. Add 40 ml of ethanol and let the plasmid precipitate overnight at -20 °C. Centrifuge the solution at 10,000 rpm for 30 min, wash with 80% ethanol, and centrifuge again at 10,000 rpm for 60 min.

13. Discard the supernatant, dry the plasmid pellet, and resuspend it gently in 1.5 ml of Buffer E. Determine the concentration of the plasmid using a nanodrop.

Run-off RNA transcription

1. Linearize template plasmid appropriately.

2. Set up the following reaction mixture in the listed order.

   200 mM HEPES at pH 7.5
   0.1 mg/ml BSA
   25 mM $MgCl_2$
   40 mM DTT
   2 mM Spermidine
   6 mM NTP at pH 7.5
   50ug/mL template DNA
   1x RNAsecure inhibitor
   T7 RNA polymerase

3. Incubate the mixture without the T7 RNA polymerase at 37 °C for 10 min prior to starting the transcription reaction. Incubate the reaction at 37 °C for an additional 1-2 h. Its best to optimize for Mg^{2+}, template DNA, and RNA polymerase concentrations.

4. Add EDTA to 50 mM to get rid of the pyrophosphate precipitation.

5. The products can be verified by running a small sample on denaturing urea gel.

UV-crosslinking of Cascade with DNA

1. Assemble the mixture below and incubate at room temperature for 30 minutes.

Components	Concentration, nM	
	Control	Cascade
H2O	-	-
5x Gel shift buffer	1x	1x
Cascade	0	300
DNA	0.5	0.5
Total	20 µl	20 µl

2. Make dilutions of Cascade as necessary.

3. Expose to UV source for 3 minutes followed by boiling 16 µl of solution from (1) with 10 µl of 5x SDS buffer.

4. Run an appropriate SDS-PAGE gel to separate the samples.

Expression of E. coli Cascade in EZ-rich defined media

1. Transform Cascade components (*casBCDE* in pHAT4, *casA* and crRNA in pRSF-duet vector) into T7EXPRESS strain of *E. coli* cells and plane on LB-Agar media with appropriate antibiotics. Incubate the plate overnight at 37 °C.

2. Generously swipe-off some *E. coli* cells using a fresh pipette tip and transfer them into ~100 ml of EZ-ΔMet media with appropriate antibiotics. Let the cells grow overnight at 37 °C. Starter cultures can also be made from glycerol stocks.

3. Transfer the overnight culture into ~900 ml of prewarmed EZ-ΔMet media and let the culture grow until OD_{600} of ~0.5 is reached.

4. Dilute the culture from the last step into half using an additional 1 L of EZ-ΔMet media. Let the cultures get to OD_{600} of ~0.5 so as to have 2 L of starter culture in the exponential growth phase.

5. Add 120 ml of culture from step 5 into 16 x 1 L of prewarmed media and grow the cultures at 37 °C until $OD6_{00}$ of 0.3 is reached. This will result in about 18 L of culture.

6. Drop the temperature of the shaker to 20 °C and add 50 mg/l SeMet. Induce protein expression at OD_{600} of ~0.4 by adding IPTG to 0.5 mM.

7. Harvest cells after an additional 20 hours and flash-freeze the cell pellets in liquid N_2. Store at -20 °C until ready for protein purification.

EZ-ΔMet media	Volume (ml)
10x MOPS mixture	100
0.132 M K_2HP0_4 (autoclaved)	10
10x ACGU solution	100
10x Supplement EZ-ΔMet solution	100
40 % Glucose	10
Autoclaved H_2O	680
Antibiotics	1x
Total	1000 ml

10x MOPS mixture

1. Mix the following in ~600 ml of H_2O.

Component	Formula weight	Grams
MOPS	209.3	167.5
Tricine	179.2	14.4

2. Add 10 M KOH to the solution until the final pH is 7.4 and bring the total volume to 880 ml.

3. Add 20 ml of freshly prepared 0.1 M $FeSO_4$.

4. Add the following solutions in order.

Component	**Volume (ml)**
1.9 M NH_4Cl	100
0.276 M K_2SO_4	20
0.02 M $CaCl_2$	0.5
2.5 M $MgCl_2$	4.2
5 M NaCl	200
Micronutrient stock	0.4
H_2O	~774

5. Filter sterilize with 0.2 micron filter. Use fresh or aliquot in sterile bottles and freeze at -20 °C to be used later.

6. Mix the following ingredients to make 50 ml of the micronutrient stock. Store at room temperature.

Component	**Formula weight**	**Amount (mg)**
$(NH_4)_6Mo_7O_{24}.4H_2O$	1235.9	9
H_3BO_3	51.83	62
$CoCl_2$	237.9	18
$CuSO_4$	249.7	6
$MnCl_2$	197.9	40
$ZnSO_4$	287.5	7

10x ACGU solution

1. Add the following ingredients in 2000 ml of 15 mM KOH. The solution can be heated gently to facilitate dissolution.

Component	Formula weight	Amount (g)
Adenine	135.13	0.540
Cytosine	111.1	0.444
Uracil	112.09	0.448
Guanine	151.13	0.604

2. Filter-sterilize the solution with 0.2 micron filters and freeze at -20 °C in appropriate aliquots, if not being used immediately.

Notes

1. I usually keep frozen stocks of all the solutions to make 1 L of EZ-ΔMet media for starter cultures.
2. I make the media on the same day I do protein expression.

10x Supplement EZ-ΔMet solution

1. Dissolve the following in order in ~400 ml of 10 mM KOH.

Amino acid	**Amount (mg)**	**Amino acid**	**Amount (mg)**
Tyr (free)	100	Leu (free)	100
Ala (free)	100	Lys (HCl)	200
Arg (HCl)	2000	**Met**	**0**
Asn (free)	100	Phe (free)	200
Asp (free)	100	Pro (free)	100
Glu (K salt)	200	Ser (free)	2000
Gln (free)	200	Thr (free)	200
Gly (free)	100	Trp (free)	40
His (free)	100	Val (free)	200
Ile (free)	20	Cys[#] (free)	100

[#] Add separately to ~20 ml of H_20. Dissolve by adding 12 M HCl drop-wise until solution is clear.

2. Add 100 ml of 100x VA Vitamin solution before adding H_2O to 2000 ml.

3. Add the following components to make 500 ml of 100x VA Vitamin solution. The solution should be stored at -20 °C.

Stock solution	**Amount (g)**	**Solvent**
0.02 M *p*-aminobenzoic acid	0.069	40 ml of 0.01 M KOH
0.02 M *p*-hydroxybenzoic acid	0.069	40 ml of 0.01 M KOH
0.02 M 2,3-dihydroxybenzoic acid	0.077	40 ml of 0.01 M KOH
0.02 M thiamine HCl	0.169	
0.02 M calcium pantothenate	0.238	
Total	**500 ml**	

Crystallization of Cascade-DNA complex

1. Resuspended DNA in dH_20 such that the concentration is ~500 μM. Verify DNA concentration with the nanodrop.

2. Heat the following mixtures at 95 °C for 3 minutes and allow for slow cooling on the heat block over 2 hours.

Component	Concentration
H_2O	-
10x Annealing buffer[#]	1x
Complementary strand	200 μM
Non-complementary strand	230 μM
Total Volume	50 μl

[#] 10x Annealing buffer: 20 mM Tris-HCl pH 7.5, 100 mM NaCl and 0.5 mM of EDTA

3. Incubate the following mixture at 37 °C for 30 minutes followed by chilling on ice for ~5 minutes. Spin mixture at 14,000 rpm for 1 minute to pellet any precipitation.

Component	Final concentration (μM)
dsDNA target	30
Cascade	20
Buffer C	to 60 μl
Total	**60 μl**

4. To get initial crystals, prepare precipitant solution consisting of 0.08 M calcium acetate, 0.1 M sodium cacodylate at pH 5.0, and 8-11 % of PEG 8,000. I usually vary the PEG concentration in 0.2% increments to make sure that I get the initial crystals.

5. Mix 4.1 μL of precipitant solution with 8.2 μL of the Cascade-DNA complex thoroughly without getting any bubbles. Pipette 3 μL (x 4 wells) of the resulting solution on sitting-drop well with a 500 μl reservoir solution. I make the reservoir solution in separate 15 ml tubes and pipette them into the reservoir right before I seal the wells (one column at a time). Once, I have all the columns sealed, I put an additional layer of tape.

6. Store the crystals in the 20 °C room. Smaller crystals appear within 24 h but the larger crystals (~500 μm x 300 μm x 300 μm) take about a week to grow. These crystals should be harvested within 2 weeks.

7. While possible, the likelihood of getting these larger crystals without microseeding is very low. Hence, the following microseeding procedure can be used to grow larger crystals on a consistent basis.

Microseeding

1. Identify drops with crystals. The crystals can be of any size but bigger starting crystals most often give better crystals.

2. Remove the skin that forms on the surface of the crystal drop with a loop. Stabilize the drop using additional buffer from the reservoir solution. At this point, protein that has not crystallized will precipitate. Exchange buffer from the drop until the solution is clear and leave about 10 uL of buffer for the next step.

3. Crush the crystals with a pipette tip and transfer them into a 1.5 ml centrifuge tube with the help of 250 ml of a solution consisting of 0.08 M calcium acetate, 0.1 M sodium cacodylate at pH 5.0, and 8.7 % PEG 8,000. Add a seed bead to the same tube (Hampton).

4. Vortex the tube at the highest setting for 5 minutes and make a 1/10 dilution stock. Make six additional two-fold dilutions.

5. Mix 4.1 μl of seed solutions from steps 4 or 5 with 8.2 μl of Cascade-DNA complex. Add 3 μl of the resulting mixture to each well with the reservoir solution always consisting of 0.08 M calcium acetate, 0.1 M sodium cacodylate at pH 5.0, and 8.7 % PEG 8,000.

6. Crystals should appear in 25-48 hours and should be ready to be harvested in approximately 10 days. Crystals that appear after ~2 days grow into bigger crystals.

Harvesting and stabilization of Cascade-DNA crystals

Most of the Cascade crystals grow stuck to the surface of the well. Some crystals also grow stuck to the layer of skin that forms on the surface of the drop. Crystal harvesting, stabilization and soaking can be carried out as follows:

1. Prepare a well on a separate tray with 500 μl and 100 μl of precipitant solutions in the reservoir and the crystal well respectively.

2. In the drop with crystals, remove the skin covering the crystal drop with the help of a crystal loop. Stabilize the drop with additional solution from the reservoir until most of the precipitation has been removed.

3. Dislodge crystals stuck to the bottom of the well with gentle pushes near the area of the crystal with a sharp acupuncture needle.

4. Once the crystal is free for manipulation, transfer the crystal to the stabilization well.

5. Buffer exchange into a cryo solution consisting of 0.1 M sodium cacodylate at pH 5.0, 0.1 M calcium acetate, 10 % PEG 8,0000, and 5.0 % each of (PEG 400, ethylene glycol, glycerol, and sucrose) in >5% steps. I usually have the crystals (~ 20-40) in ~ 200 µl of the stabilization solution before slowly adding the cryo solution in 10 µl increments up to 100 % of the cryo solution.

6. Heavy atom soaks can be carried out with the crystals stabilized in the cryo solution.

Purification of CRISPR-components from different species

Purification of CasB subunit of the Cascade complex from Thermus thermophilus

1. Gene encoding CasB was cloned into mRSF plasmid (derivative of pRSF-duet plasmid with gene encoding for maltose-binding protein (MBP) proximal to the first multiple cloning site) so as to express CasB with an N-terminal MBP tag.

2. The resulting plasmid was transformed into T7EXPRESS strain of *E. coli* cells (NEB) and plated on LB-Agar media plates (*kan*^{r}).

3. A handful of colonies were used to grow cell cultures in LB media (1L x3). Cells were grown at 37 °C to OD_{600} of ~0.6 before temperature was lowered to 20 °C and IPTG was added to 0.1 mM. Cells were allowed to grow overnight.

4. The cells were harvested the next day (after ~18 h post-induction) and resuspended in Buffer A.

5. Cells pellets were lysed and loaded on a 5 ml amylose column (GE Healthcare). The column was subsequently washed with ~50 ml of Buffer D supplemented with 1 M urea, after which, the column was rewashed with ~25 ml of Buffer A.

6. Protein was eluted off the amylose column with ~15 ml of Buffer E. The eluted protein was incubated with $1/10^{th}$ (by mass) of TEV protease overnight at 4 °C.

7. The following day, protein mixture was first heated at 70 °C for 10 min and then placed on ice for 5 min. The resulting precipitation was pelleted by centrifugation at 18,000 rpm for 30 min.

8. The supernatant was collected and loaded on a HiLoad 26/60 S200 column pre-equilibrated in Buffer F in 10-ml batches. Protein eluted as a dimmer still containing some uncleaved (MBP) CasB.

9. Appropriate fractions were collected and through a 5-ml amylose column once more. The flow-through was collected and concentrated to ~ 5 mg/mL. The final protein sample was >95% pure as judge by Coomassie staining of an SDS-PAGE gel and had negligible nucleic acid contamination based on its low A_{260}:A_{280} (~0.65).

Purification of Thermotoga maritima MSB8 Cas3

1. *T. maritima cas3* was cloned into pHAT2 vector and transformed into T7EXPRESS strain of *E. coli.*

2. Cells were grown in LB media (6 L) at 37 °C until an OD_{600} of 0.3. Temperature was turned down to 20 °C and the cells were allowed to grow to an OD_{600} of 0.5. At this point, IPTG was added to 0.2 mM to induce protein expression. Cells were subsequently harvested after 20 hours.

3. Cells were lysed in *buffer A* and the soluble fraction was loaded on a 5-ml IMAC affinity column. The loaded column was sequentially washed with 100 ml each of *buffer A* and 10% of *buffer B*. About 15 mgs of *T. maritima* Cas3 was eluted with 100% of *buffer B.*

4. Protein was loaded on a HiLoad 26/60 S200 column. Cas3 eluted at a volume consistent with its size of ~84.3 kDa. Appropriate fractions were pooled and the NaCl concentration was diluted down to 100 mM before being loaded on a HT 5ml-Q-affinity column. Protein was eluted with a linear gradient between 0-50 % of *buffer D* over a volume of 25 ml. Appropriate fractions were pooled and dialyzed back into *buffer C.*

5. The final Cas3 sample was >95% pure as judged by Coomassie staining of an SDS-PAGE gel. The protein sample was concentrated to ~12 mg/ml and was used to setup the Classics and JCSG crystal screens. Crystals were readily seen in multiple drops after 4-5 days.

Cloning and purification of CFP-Cas3-YFP fusion protein

1. A modified prSET-b plasmid with sequences encoding for CFP and YFP separated by SphI and BglII restriction sites were used as template.

2. *E. coli* Cas3 was PCR amplified with SphI and BglII restriction sites on its 5'- and 3'- sites and cloned into the prSET plasmid such that the resulting coding region had sequences for *CFP, cas3, and YFP* in that order.

3. Purification was carried out essentially as in case of wild-type *E. coli* Cas3.

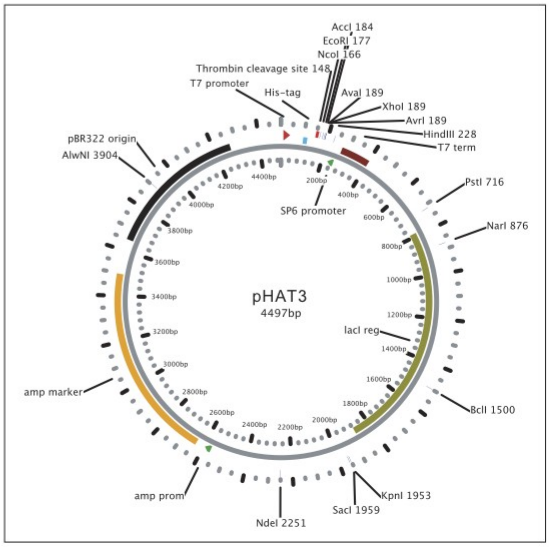
Appendix B

pHAT His-tag fusion vectors

```
                T7 promoter          lac operator          (XbaI) T7 ribosome binding site
1   TAATACGACTCACTATAGGGAATTGTGAGCGGATAACAATTCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATCC  85

                                                  Eco47III   BamHI
                                          SpeI          NcoI                  NotI    NsiI HindIII
pHAT   TATCCATGAACACCATTCATCACCATCACCATCACAACACTAGTAGCGCTACCATGGATCCGGGCCCTCTAGATGCGGCCGCATGCATAAGCTT
           M  N  T  I  H  H  H  H  H  H  N  T  S  S  A  T  M  D  P

                                  NcoI       EcoRI SalI XhoI   BamHI                NotI    NsiI HindIII
pHAT2  TATCCATGAGTCATCACCATCACCATCACTCCATGGATATCGAATTCGTCGACCTCGAGGGATCCGGGCCCTCTAGATGCGGCCGCATGCATAAGCTT
           M  S  H  H  H  H  H  H  S  M  D  I  E  F  V  D  L  E  G  S

                                                                                          NcoI
                                          SpeI                                        BamHI          EcoRI
pHAT3  TATCCATGAACACCATTCATCACCATCACCATCACAACACTAGTGGATCTGGTGGTGGTGGCGGCCGCCTGGTTCCGAGGGGATCCATGGATATCGAATTC
           M  N  T  I  H  H  H  H  H  H  N  T  S  G  S  G  G  G  G  G  R  L  V  P  R△G  S  M  D  I  E  F
                                                                                  Thrombin
       SalI XhoI       BamHI                      NsiI  HindIII
       GTCGACCGACCTCGAGGGATCCGGGCCCTCTAGATGCGGCCGCATGCATAAGCTT
       V  D  L  E  G  S

                                          SpeI                                        BamHI
pHAT4  TATCCATGAACACCATTCATCACCATCACCATCACAACACTAGTGGATCTGGTGGTGGTGGCGGCCGCCTGGTTCCGAGGGGATCCATGTCTGAAAACCTG
           M  N  T  I  H  H  H  H  H  H  N  T  S  G  S  G  G  G  G  G  R  L  V  P  R△G  S  M  S  E  N  L
                                                                                  Thrombin
                  NcoI       EcoRI SalI  XhoI  BamHI                      NsiI  HindIII
       TACTTCCAGGGTTCCATGGATATCGAATTCGTCGACCTCGAGGGATCCGGGCCCTCTAGATGCGGCCGCATGCATAAGCTT
       Y  F  Q△G  S  M  D  I  E  F  V
              TEV

             EcoRV       SalI
          NcoI       EcoRI       XhoI  BamHI                    HindIII
pHAT5  TATCCATGGATATCGAATTCGTCGACCTCGAGGGATCCCACCATCACCATCACCATTAAGCTT
           M  D  I  E  F  V  D  L  E  G  S  H  H  H  H  H  H  *

            SP6 promoter
GAGTATTTCTATAGTGTCACCTAAATCCCAGCTTGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACC

             T7 transcription terminator
GCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGG
```

Feature table

	Position in pHAT2
T7 promoter	1 - 17
lac operator	21 - 42
T7 gene 10 RBS	50 - 80
His-tag	98 - 115
Polylinker	83 - 175
SP6 promoter	180 - 198
T7 transcription terminator	277 - 318
LacI gene (coding region)	728 - 1809
β-lactamase (coding region)	2579 - 3429
Origin of replication	3504 - 4199



References:

Reference for pHAT is:
Peränen J., Rikkonen M.,Hyvönen M., Kääriäinen L. (1996)
T7 vectors with modified T7*lac* promoter for expression of proteins in *Escherichia coli*. Anal. Biochem., 236:371-373

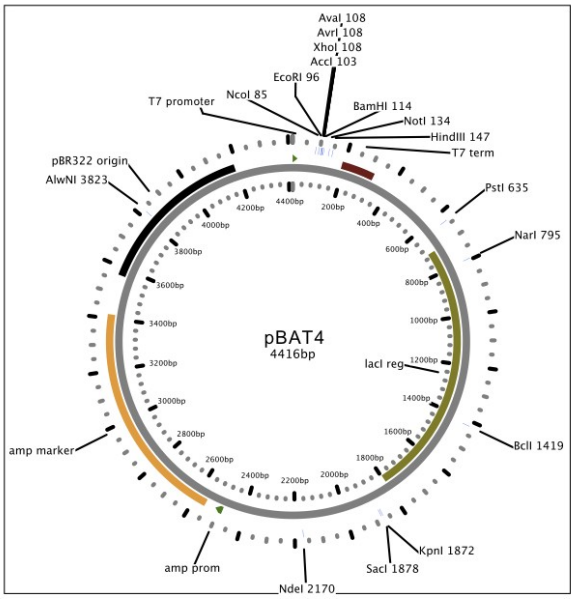
pHAT2 and pHAT5 were created by Marko Hyvönen.
pHAT3 was cloned from pTAT8 by SpeI deletion by David Story.
pHAT4 was cloned by SpeI deletion from pGAT3 by Carina Lobley and Mairi Kilkenny.
For more information see http://www-cryst.bioc.cam.ac.uk/~marko/vector, or email marko@cryst.bioc.cam.ac.uk

pBAT expression vectors

```
                T7 promoter          lac operator          (XbaI) T7 ribosome binding site
1   TAATACGACTCACTATAGGGAATTGTGAGCGGATAACAATTCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATA  80

pBAT4   NcoI       EcoRI SalI  XhoI  BamHI              NotI    NsiI  HindIII
        TATCCATGGATATCGAATTCGTCGACCTCGAGGGATCCGGGCCCTCTAGATGCGGCCGCATGCATAAGCTT
            M  D  I  E  F  V  D  L  E  G  S

pBAT5   SnaBI BglII  SalI  NotI    NsiI  HindIII
        TACGTACAGATCTGTCGACGCGGCCGCATGCATAAGCTT

                SP6 promoter
        GAGTATTCTATAGTGTCACCTAAATCCCAGCTTGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACC

                T7 transcription terminator
        GCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGG
```

Feature table

	Position in pBAT4
T7 promoter	1 - 17
lac operator	21 - 42
T7 gene 10 RBS	50 - 80
Polylinker	83-152
SP6 promoter	157-175
T7 transcriptional terminator	254-295
lacI gene (coding region)	705-1786
b-lactamase gene (coding region)	2556-3406
origin of replication	3481-4176
Total length	4416

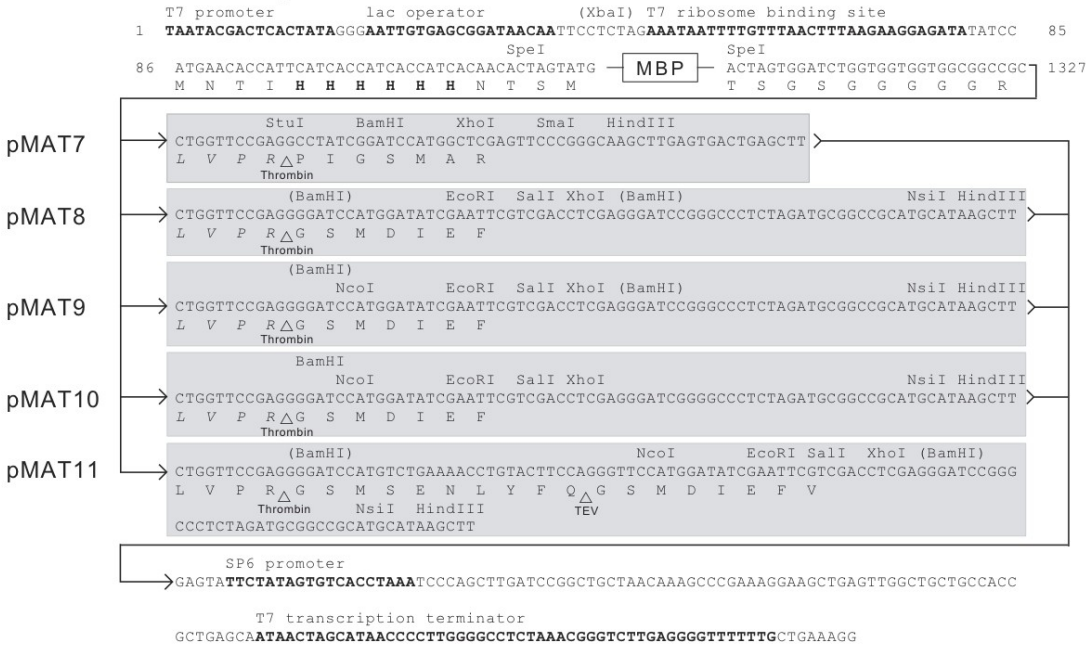


References:

Reference for these original non-fusion vectors is:
J., Rikkonen M.,Hyvönen M., Kääriäinen L. (1996)
T7 vectors with modified T7*lac* promoter for expression of proteins in *Escherichia coli*. Anal. Biochem., 236:371-373

For more information, see http://www-cryst.bioc.cam.ac.uk/~marko/vector, or email marko@cryst.bioc.cam.ac.uk

pMAT MBP fusion vectors



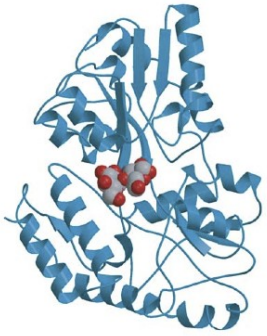
Feature table

	Position in pMAT10
T7 promoter	1 - 17
lac operator	21 - 42
T7 gene 10 RBS	50 - 80
His-tag	98 - 115
Maltose binding protein	125 - 1237
Polylinker	1280 - 1351
SP6 promoter	1358 - 1375
T7 transcription terminator	1454 - 1495
LacI gene (coding region)	1905 - 2986
β-lactamase (coding region)	3756 - 460
Origin of replication	4681 - 5376

Fusion protein data

(until thrombin cleavage site in pMAT10)

Length	398 aa
Molecular weight	43540.14
pI	5.66
ε_{280} ($M^{-1}cm^{-1}$)	64720



Maltose binding protein, PDB:1MDQ

References:

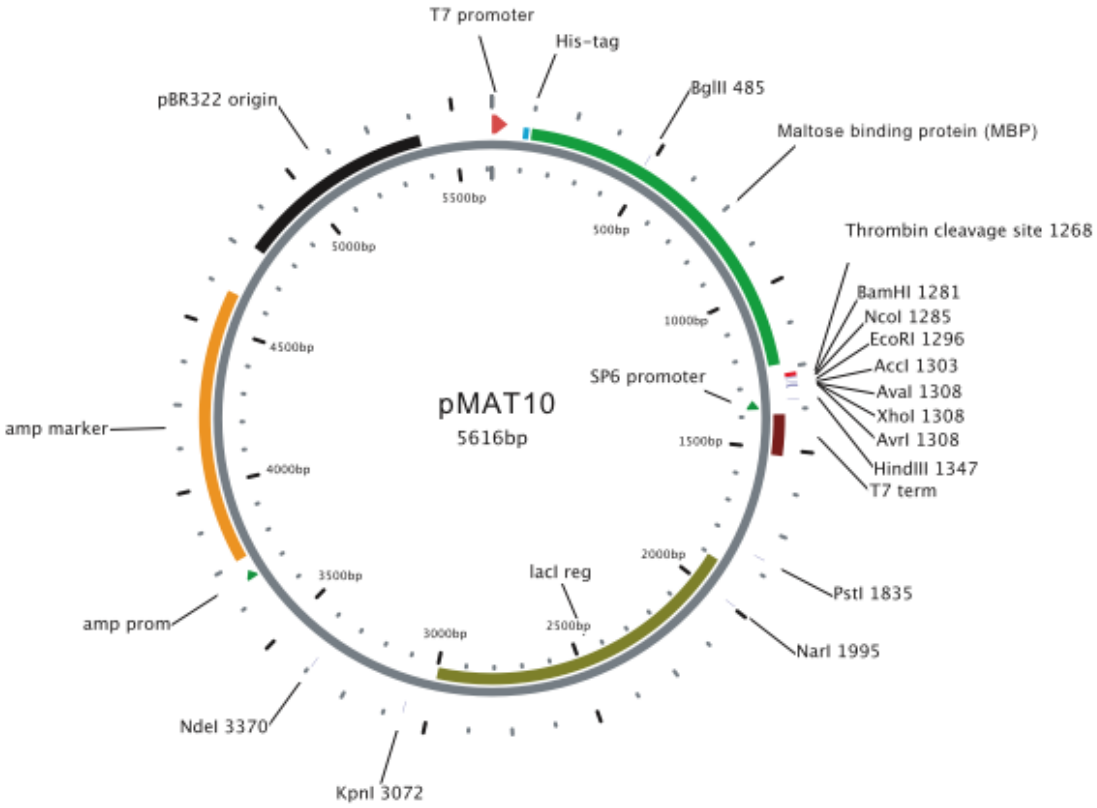
pMAT7 was created by Johan Peränen, and should be cited as: J. Peränen, unpublished.
pMAT8 was created by Elliott Stollar from pMAT7, and should be cited as: E. Stollar, unpublished.
pMAT9 was created by Jenny Reed and Marko Hyvönen from pMAT8 by mutating the internal NcoI site from MBP, and should be cited as: J. Reed and M. Hyvönen, unpublished.
pMAT10 was created by Keily Littlefield and Darerca Owen from pMAT9 by mutating the second BamHI site in the middle of the polylinker, and should be cited as: Keily Littlefield and Darerca Owen, unpublished.
pMAT11 was created by Sheena Gordon by inserting MBP to pHAT4 and should be cited as Sheena Gordon, unpublished

Reference for the original non-fusion and His-tag vectors is:
Peränen J., Rikkonen M.,Hyvönen M., Kääriäinen L. (1996)
T7 vectors with modified T7*lac* promoter for expression of proteins in *Escherichia coli*. Anal. Biochem., 236:371-373

For more information see http://www-cryst.bioc.cam.ac.uk/~marko/vector, or email marko@cryst.bioc.cam.ac.uk

Notes:
All the data on this page is correct to the best of our knowledge, yet you should use it at your own risk.
Restriction sites indicated in paranthesis cut the vector twice, but are thought to be useful in some instances and shown to your benefit.
Plasmid map on page 2 was orginally created using Plasmapper web server at http://wishart.biology.ualberta.ca/PlasMapper/ and edited by hand.

Marko Hyvonen, University of Cambridge


T7 promoter
His-tag
BglII 485
Maltose binding protein (MBP)
Thrombin cleavage site 1268
BamHI 1281
NcoI 1285
EcoRI 1296
AccI 1303
AvaI 1308
XhoI 1308
AvrI 1308
HindIII 1347
T7 term
PstI 1835
NarI 1995
KpnI 3072
NdeI 3370
amp prom
amp marker
pBR322 origin
SP6 promoter
lacI reg
pMAT10
5616bp
1
500bp
1000bp
1500bp
2000bp
2500bp
3000bp
3500bp
4000bp
4500bp
5000bp
5500bp

11200 Rockville Pike
Suite 302
Rockville, Maryland 20852

August 19, 2011

American Society for Biochemistry and Molecular Biology

To whom it may concern,

It is the policy of the American Society for Biochemistry and Molecular Biology to allow reuse of any material published in its journals (the Journal of Biological Chemistry, Molecular & Cellular Proteomics and the Journal of Lipid Research) in a thesis or dissertation at no cost and with no explicit permission needed. Please see our copyright permissions page on the journal site for more information.

Best wishes,

Sarah Crespi

American Society for Biochemistry and Molecular Biology
11200 Rockville Pike, Rockville, MD
Suite 302
240-283-6616
JBC | MCP | JLR

Tel: 240-283-6600 • Fax: 240-881-2080 • E-mail: asbmb@asbmb.org

Bibliography

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr*. *D Biol. Crystallogr*. **66**, 213-221

Afonine, P. V., Grosse-Kunstleve, R. W., and Adams, P. D. (2005) A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr D. Biol. Crystallogr*. **61**, 850-855

Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Headd, J. J., Moriarty, N. W., Richardson, J. S., Richardson, D. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2010) *J. Appl Crystallogr*. **43**, 669-676

Agari, Y., Yokoyama, S., Kuramitsu, S., and Shinkai, A. (2008) X-ray crystal structure of a CRISPR-associated protein, Cse2, from *Thermus thermophilus* HB8. *Proteins* **73**, 1063-1067

Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S., and Shinkai, A. (2010) Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. *J. Mol. Biol.* **395**, 270-281

An, G., Justesen, J., Watson, R. J., and Friesen, J. D. (1979) Cloning the spot gene of *Escherichia coli:* identification of the spot gene product. *J. Bacteriol*. **137**, 1100-1110

Anjem, A., Varghese, S., and Imlay, J. A. (2009) Manganese import is a key element of the OxyR response to hydrogen peroxide in *Escherichia coli*. *Mol. Microbiol*. **72**, 844-858

Aravind, L., and Koonin, E. V. (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci*. **23**,469-472

Armon, A., Graur, D., and Ben-Tal, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447-463

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

Bailey, S. (2013) The Cmr complex: an RNA-guided endoribonuclease. *BIochem. Soc. Transac.* **41**, 1464-1467

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007) CRISPR provides acquired resistance against

viruses in prokaryotes. *Science* **315**, 1709-1712

Beese, L. S., and Steitz, T. A. (1991) Structural basis for the 3'-5' exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. *EMBO J.* **10**, 25-33

Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A. M., and Savchenko, A. (2008) A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283**, 20361-20371

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

Breitbart, M., and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends in Microb.* **13**, 278-284

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V., and van der Oost, J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964

Cady, K. C., and O'Toole, G. A. (2011) Non-identity-mediated CRISPR-bacteriophage

interaction mediated via the Csy and Cas3 proteins. *J. Bacteriol*. **193**, 3333-3445

Cai, L. (2001) Topological testing of the mechanism of homology search promoted by RecA protein. *Nucleic Acids Res.* **29**, 1389-1398

Carthew, R. W., and Sontheimer, E. J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642-655

Churchill, M. E. A., Klass, J., and Zoetewey, D. L. (2010) Structural analysis of HMGD-DNA complexes reveals influence of intercalation on sequence selectivity and DNA bending. *J. Mol. Biol.* **403**, 88-102

Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K., Semenova, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **945**, 1-7

Delano, W. L. (2010) The PyMOL Molecular Graphics System, Version 1.3. Schrödinger, LLC, New York

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

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

Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J., and Mojica, J. M. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*. **10**, 792-802

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

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

Firczuk, M., Wojciechowski, M., Czapinska, H., and Bochtler, M. (2011) DNA intercalation without flipping in the specific ThaI-DNA complex. *Nuc. Aci. Res.* **29**, 744-754

Fineran P. C., Gerritzen, M. J. H., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S. A. F. T., Staals, R. H. J., and Brouns, S. J. J. (2014) Degeenrate target sites mediate

rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E1629-1638

Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R., and Steitz, T. A. (1988) Cocrystal structure of an editing complex of Klenow fragment with DNA. *Proc. Natl. Acad. Sci. U.S.A*. **85**, 8924-8928

, A. H., and Moineau, S. (2010) *Nature* **468**, 67-71

Graham, A. I., Hunt, S., Stokes, S. L., Bramall, N., Bunch, J., Cox, A. G., McLeod, C. W., and Poole, R. K. (2009) Severe zinc depletion of *Escherichia coli*: Roles for high affinity zinc binding by ZinT, zinc transport and zinc-independent proteins. *J. Biol. Chem.* **284**, 18377-18389

Goddard, T. D., Huang, C. C., and Ferrin, T. E. (2007) Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157**, 281-287

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

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M.,

and Terns, M. P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956

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

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

Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J. A. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355-1358

Hochstrasser, M. L., Taylor, D. W., Bhat, P., Guegler, C. K., Sternberg, S. H., Nogales, E., and Doudna, J. A. (2014) CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci. USA*. Online: www.pnas.org/cgi/doi/10.1073/pnas.1405079111

Hogg, T., Mechold, U., Malke, H., Cashel, M., and Hilgenfeld, R. (2004) Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response. *Cell* **117**, 57-68

Holm, L., and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*. **233**, 123-138

Horton, N. C., and Finzel, B. C. (1996) The structure of an RNA/DNA hybrid: A substrate of the ribonuclease activity of HIV-1 Reverse transcriptase. *J. Mol. Biol.* **264**, 521-533

Howard, J. A., Delmas, S., Ivančić-Baće, I., and Bolt, E. L. (2011) Helicase dissociation and annealing of RNA-DNA hybrids by Escherichia coli Cas3 protein. *Biochem. J.* **439**, 85-95

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

Jansen, R., Embden, J. D., Gaastra, W., and Schouls, L. M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. A. (2013) RNA-

programmed genome editing in human cells. *Elife* **2**, e00471

Jore, M. M., Lundgren, M., van Duijn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., Beijer, M. R., Barendregt, A., Zhou, K., Snijders, A. P., Dickman, M. J., Doudna, J. A., Boekema, E. J., Heck, A. J., van der Oost, J., and Brouns, S. J. (2011) *Nat. Struct. Mol. Biol*. **18**, 529-536

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., and Crook, D. W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Micribiol. Rev.* **33**, 376-393

Juranek, S., Eban, T., Altuvia, Y., Brown, M., Morozov, P., Tuschl, T., and Margalit, H. (2012) A genome-wide view of the expression and processing patterns of *Thermus thermophilus* HB8 CRISPR RNAs. *RNA* **4**, 783-794

Kabsch, W. (2010) XDS. *Acta Crystallogr. D. Biol. Crystallogr*. **D66**, 125-132

Kiianitsa, K., Solinger, J. A., and Heyer, W. D. (2003) NADH-coupled microplate photometric assay for kinetic studies of ATP-hydrolyzing enzymes with low and high specific activities. *Anal. Biochem.* **321**, 266-271

Koo, Y., Ka, D., Kim, E., Suh, N., and Bae, E. (2013) Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial

immune system. *J. Mol. Biol.* **425**, 3799-3810

Koonin, E. V., and Wolf, Y. I. (2008) Genomis of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nuc. Aci. Res.* **36**, 6688-6719

Kondo, N., Nakagawa, N., Ebihara, A., Chen, L., Liu, Z. J., Wang, B. C., Yokoyama, S., Kuramitsu, S., and Masui, R. (2007) *Acta Crystallogr. D. Biol. Crystallogr*. **63**, 230-239

Kondo, N., Nishikubo, T., Wakamatsu, T., Ishikawa, H., Nakagawa, N., Kuramitsu, S., and Masui, R. (2008) Insights into different dependence of dNTP triphosphohydrolase on metal ion species from intracellular ion concentrations in *Thermus thermophilus. Extremophiles* **12**, 217-223

Lintner, N. G., Kerou, M., Brumfield, S. K., Graham, S., Liu, H., Naismith, J. H., Sdano, M., Peng, N., She, Q., Copié, V., Young, M. J., White, M. F., and Lawrence, M. (2011) Structural and functional characterization of an archaeal Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J. Biol. Chem.* **286**, 21643-21656

Lo, M. C., Aulabaugh, A., Jin, G., Cowling, R., Bard, J., Malamas, M., and Ellestad, G. (2004) Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal. Biochem*. **332**, 153-159

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708-720

Macomber, L., Elsey, S. P., and Hausinger, R. P. (2011) Fructose-1,6-bisphosphate aldolase (class II) is the primary site of nickel toxicity in *Escherichia coli*. *Mol. Microbiol.* **82**, 1291-1300

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., Yakunin, A. F., van der Oost, J., and Koonin, E. V. (2011*) Nat. Rev. Microbiol*. **9**, 467-477

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

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

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

Marraffini, L. A., and Sontheimer, E. J. (2010) CRISPR interference: RNA-directed

adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181-190

Matulis, D., Kranz, J. K., Salemme, F. R., and Todd, M. J. (2005) Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using ThermoFluor. *Biochemistry* **44**, 5258-5266

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

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

Mojica, F. J., Díez-Villaseñor, C., Soria, E., Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36**, 244-246

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

Mulepati, S., Orr, A., and Bailey, S. (2012) Crystal structure of the largest subunit of a

bacterial RNA-guided immune complex and its role in DNA target binding. *J. Biol. Chem.* **287**, 22445-22449

Mulepati, S., and Bailey, S. (2013) *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.* **288**, 22184-22192

Nakata, A., Amemura, M., and Makino, K. (1989) Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *J. Bacteriol.* **171**, 3553-3556

Nam, K. H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M. P., and Ke, A. (2012) Double-stranded endonuclease activity in *Bacillus halodurans* Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943-35952

Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M. P., and Ke, A. (2012) Cas5d protein processes pre-crRNA and assembles into a Cascade-like Interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* **20**, 1574-1584

Nam, K. H., Huang, Q., and Ke, A. (2012) Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. *FEBS Lett.* **586**, 3956-3961

Neidhardt, F. C., Bloch, P. L., and Smith, D. F. (1974) Culture medium for enterobacteria. *J Bacteriol.* **119**, 736-747

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 1-15

Otwinowski, Z., and Minor, W. (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode. *Method Enzymol.* **276**, 307-326

Pantoliano, M. W., Petrella, E. C., Kwasnoski, J. D., Lobanov, V. S., Myslik, J., Graf, E., Carver, T., Asel, E., Springer, B. A., Lane, P., and Salemme, F. R. (2001) High-density miniaturized thermal shift assays as a general strategy for drug discovery. *J. Biomol. Screen* **6**, 429-440

Peränen, J., Rikkonen, M., Hyvönen, M., and Kääriäinen, L. (1996) T7 vectors with a modified T7lac promoter for expression of proteins in *Escherichia coli*. *Anal. Biochem.* **236**, 371-373

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K. H., Bernardis, S., Ke, A., and DeLisa, M. P. (2010) Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol. Microbio.* **79**, 584-599

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014) *In vitro* assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res.* Epub.

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

Proudfoot, M., Kuznetsova, E., Brown, G., Rao, N. N., Kitagawa, M., Mori, H., Savchenko, A., and Yakunin, A. F. (2004) General enzymatic screens identify three new nucleotidases in *Escherichia coli*. Biochemical characterization of SurE, YfbR, and YjjG. *J. Biol. Chem*. **279**, 54687-54694

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

Read, R.J. (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst.* **A42**, 140-149

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C. V., Spagnolo, L., and White, W. F. (2013) Structure of the

CRISPR interference complex CSM reveals key similarities with Cascade. *Mol. Cell* **52**, 124-134

Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725-738

Seto, D., Bhatnagar, S. K., and Bessman, M. J. (1988) The purification and properties of deoxyguanosine triphosphate triphosphohydrolase from *Escherichia coli*. *J. Biol. Chem.* **263**, 1494-1499

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

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

Sheldrick, G. M. (2008) A short history of *SHELX. Acta Crystallogr. A.* **64**, 112-122

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

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

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

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

Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genetics,* **26**, 335-340

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

Swarts, D. C., Mosterd, C., van Passel, M. W., and Brouns, S. J. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**, e35888

Terwilliger, T. (2004) SOLVE and RESOLVE: Automated structure solution, density

modification, and model building. *J. Synchrotron. Radiat*. **11**, 49-52

Vedadi, M., Niesen, F. H., Allali-Hassani, A., Fedorov, O. Y., Finerty, P. J., Jr., Wasney, G. A., Yeung, R., Arrowsmith, C., Ball, L. J., Berglund, H., Hui, R., Marsden, B. D., Nordlund, P., Sundstrom, M., Weigelt, J., and Edwards,A. M. (2006) Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. U.S.A*. **103**, 15835-15840

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

Westra, E. R., van Erp, P. B., Künne, T., Wong, S. P., Staals, R. H. J., Seegers, C. L., Bollen, S., Jore, M. M., Semenova, S., Severinov, K., de Vos, W. M., Dame, R. R., de Vries, R., Brouns, S. J., and van der Oost, J. (2012) CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell*. **46**, 595-605

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

Wiedenheft, B., van Duijn, E., Bultema, J. B., Waghmare, S. P., Zhou, K., Barendregt,

A., Westphal, W., Heck, A. J., Boekema, E. J., Dickman, M. J., and Doudna, J. A. (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. USA*. **108**, 10092-10097

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

Wigley, D. B. (2013) Bacterial DNA repair. Recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat. Rev. Microbiol.* **11**, 9-13

Yang, W., Lee, J. Y., and Nowotny, M. (2006) Making and breaking nucleic acids: two-Mg^{2+}-ion catalysis and substrate specificity. *Mol. Cell* **22**, 5-13

Yakunin, A. F., Proudfoot, M., Kuznetsova, E., Savchenko, A., Brown, G., Arrowsmith, C. H., and Edwards, A. M. (2004) The HD domain of the *Escherichia coli* tRNA nucleotidyltransferase has 2', 3'-cyclic phosphodiesterase, 2'-nucleotidase, and phosphatase activities. *J. Biol. Chem*. **279**, 36819-36827

Yosef, I., Goren, M. G., Kiro, R., Edgar, R., and Qimron, U. (2011) High-temperature protein G is essential for acticity of the *Escherichia coli* clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20136-20141

Yosef, I., Goren, M. G., Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escerichia coli*. *Nucleic Acids Res.* **12**, 5569-5576

Yosef, I., Shitrit, D., Goren, M. G., Burstein, D., Pupko, T., and Qimron, U. (2013) DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl. Acad. Sci.* **110**, 14396-14401

Young, J. C., Dill, B. D., Pan, C., Hettich, R. L., Banfield, J. F., Shah, M., Fremaux, C., Horvath, P., Barrangou, R., and Verberkmoes, N. C. (2012) Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in *Streptococcus thermophilus. PLoS One* **5**, e38077

Zimmerman, M. D., Proudfoot, M., Yakunin, A., and Minor, W. (2008) Structural insight into the mechanism of substrate specificity and catalytic activity of an HD-domain phosphohydrolase: the 5'-deoxyribonucleotidase YfbR from *Escherichia coli*. *J. Mol. Biol*. **378**, 215-226

Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G. Huanting, L., Sonja-Verena, A., Naismith, J. H., Spagnolo, L., and White, M. F. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol Cell* **45**, 303-313

Curriculum Vitae Sabin Mulepati

BIRTH

Kathmandu, Nepal May 31, 1985

EDUCATION

Johns Hopkins University, Baltimore, MD 2008-2014
Ph.D., Biophysics
Thesis Advisor: Scott Bailey

Susquehanna University, Selinsgrove, PA 2004-2008
B. S., Biochemistry
Magna cum laude, University and Departmental Honors

HONORS AND AWARDS

The Owens Scholars Graduate Fellowship 2008-2011
Johns Hopkins University
The Presser Full Undergraduate Scholarship 2004-2008
Susquehanna University
Amgen Scholar 2006
Best Undergraduate Chemistry Research Award 2008
Gamma Sigma Epsilon Chemistry Honor Society 2007
Susquehanna University Presidential Fellow 2005

PUBLICATIONS

Mulepati, S., Mathews, I. I., and Bailey, S. (2014). Crystal structure of the type-I Cascade complex from *Escherichia coli* bound to its target DNA. *Manuscript in preparation.*

Mulepati, S. (2014). Mechanism of target DNA recognition and degradation by the type I CRISPR system. *Ph.D. thesis in preparation.*

Mulepati, S., and Bailey, S. (2013). *In vitro* reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem*. **287**, 22184-22192

Mulepati, S., Orr, A., and Bailey, S. (2012). Crystal structure of the largest subunit of a bacterial RNA-guided immune complex and its role in DNA target binding. *J. Biol. Chem.* **287**, 22445-22449

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

Kavran, J. M., Ward, M. D., Oladosu, O. O., **Mulepati, S.**, and Leahy, D. J. (2010). All mammalian hedgehog proteins interact with cell-adhesion molecule, down-regulated by oncogenes (CDO) and brother of CDO (BOC) in a conserved manner. *J. Biol. Chem.* **285**, 24584-24590