University of Alberta

Structural and Functional Characterization of *T. thermophilus* CasE

by

Emily Martha Gesner

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

Spring 2011
Edmonton, Alberta

Dedication

This thesis is dedicated to George, the love of my life and my best friend, without whom I could never have completed this task. I wish to express my deepest gratitude for your unconditional love and support, and the many sacrifices you made throughout my graduate program. Thank you, it has meant the world to me.

Abstract:

Powerful mechanisms of genetic interference in both unicellular and multicellular organisms are based on the sequence-directed targeting of DNA or RNA by small effector RNAs. In many bacteria and almost all archaea, RNAs derived from clustered, regularly interspaced, short palindromic repeat (CRISPR) loci are involved in an adaptable and heritable gene-silencing pathway. Resistance to phage infection is conferred by the incorporation of short invading DNA sequences into the prokaryotic genome as CRISPR spacer elements separated by short repeat sequences. A central aspect to this pathway is the processing of a long primary transcript (pre-crRNA) containing these repeats by crRNA endonucleases to generate the mature effector RNAs that interfere with phage or plasmid gene expression. Here we describe a structural and functional analysis of the CasE endonuclease of *T. thermophilus* a member of the Ecoli CRISPR sub-type. High resolution X-ray structures of CasE bound to repeat RNAs model both the pre-and post-cleavage complexes associated with processing the pre-crRNA. These structures establish the molecular basis of a specific CRISPR RNA recognition and suggest the mechanism for generation of effector RNAs responsible for gene-silencing.

Acknowledgments

First and foremost, I wish to thank George Johnson, to whom this thesis is dedicated. I also wish to express my tremendous gratitude to my parents and grandparents, Garret and Deborah Gesner, and John and Marion Nicholson who supported me emotionally and financially throughout my graduate program. I also wish to thank Mary Ann and David Johnson for their generous support over the years.

I would like to acknowledge the support of past and present MacMillan laboratory members including: Oliver Kent, Stephen Chaulk, Kaari Lynch, Dustin Ritchie, Matthew Schellenberg, Edyta Sieminska, Tao Wu, Erin Garside, Karim Atta, and Mark George. I would like to thank Edan Foley, Andrew Simmonds, and Sarah Hughes for their useful advice and insight. I would like to acknowledge members of the technical staff for their assistance including Tracy Sawchuk, Lillian Cook, and Troy Locke. I would like to thank Susan Smith, Kim Arndt, and Marion Benedict for their tremendous administrative support. Finally, I must thank my supervisor, Dr Andrew MacMillan and my committee members, Dr Chris Bleackley and Dr Mark Glover for their support and advice throughout my graduate career.

Table of Contents:

List of Tables:

List of Figures:

List of Abbreviations:

A	Adenosine
aa	Amino acid
ATP	Adenosine triphosphate
bp	Base-pair
C	Cytosine
Cas	CRISPR Associated
Cascade	CRISPR Associated Complex for Anti-Viral Defence
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
DNA	Deoxyribonucleic Acid
ds	Double Stranded
DTT	Dithiolthreitol
EDTA	Ethylene diamine tetraacetic acid
G	Guanidine
K_d	Dissociation constant
K_M	Michaelis constant - inverse of enzyme affinity
LGT	Lateral Gene Transfer
mRNA	Messenger RNA
nt	Nucleotide
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PDB	Protein Data Bank
RAMP	Repeat Associated Mysterious Protein
RRM	RNA recognition motif
RNA	Ribonucleic Acid
RNAi	RNA interference
SDS	Sodium dodecyl sulfate
Spoligotyping	Spacer oligonucleotide typing
ss	Single Stranded
TEV	Tobacco etch virus
T	Thymidine
U	Uridine
V_0	Initial reaction rate
V_{max}	Maximum reaction rate
WT	Wild-type

Chapter 1
The CRISPR-Cas Genetic Defence System

1-1. Prokaryotic Lateral Gene Transfer

The transfer of genetic material between two organisms that are not parent and progeny is referred to as lateral gene transfer (LGT). This genetic transfer can occur through the uptake of free DNA from the surrounding environment (transformation), the transfer of genetic material from conjugative plasmids (conjugation), or bacteriophage infection (transduction) (Thomas & Nielsen, 2005). Bacterial genomes contain a significant amount of DNA derived from LGT attesting to the process's significant contribution to prokaryotic evolution (Nakamura *et al.*, 2004). For example, sharing plasmids allows for rapid adaptation and the spread of pathogenicity and antibiotic resistance. The majority of transferred genes, however, do not hold an immediate selective advantage. A balance must be struck between the permission and defence against gene uptake to continually allow diversification and adaptation while maintaining genetic integrity (Goodier & Kazazian, 2008; Keeling & Palmer, 2008; Koonin & Wolf, 2008; Schaack *et al.*, 2010; Wozniak & Waldor, 2010). Mechanisms have thus evolved to defend against invasive genetic elements such as the restriction/modification system (Comeau & Krisch, 2005; Heidelberg *et al.*, 2009; Thomas & Nielsen, 2005).

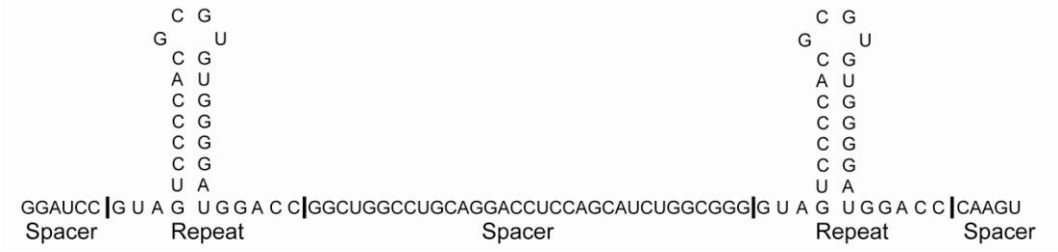

Figure 1-1. CRISPR loci schematic. The 28 nucleotide repeat sequence (example from *T. thermophilus* HB8 cluster 7) with a 33 nucleotide spacer sequence. CRISPR loci are characterized by 24-47 nucleotide palindromic repeat sequences interspaced by phage/plasmid derived 26-70 nucleotide spacer sequences. They are transcribed into long RNAs that are not translated, the palindromic repeats often fold into RNA hairpins.

Recently, a new prokaryotic genetic defence mechanism, the CRISPR-Cas system, has been discovered. CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) are genetic loci that contain variable spacer sequences separated by conserved repeat sequences. The loci are transcribed as a long precursor CRISPR RNA (pre-crRNA) and processed into short crRNAs containing a single spacer sequence (Brouns *et al*., 2008; Carte *et al*., 2008; Haurwitz *et al.*, 2010). These effector RNAs contain a single spacer sequence homologous to phage or plasmid DNA which act as guides to identify and target invasive genetic elements for degradation (Horvath & Barrangou, 2010; Karginov & Hannon, 2010; Sorek *et al.*, 2008). This adaptable system provides acquired, heritable immunity to phage or plasmids.

1-2. Discovery of the CRISPR-Cas System

CRISPRs were first observed in *E. coli* K12 by Ishino and colleagues in the late 1980s as a highly conserved 29 base-pair (bp) inverted repeat sequence at intervals of 32-33 bp (Figure 1-1; Ishino *et al.*, 1987; Mojica *et al.*, 2005; Nakata *et al.*, 1989). Upon further investigation, this peculiar repeat sequence was also found in the genomes of *S. dysenteria* and *S. typhimurium*. Presently, these loci have been identified in over 90% of archaea and 40% of bacteria (Grissa *et al.*, 2007; Nakata *et al.*, 1989). In 2002, Jansen and colleagues coined the term CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) to describe these loci which were, at the time, believed to be involved in DNA repair (Godde & Bickerton, 2006; Jansen *et al.*, 2002; Lillestol *et al.*, 2009; Mojica *et al.*, 1995).

The observation that many of the sequenced spacers from bacterial genomes were homologous to viral and plasmid sequences gave rise to the hypothesis that the CRISPR system was involved in genetic defence (Bolotin *et al.*, 2005; Mojica *et al.*, 2005). Indeed, bacterial strains containing spacers homologous to phage and plasmid DNA cannot be invaded by the respective phage or plasmid suggesting that the spacers are directly involved in protection against mobile genetic elements (Barrangou *et al.*, 2007; Marraffini & Sontheimer, 2008; Tyson & Banfield, 2008). After phage infection, 1-3 new spacers are integrated into the CRISPR locus upstream of the existing spacers conferring resistance to the surviving fraction of the bacterium and its progeny.

In fact, the number of spacers a bacterium has against a phage is inversely correlated with phage sensitivity (Barrangou *et al.*, 2007; Pourcel *et al.*, 2005; van der Ploeg, 2009).

CRISPR loci are characterized by palindromic repeats ranging from 24 to 47 bp, often predicted to form an RNA hairpin when transcribed, separated by unique spacer sequences of 26 to 70 bp (Figure 1-1; Grissa *et al.*, 2007). A 150-550 bp conserved leader sequence upstream of the CRISPR loci contains a promoter region required for transcription of CRISPR RNAs (Pul *et al.*, 2010b). The long non-coding CRISPR transcript is quickly processed into short ~60 nucleotide (nt) effector RNAs containing a spacer sequence and a portion of the repeat on both the 5' and 3' ends (Brouns *et al*, 2008).

In prokaryotes, genes involved in the same biological process are often found in close proximity in the genome, typically as a part of an operon (Overbeek *et al.*, 1999). Therefore, it was reasonable to explore the role of the protein coding genes that surrounded CRISPR loci. A family of CRISPR-Associated (*cas*) genes are often found adjacent to CRISPR loci (< 1kb). These genes code for proteins that are predicted endonucleases, exonucleases, helicases, and nucleic binding proteins (Jansen *et al.*, 2002; Makarova *et al.*, 2006; Makarova *et al.*, 2009). The CRISPR loci and the associated *cas* genes make up the CRISPR-Cas system.

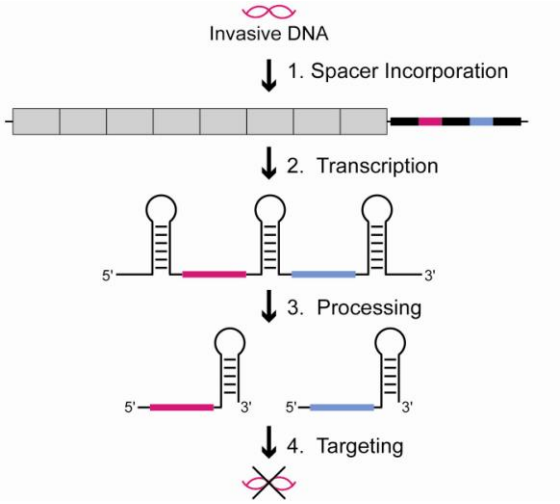

Figure 1-2. The CRISPR-Cas System. The critical steps of the CRISPR-Cas pathway are: (1) Spacer acquisition and integration into the CRISPR loci (new pink spacer), (2) Transcription and (3) processing of CRISPR transcript, and (4) Targeting of invasive genetic elements (pink). The CRISPR loci are made up of short palindromic repeat sequences (black boxes) separated by spacers (pink and blue boxes) derived from phage or plasmid DNA. CRISPR-associated (Cas) genes are in close proximity to the CRISPR loci (grey boxes).

1-3. CRISPR-Cas System Overview.

Our current understanding separates the essential phases of the CRISPR-Cas pathway into spacer acquisition, transcription and processing of effector crRNAs, and invasive element targeting (Figure 1-2; Marraffini & Sontheimer, 2010b). Foreign DNA is processed into short fragments that are incorporated into the CRISPR locus between repeat sequences (Horvath *et al*., 2008; Mojica *et al*., 2005; Van der Ploeg *et al*., 2009). This programs the host and its progeny to target homologous genetic elements upon future encounters. The CRISPR locus is transcribed into a long transcript that is processed into short effector crRNAs containing a single spacer sequence flanked by portions of the

repeat (Brouns *et al*., 2008; Carte *et al*., 2008; Haurwitz *et al*., 2010). Upon repeat exposure to the invasive element, these effector crRNAs may bind to homologous viral or plasmid DNA through base-pairing to mark it for degradation (Figure 1-2). The actions of the CRISPR system are facilitated by a highly variable combination of Cas proteins.

1-4. CRISPR-Associated Genes

CRISPR-associated genes are found in close proximity to CRISPR loci and code for proteins that are directly involved in the pathway (Haft *et al*., 2005; Makarova *et al*., 2002, Makarova *et al*., 2006). Six of these, the core *cas* genes (*cas*1-6) are most commonly found in association with CRISPR loci in CRISPR containing genomes (Makarova *et al.*, 2006). In addition, particular combinations of *cas* genes are often found together defined by a conserved genetic order and association with a specific repeat type (Haft *et al*., 2005). This has given rise to the proposed delineation of *cas* genes into eight subtypes (Table 1-1; Haft *et al.*, 2005; Kunin *et al.*, 2007).

1-4.1. Cas1

Of the six core *cas* genes, only Cas1 and Cas2 are found in all CRISPR systems (Haft *et al*., 2005; Makarova *et al*., 2006). They are proposed to be involved in the processing and acquisition of spacers as they are not necessary for phage resistance once a relevant spacer has been acquired (Brouns *et al.*, 2008;

Makarova *et al.*, 2006). The *P. aeruginosa* Cas1 homologue is a divalent metal-dependent, DNA-specific endonuclease that cleaves both dsDNA and ssDNA into ~80 base-pair fragments (Wiedenheft *et al.,* 2009). Unlike most metal-dependent endonucleases, the enzyme has a preference for Mn^{2+} over Mg^{2+}. In contrast, the *S. solfataricus* homologue possesses ss/dsDNA and ss/dsRNA binding and annealing activity, but does not have cleavage activity (Han *et al.*, 2009).

Recently, a role for *E. coli* Cas1 in DNA repair has been proposed. This homologue possesses Mg^{2+} dependent ssDNA, dsDNA and ssRNA endonuclease activity (Babu *et al*., 2011). The products of ssDNA and dsDNA cleavage are ~24 and ~30 nt respectively, which again is incompatible with the 32 nt spacer length found in *E. coli*. In addition, Cas1 has endonucleolytic activity against several DNA recombination and repair intermediates such as holliday junctions, replication forks and 5' flaps (Babu *et al*., 2011). Physical and genetic association with DNA repair machinery (RecB, RecC, and RuvB) suggest a connection with DNA repair. Furthermore, Cas1 mutants displayed increased sensitivity to DNA damage and impaired chromosomal segregation. This suggests that Cas1 not only plays a role in spacer integration, but is also important for DNA damage repair (Babu *et al*., 2011).

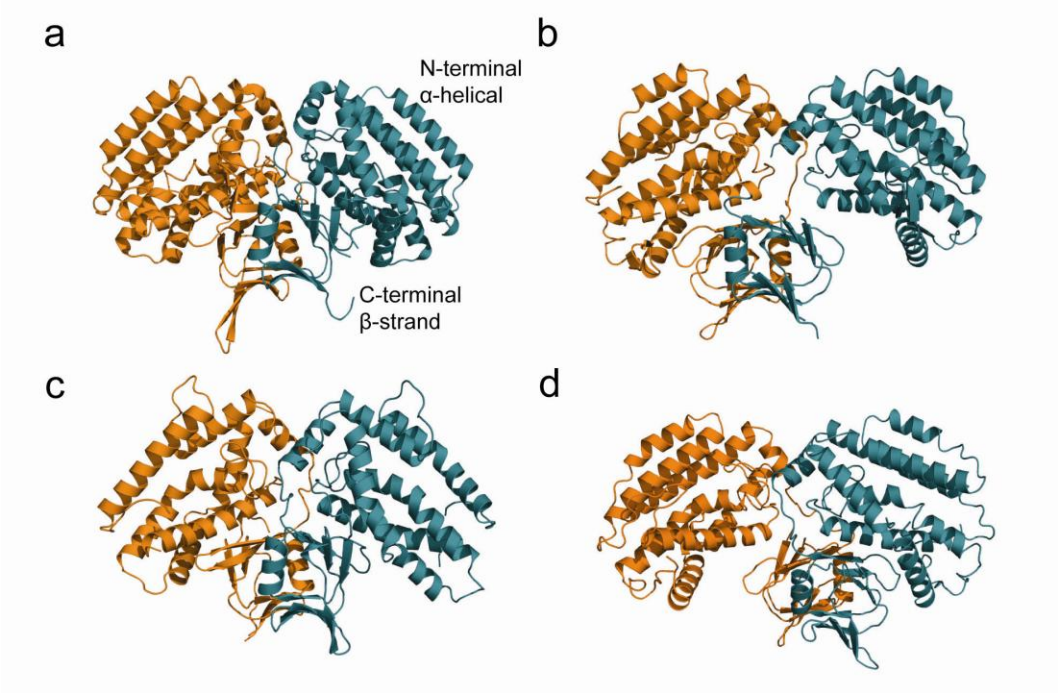

Figure 1-3. Crystal structures of Cas1 homologue. Ribbon diagram of Cas1 homologues aligned using iSuperpose and translated horizontally. They share a common N-terminal β-strand domain and a C-terminal α-helical domain which is shaped like a butterfly. (a) *P .aeruginosa* (PDB 3GOD; Wiedenhaft *et al*., 2009). (b) *A. aeolicus* (PDB 2YZS). (c) *E. coli* (PDB 3NKD; Babu *et al*., 2011). (d) *T. maritima* (PDB 3LFX).

There are four X-ray crystal structures deposited in the protein data bank (PDB) of the Cas1 protein homologues from *P. aeruginosa* (PDB 3GOD; Wiedenheft *et al.*, 2009), *A. aeolicus* (PDB 2YZS), *E. coli* K12 (PDB 3NKD; Babu *et al*., 2011), and *T. maritima* (PDB 3LFX) (Figure 1-3). These structures share a common dimeric fold resembling a butterfly consisting of an N-terminal β-strand domain and a C-terminal α-helical domain (the α-helical domain and β-strand domain making up the upper and lower lobes of each wing respectively). The face of the α-helical domain contains a conserved, solvent exposed, divalent

metal binding site (Glu/Asp, His, and Asp) on the face of the α-helical domain (Figure 1-4). Surrounding this metal binding site is a positively charged surface which favours nucleic acid binding. Several solvent exposed Arg and Lys residues in close proximity to the active site may serve to position the nucleic acid in a catalytically favourable conformation (Figure 1-4).

Unanswered questions remain regarding Cas1's role in the CRISPR pathway and in DNA repair. It is difficult to rationalize the length of *in vitro* cleavage products from *P. aeruginosa* and *E. coli* (~80 nt and 30 nt respectively) and the average spacer size of 32 nucleotides suggesting that one or more co-factors must be involved in the processing and integration stages of spacer acquisition (Wiedenheft *et al.*, 2009; Babu *et al.*, 2011). Because Cas1 acts on methylated and un-methylated DNA alike it is unclear how the enzyme distinguishes between self and non-self nucleic acid during spacer acquisition. The role of *E. coli* Cas1 in DNA repair pathways suggests that the CRISPR pathway is linked to DNA damage response. This is interesting because Cas genes were originally identified as a possible DNA repair system, however, the details of the relationship between the DNA repair system and the genetic interference system have yet to be described (Babu *et al.*, 2011; Makarova *et al.*, 2006; Makarova *et al.*, 2002).

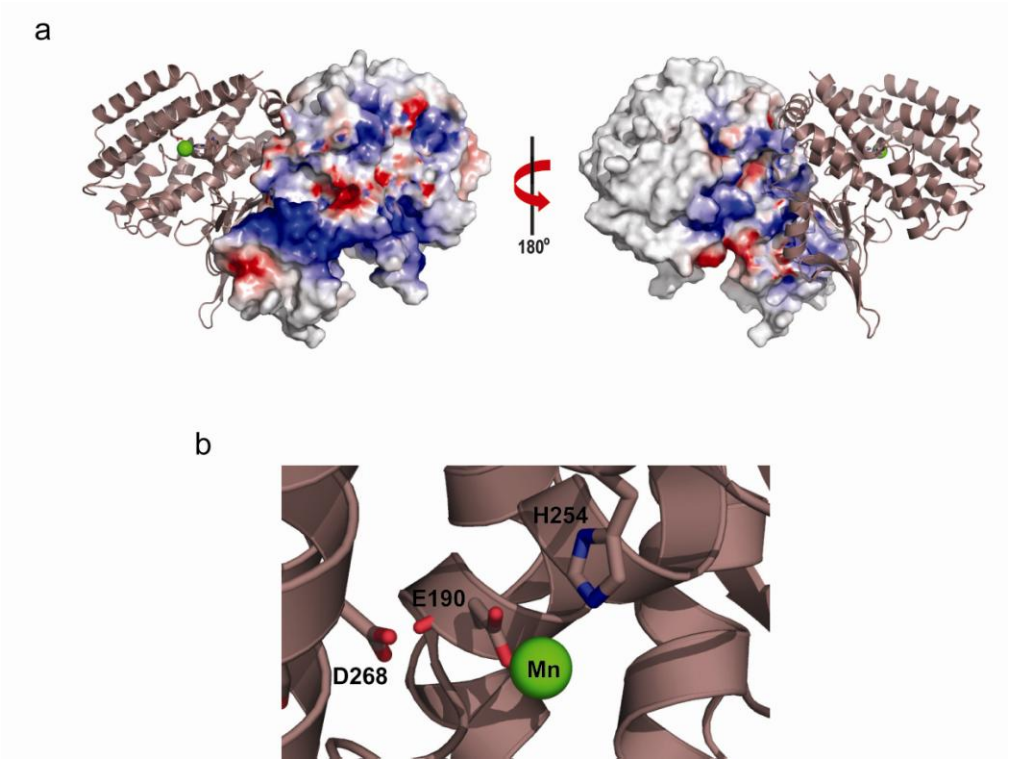

Figure 1-4. Mn^{2+} binding site of *P. aeruginosa* Cas1 homologue. (a) Two views of *P. aeruginosa* homodimer rotated by 180°. Electrostatic surface potential representation is shown on one monomer to display the basic patch and the active site residues and Mn^{2+} binding site are displayed on the ribbon diagram on the other monomer (PDB 3GOD). (b) Mn^{2+} binding site shown with conserved E190, H254, and D268 residues shown (Wiedenheft *et al.*, 2009).

1-4.2. Cas2

Cas2 is a predicted endonuclease found in all CRISPR systems, and like Cas1, is not required for phage resistance after spacer incorporation (Brouns *et al.*, 2008). Three X-ray crystal structures of Cas2 homologues have been deposited in the PDB: *S. solfataricus* (PDB 2I8E; Beloglazova *et al.*, 2008), *P. furiosus* (PDB 2IOX), and *T. thermophilus* (PDB 1ZPW) which share a common fold. In all three cases the structures are homodimeric ferredoxin-like folds with a dimeric

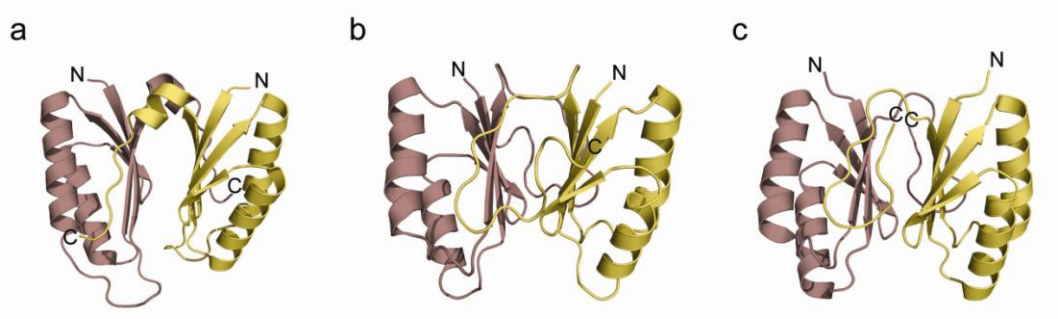

Figure 1-5. X-ray crystal structures of Cas2 homologues. Ribbon diagram of Cas2 homologues. Structures were superimposed using iSuperpose and translated laterally. X-ray crystal structures share a common homodimeric ferredoxin-like fold (shown in yellow and violet) to form two five-stranded antiparallel β-sheets: (a) *S. solfataricus* (PDB 2I8E), (b) – *P. furiosus* (PDB 2IOX), and (c) *T. thermophilus* (PDB 1ZPW).

interface formed between the two β-sheets (Figure 1-5). β5 of each monomer is exchanged with the other monomer forming two five-stranded, anti-parallel β-sheets.

In vitro experiments have shown that homologues from *S. solataricus*, *A. fulgidis, T. maritima, M. thermoautotrophicum,* and *N. europuea* possess Mg^{2+}-dependent ssRNA endonucleolytic activity with a preference for U-rich regions of RNA (Beloglazova *et al.*, 2008). The nuclease activity of the *T. thermophilus* and the *P. furiosus* homologues has not been tested. Although it is known that Cas2 does not play a role in the crRNA processing and DNA targeting steps, it remains unclear how Cas2 may function in the CRISPR-Cas system (Brouns *et al*., 2008).

1-4.3. Cas3

The Cas3 protein is required for phage resistance, most likely at the targeting stage (Brouns *et al.*, 2008). Its domain structure is comprised of a conserved C-terminal DEAD/DEAH box helicase domain fused to an N-terminal HD-nuclease domain. In some instances the domains have undergone a fission event to yield two proteins (Makarova *et al.*, 2006; Makarova *et al.*, 2009). The *S. solataricus* Cas3 homologue possesses Mg^{2+} dependent dsDNA and dsRNA hydrolysis activity with a preference for cleaving after G:C base-pairs (Han & Krauss, 2009). In this case, the Cas3 homologue contains only the HD-nuclease domain and the absence of the corresponding helicase domain containing protein may contribute to its lack of substrate specificity. Recently, in *S. thermophilus*, an Ecoli subtype Cas3 homologue has been shown to possess ssDNA-stimulated ATPase activity coupled to dsDNA and DNA/RNA duplex unwinding activity and also ssDNA nuclease activity, suggesting that it is involved in CRISPR-mediated DNA targeting (Sinkuna *et al*., 2011).

A 2.3 Å X-ray crystal structure of the HD-nuclease domain of Cas3 from *M. jannaschii* (Figure 1-6; PDB 3M5F) has been deposited in the protein data bank by the Midwest Protein Structure Initiative, but has not been published. The structure reveals a canonical HD-nuclease fold which is mostly α-helical (Figure 1-6 a,b). Although two calcium ions bound by conserved histidine and aspartate residues are modelled (H65, H90, H122, H123 and D66), there are issues with

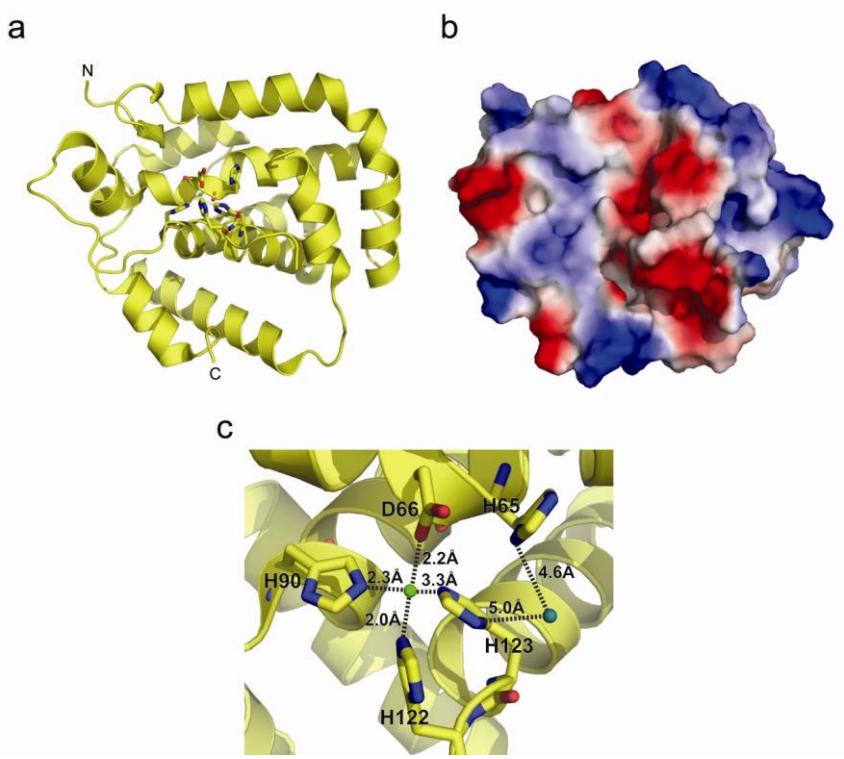

Figure 1-6. X-ray crystal structure of *M. jannaschii* Cas3 (PDB 3M5F). (a) Ribbon diagram of HD-nuclease domain of *M. jannaschii* Cas3, shown are the calcium ions in green/cyan bound by the conserved histidine and aspartate residues. (b) Electrostatic surface representation of the same view shown in (a). (c) Close up view of the active site residues showing the bound calcium ion A in green and the improperly modelled calcium ion B in cyan, histidine and aspartate residues are shown with hydrogen bonding distances indicated.

respect to refinement of the structure which suggest that only one of these calcium ions is actually present (Figure 1-6 c). At this resolution, identifying an atom as a calcium ion is somewhat ambiguous and there is no mention of using an anomalous scattering metal ion to confirm its placement. Furthermore, although calcium ion A has suitable proposed amino acid ligands (D66, H90, H122, H123), calcium ion B is located too distant from the proposed ligands (H65, H123), making this assignment unlikely (Figure 1-6c). A deficiency in the number of

water molecules modelled, 30, rather than the expected 200-300 of a protein of this size, further suggests that this structure is not well refined.

1-4.4. Cas4

The *cas4* gene encodes a predicted RecB-like nuclease containing three conserved C-terminal cysteines that may coordinate zinc (Haft *et al.*, 2005; Makarova *et al.*, 2006). RecB contains an ATP dependent 3'-5' helicase domain and a nuclease domain which, along with RecC and RecD makes up the Exonuclease V complex involved in the recombinational repair of dsDNA breaks (Kowalczykowski, 2000). Cas4 is predicted to contain a nuclease domain like RecB, but not the helicase domain. Because of its similarity to RecB, Cas4 has been speculated to be involved in spacer integration. Another possibility is that Cas4 may, like Cas1, be involved in DNA damage repair. Further molecular, biochemical and structural work is required to define the role of Cas4 in the CRISPR pathway.

1-4.5. Cas5 & Cas6

Makarova and colleagues have grouped Cas5 and Cas6 homologues into the RAMP family (Repeat Associated Mysterious Protein) (Makarova *et al.*, 2009). This family is a diverse class of predicted RNA binding proteins characterized by a C-terminal glycine-rich loop (Makarova *et al.*, 2009). The *cas5* genes are divided into eight genetically distinct subtypes in congruence with the subtypes as

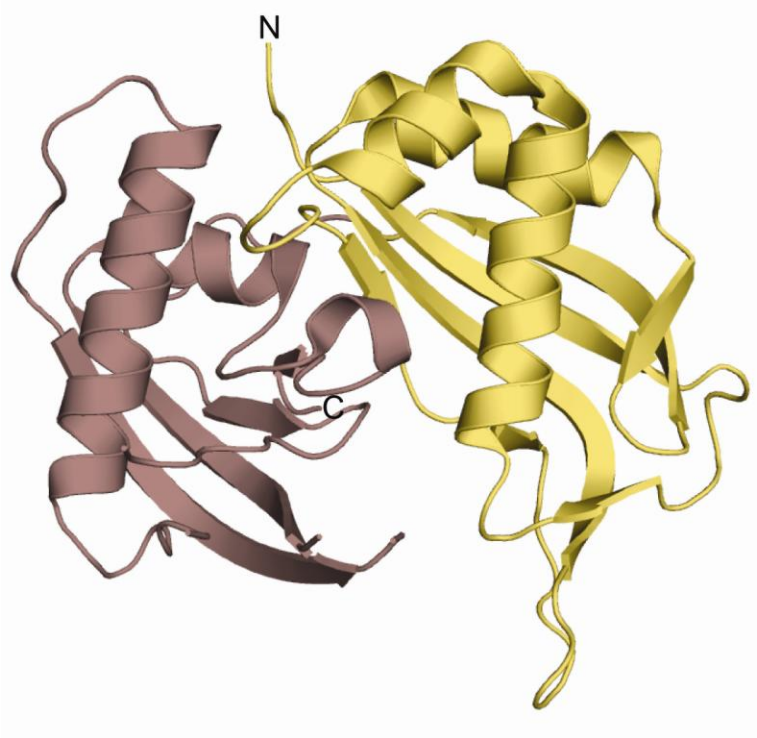

Figure 1-7. *P. furiosus* Cas6 X-ray crystal structure. Ribbon diagram of *P. furiosus* Cas6 (PDB 3I4H), which is comprised of two tandem ferredoxin-like folds (coloured yellow and violet) (Carte *et al.*, 2008). Cas6 is responsible for crRNA endonucleation in *P. furiosus*.

described below (i.e. Cas5a, Cas5d, Cas5n, Cas5y, Cas5h, Cas5t, and Cas5m). However, the role of Cas5 proteins in the CRISPR pathway is unknown.

The Terns laboratory has demonstrated that Cas6 is a sequence specific endoribonuclease responsible for processing the CRISPR transcript into effector crRNAs in *P. furiosus*, suggesting that the core Cas6 homologues are a conserved family of crRNA endonuclease (Carte *et al.*, 2008). The X-ray crystal structure reveals two tandem ferredoxin-like folds that contain the typical βαββαβ fold

arranged so that the β-sheets form an interface buttressed by two α-helices on either side (Figure 1-7; PDB 3I4H; Carte *et al.,* 2008). The Cas6 protein and its structure will be discussed in greater detail in Chapters 3 and 4.

1-4.6. Cas Subtypes

Besides the core genes, there are eight *cas* subtypes named after the species in which they were initially identified: Ecoli, Ypest, Nmeni, Apern, Dvulg, Hmari, Tneap, and Mtube (Table 1-1; Haft *et al.*, 2005). Specific subtypes of CRISPR systems are defined by the structure and organisation of repeats as well as by the particular set of accompanying protein coding *cas* genes (Haft *et al.*, 2005; Kunin *et al.*, 2007). For example, the Ecoli subtype, discussed in detail later, consists of the core *cas*1-3 genes and the subtype specific *cse*1-4 and *cas*5e genes. In contrast, the Dvulg subtype consists of the core *cas*1-4 genes and the subtype specific *csd*1-2 and *cas*5d genes.

The members of Cmr, a ninth subtype of *cas* genes (also known as RAMP) is comprised of RAMP domain containing proteins characterized by a conserved glycine-rich loop and predicted nucleic acid binding activity (Makarova *et al.,* 2006). The Cmr subtype is considered to be evolutionarily distinct from the rest of the Cas subtypes and is more closely associated with thermophilic bacteria and crenarchaea. This subtype is not always found in close proximity to a CRISPR locus and is never found in a prokaryotic genome in the

absence of other *cas* subtypes suggesting that it is not sufficient to provide resistance alone (Haft *et al.*, 2005; Makarova *et al*., 2006).

Table 1-1. CRISPR-Cas Subtypes with associated Core and Subtype-specific *cas* Genes

Subtype	Reference Organism	Core *cas* Genes	Subtype-specific *cas* Genes
Ecoli	*Escherichia coli*	Cas 1-3	cse 1-4, cas5e
Ypest	*Yersinia pestis*	Cas 1-3	csy 1-4
Nmeni	*Neisseria meningitidis*	Cas 1-2	csn1-2
Dvulg	*Desulfovibrio vulgaris*	Cas 1-4	csd 1-2, cas5d
Tneap	*Thermatoga neapolitana*	Cas 1-4, 6	cst1-2, cas5t
Hmari	*Haloarcula marismortui*	Cas 1-4, 6	csh 1-2, cas5h
Apern	*Aeropyrum pernix*	Cas 1-6	csa1-5
Mtube	*Mycobacterium tuberculosis*	Cas 1, 2, 6	csm1-5
Cmr	N/A	None	cmr1-6

*Adapted from Marraffini and Sontheimer, 2010

1-5. CRISPR-Cas Mobility

Intriguingly, both the CRISPR loci and the associated *cas* genes themselves are mobile genetic elements (Godde & Bickerton, 2006; Portillo & Gonzalez, 2009). The *cas* genes are found in a myriad of combinations amongst prokaryotes. Subtypes are based on similarities within cas genes rather than among phylogentically related organisms (Haft *et al*., 2005). Different *cas* gene subtypes can be found in the same species for example, *T. thermophilus* HB8 contains Ecoli, Mtube, and Cmr subtypes (Haft *et al*., 2005). Furthermore, the same subtypes can be present in phylogenetically distant species, for example, the Ecoli subtype is found in *C. diptheriae* (phylum Actinobacteria), *C. tepidium* (phylum Chlorobia), *T. thermophilus* (phylum Deinococcus thermus), and *E. coli* (phylum Proteobacteria) (Haft *et al.*, 2005). The CRISPR-Cas systems tend to be

localized to variable chromosomal regions with diverse gene combinations suggesting that they undergo displacement by genomic shuffling mechanisms such as transposition and are readily exchanged probably via conjugative plasmids or chromosomal conjugation (Horvath *et al.*, 2008; Portillo & Gonzalez, 2009). This genetic mobility has contributed to the diversification of the CRISPR system across phylogenetic boundaries (Haft *et al*., 2005; Makarova *et al.*, 2006).

1-6. CRISPR transcript

CRISPR transcription is directed by a leader sequence upstream of the CRISPR loci which has been shown to contain a promoter in some species (Pul *et al.*, 2010b; Viswanathan *et al.*, 2007). In *E. coli* K12, CRISPR transcription is repressed by heat-stable nucleoid structuring protein (H-NS), a global repressor of transcription often associated with genes of lateral origin (Pul *et al*., 2010; Westra *et al*., 2010). As the loci and associated genes are indeed of lateral origin, it is not surprising that the CRISPR-Cas system is regulated in this manner. This regulation can be relieved through LeuO expression, a global activator of transcription implicated in the stress response which was initially identified as a transcriptional activator of the Leucine synthesis operon (Pougach *et al.*, 2010; Pul *et al.*, 2010b; Westra *et al.*, 2010). This suggests that in *E. coli*, CRISPR-Cas system activation is coupled to the stress response (Pul *et al*., 2010; Westra *et al*., 2010). Protection provided by CRISPR-Cas system activation may increase

survival in stressful environments such as nutrient deficiencies, when bacteria are more vulnerable to phage infection.

Table 1-2. Association of Repeat Clusters with Cas Subtypes

Subtype	Repeat Cluster	Hairpin Folding Score
Ecoli	2	Strong
Ypest	4	Intermediate
Nmeni	10	Poor
Dvulg	3	Strong
Tneap	1, 6	Poor
Hmari	1, 9	Poor
Apern	6, 7, 11	Poor
Mtube	1, 6, 8, 2	Poor
Cmr	N/A	N/A

It was originally proposed that the repeats found in CRISPR loci were greatly divergent and strain specific (Makarova *et al*., 2006). A genetic analysis has shown, however, that these repeats can be grouped into twelve clusters based on sequence and secondary structure similarity (Table 1-2; Kunin *et al.*, 2007). The strength of hairpin folding scores was compared between the clusters; clusters 2 and 3 are predicted to form a strong hairpin; 4, 5, 8, and 12 are predicted to form weak hairpins; and 1, 6, 7, 9, 10, and 11 are likely single-stranded (Table 1-2; Figure 1-8; Kunin *et al.*, 2007). Co-variation of sequence in clusters with strong hairpin folding scores maintains secondary structure. For example, in cluster 2, G16 is predicted to base-pair with C23, but in the case of a G16 to U mutation, C23 will also be mutated to an A to maintain base-pairing.

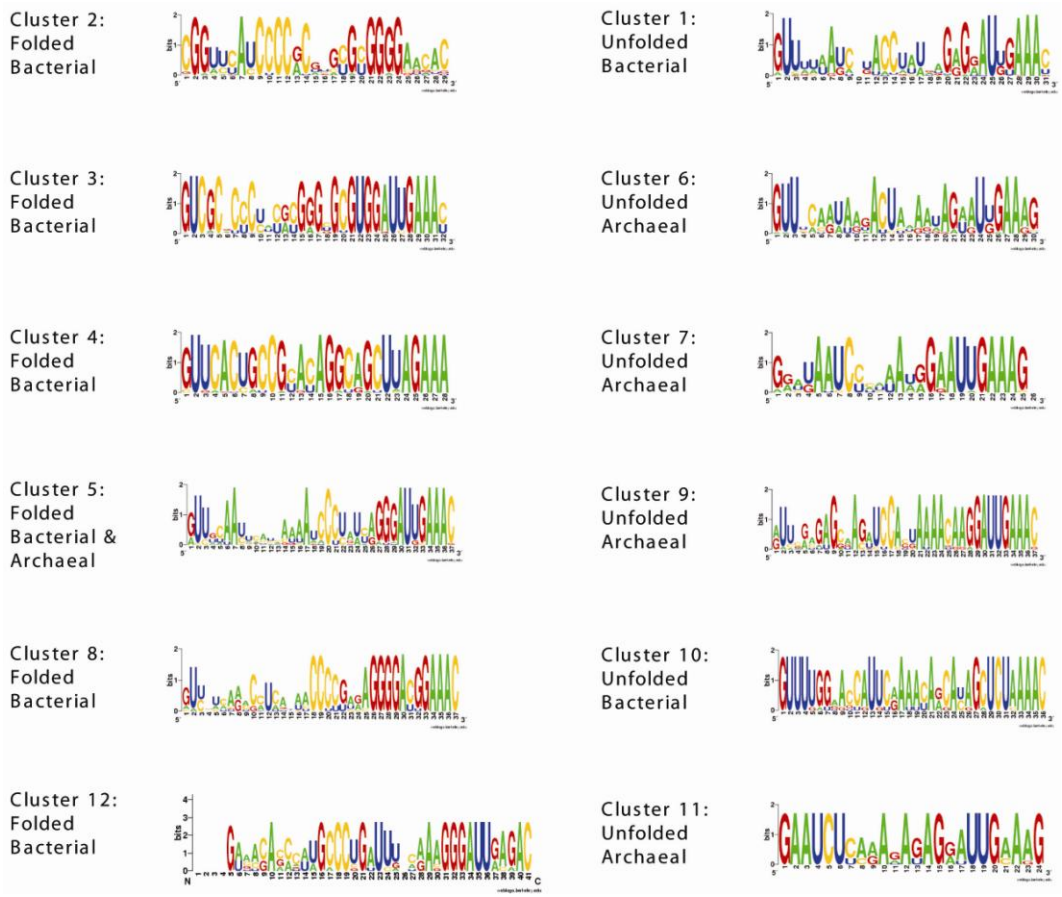

Figure 1-8. Sequence similarity of CRISPR repeats. The 12 largest CRISPR repeat clusters are shown with their sequence logos and a coarse phylogenetic composition. On the left are the folded repeats are grouped and on the right the single stranded unfolded repeats are grouped (Adapted from Kunin *et al.*, 2007).

Repeats from the same cluster, but from evolutionarily unrelated species were shown to have more sequence similarity than to repeats from more closely related species with distinct cluster types further supporting the proposed horizontal mobility of these loci (Kunin *et al.*, 2007; Sorek *et al.*, 2008). Certain clusters are found exclusively in proximity to a particular Cas subtype (Table 1-2). For example cluster 2, 3, and 4 are only found near Ecoli, Dvulg, and Ypest Cas

subtype genes respectively arguing that specific subtypes of Cas proteins are necessary for the processing of distinct repeat clusters (Table 1-2; Kunin *et al.*, 2007).

1.7 Spacer Acquisition

An interesting but poorly understood aspect of the CRISPR-Cas system is the mechanism by which spacers are selected for integration. The phage or plasmid sequence chosen to be incorporated as a new spacer is referred to as the proto-spacer which are the future targets of the crRNAs (Deveau *et al.*, 2008; Mojica *et al*., 2005). Spacers are always incorporated into the CRISPR loci nearest the leader (Mojica *et al*., 2005). A consequence of consistent spacer incorporation into the leader end of the CRISPR loci, is that a genomic history can be drawn from the order and origin of the spacers. This feature has become useful in the genomic analysis of bacterial strains such as *M. tuberculosis*, termed spoligotyping (spacer oligonucleotide typing) (Mokrousov *et al.*, 2007; Zhang *et al.*, 2010). It has been shown that CRISPR loci may have a limit to the number of spacers they can carry, and as new spacers are added to the leader end, older ones may be removed from the distal end to ensure that loci do not expand beyond their capacity (Deveau *et al.*, 2008; Deveau *et al.*, 2010; Horvath *et al.*, 2008).

Interestingly, the two to three nucleotides adjacent to the proto-spacer region in phage and plasmids appear to play a role in spacer acquisition. These

sequences, called proto-spacer adjacent motifs (PAM) are likely what targets the acquisition complex to the nucleic acid (Deveau *et al.*, 2008; Mojica *et al.*, 2009). Proto-spacer adjacent motifs are conserved between CRISPR subtypes, for example, Ecoli subtype Cas system protospacers have a [C A/T T] adjacent motif (Figure 1-9; Mojica *et al.*, 2009). The orientation of spacer integration is dependent on the polarity of the PAM. The PAM side of the spacer is incorporated nearest to the leader indicating that it is a double stranded nucleic acid (most likely DNA) that is processed for integration (Mojica *et al.*, 2009). It is unknown which proteins are responsible for recognizing this sequence and how they might direct its processing and integration. Further studies are required to explain this sequence preference.

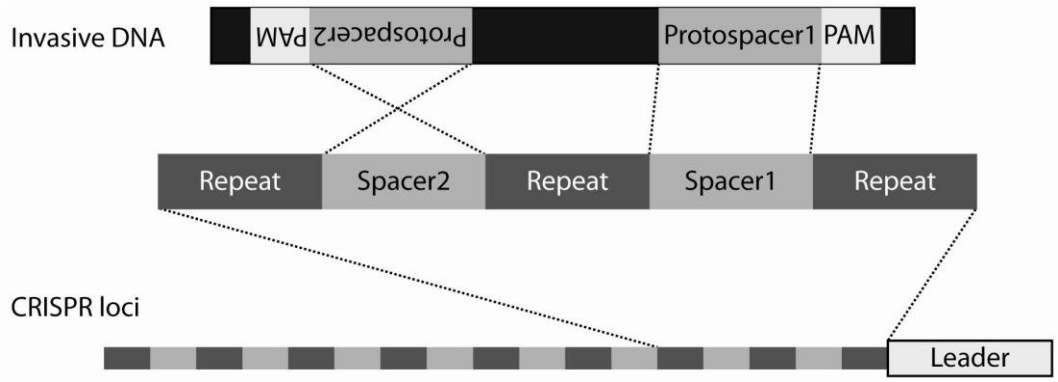

Figure 1-9. Spacer acquisition polarity. The protospacer adjacent motif (PAM) determines the orientation in the CRISPR loci. Spacers 1 and 2 are derived from opposite DNA strands with respect to each other, but the same orientation with respect to the PAM. Spacers are always incorporated in the position nearest the leader sequence.

1-8. CRISPR-Cas Targeting

The CRISPR-Cas system is analogous to the eukaryotic RNA interference (RNAi) pathway in that both systems utilize small effector RNAs as guides to target nucleic acid for degradation (Makarova *et al.*, 2006; Makarova *et al.*, 2009). Unlike eukaryotic RNAi, several lines of evidence taken together indicate that DNA is the primary target rather than messenger RNA (mRNA). No spacers have been identified to date that target RNA viruses and many spacers originate from genes expressed late in the lysogenic cycle (Deveau *et al.*, 2010; Mojica *et al.*, 2009; Semenova *et al.*, 2009; Shah *et al.*, 2009; Wiedenheft *et al.*, 2009). Spacers can originate from either coding or non-coding strands, and there are even examples of spacers originating from intergenic regions (Horvath *et al.*, 2008; Lillestol *et al.*, 2006; Semenova *et al.*, 2009; Shah *et al.*, 2009; van der Ploeg, 2009). Furthermore, in *E. coli* it has been shown *in vivo* that spacers targeting the coding or non-coding strand of lambda phage DNA are both able to confer resistance (Brouns *et al.*, 2008).

A set of telling experiments by Sontheimer's group established that in *S. epidermidis* DNA is the target of CRISPR-based interference (Marraffini & Sontheimer, 2008). The group used an isolate containing a spacer corresponding to the *nickase* (*nes*) gene found on many conjugative plasmids required for conjugation. Nickase cleaves one strand of the *oriT* locus in the donor cell before

transfer to the host cell can occur. Transcription of the *nickase* gene is not required for conjugation and therefore, nickase mRNA will only be present in the donor and not the host. Host cells harbouring the *nickase* CRISPR spacer were resistant to conjugation suggesting that the transferred plasmid DNA was targeted by the host CRISPR-Cas system (Marraffini & Sontheimer, 2008). Furthermore, interruption of the proto-spacer in the *nickase* gene with a self-splicing intron, thus changing the DNA sequence but not the mRNA sequence, abolished resistance to conjugation (Marraffini & Sontheimer, 2008). Recently, work done by Garneau and colleagues demonstrated that in *S. thermophilus* the CRISPR-Cas system rapidly cleaves plasmid and phage dsDNA site specifically within the proto-spacer region (Garneau *et al.*, 2010).

This evidence suggests that DNA is the target of the CRISPR-Cas system, however, restricted to the Cmr subtype, the CRISPR system has also been shown to target ssRNA (Hale *et al.*, 2009). A purified complex from *P. furiosus* extracts containing crRNAs cleaves target RNA, but not DNA at a fixed point from the 3' end of the spacer. The purified complex contained the complete set of Cmr proteins (cmr1-1, cmr1-2, and cmr2 - cmr6), but no core Cas proteins. *Cmr*2 encodes a protein with a polymerase domain and an HD nuclease domain, and *Cmr* 1-1, 1-2, 3, 4, and 6 all code for RAMP domain containing proteins. Reconstituted complexes comprised of recombinant proteins and synthetic crRNAs also cleaved target RNA (Hale *et al.*, 2008). Due to this conflict with

respect to nucleic acid target, CRISPR systems have been divided into sub-systems; the CRISPR-Cas system and the CRISPR-Cmr system. CRISPR-Cas targets DNA and represents the typical CRISPR system comprised of one to six core Cas proteins and some subtype-specific Cas proteins. The CRISPR-Cmr system comprised solely of Cmr subtype proteins targets RNA. Because Cmr is never the only subtype found in a bacterial genome, it may not be sufficient to confer interference alone, suggesting it may be an accessory system to the typical CRISPR-Cas system.

1-9. Targeting: Self versus Non-Self Discrimination

An obstacle faced by all cellular defence systems is the discrimination between self and non-self elements to prevent auto-immunity. Because the DNA spacer sequence is found in both target and host genomes, a mechanism must be in place to prevent self recognition. Sontheimer's group was able to demonstrate that it is the base-pairing in the sequences adjacent to the spacer that discriminates between self and non-self (Marraffini & Sontheimer, 2010). DNA sequences that do not base-pair with the crRNA beyond the spacer are targeted for degradation (Figure 1-10a). Because the effector crRNA also contains portions of the repeat on either end, the base-pairing is extended between the crRNA and the CRISPR loci protecting the DNA from targeting (Figure 1-10b). In fact, mutation of two base-pairs upstream of the protospacer can cause sensitivity to phage infection in hosts that were otherwise insensitive (Marraffini & Sontheimer, 2010).

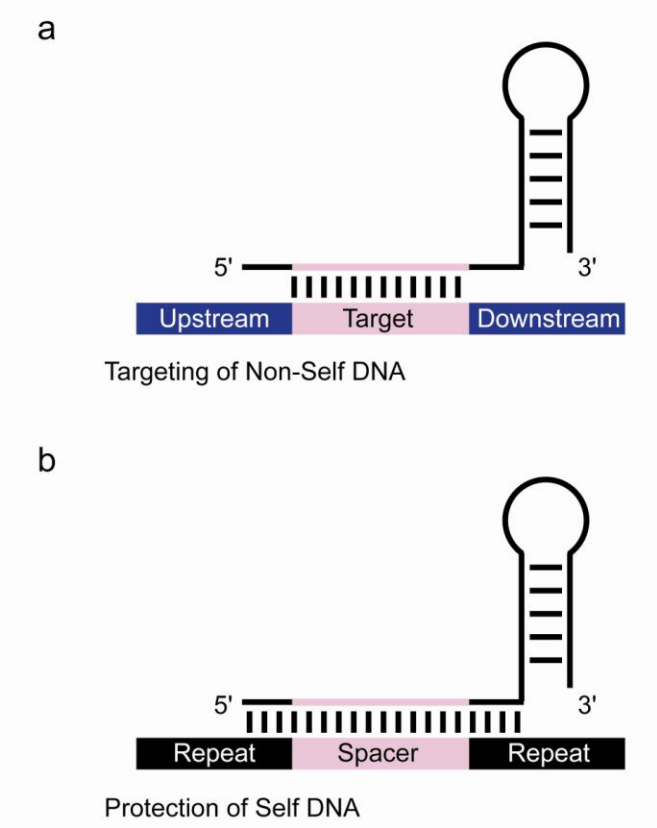

Figure 1-10. Self-versus non-self discrimination. (a) In S. epidermidis when the base-pairing between does not extended beyond the target and the spacer, CRISPR interference occurs. (b) Targets whose base-pairing does extend beyond the spacer region, for example pairing between the crRNA and the CRISPR locus, is recognized as self and is protected against degradation (Adapted from Marrafini *et al*., 2010).

This is consistent with a number of studies which have shown that phage evade CRISPR-Cas detection by a double or even a single mutation adjacent to the protospacer (Deveau *et al.*, 2008; Deveau *et al.*, 2010; Semenova *et al.*, 2009). This is yet another example of how phage and their hosts are constantly evolving responsively to promote their own survival.

1-10. Ecoli Subtype

The best studied CRISPR-Cas system is *E. coli* K12 where CRISPRs were first identified (Brouns *et al.,* 2008; Díez-Villaseñor *et al.*, 2010; Ishino *et al.*, 1987; Nakata *et al.,* 1989; Pougach *et al.*, 2010; Pul *et al.*, 2010b; Westra *et al.*, 2010). This system contains the core proteins Cas1, 2, 3, and Cas5e (herein referred to as CasD). It also contains four Cas proteins of the Ecoli subtype Cse1, Cse2, Cse3 , and Cse4, herein referred to as CasA, B, E, and C respectively (Haft *et al.,* 2005). CasA-E make up a multi-protein complex called Cascade (CRISPR-associated complex for anti-viral defence) responsible for CRISPR transcript processing (Figure 1-10; Brouns *et al.,* 2008). Recent work by the Delisa laboratory suggests that the core of the Cascade complex is comprised of a ternary complex consisting of CasC, D, and E (Perez-Rodriguez *et al.*, 2011). CasA and B are not stably associated with this ternary complex *in vitro* suggesting that they may be loosely associated accessory proteins. Analysis of both wild-type and mutant *E. coli* strains demonstrates that an initial pre-crRNA is processed to yield mature crRNA by endonucleolytic cleavage at the base of the repeat stem-loop (Brouns *et al.*, 2008). The cleavage is specifically 8 nucleotides upstream of the CRISPR repeat RNA sequence to yield mature crRNAs of ~57 nucleotides containing the spacer sequence buttressed by portions of the repeat sequence (Figure 1-11). The enzymatic activity required for this processing resides in the RAMP domain containing CasE protein (Brouns *et al.*, 2008).

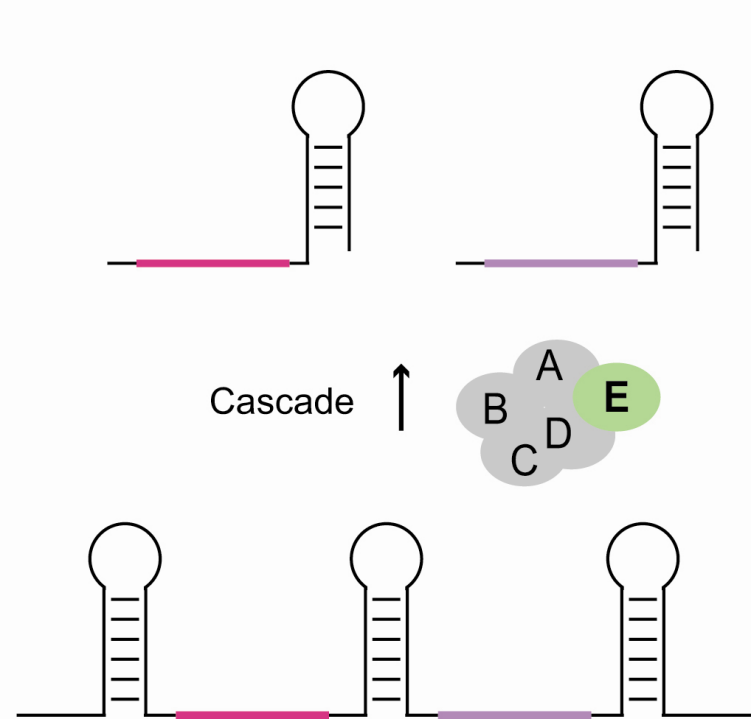

Figure 1-11. Cascade Complex. The Cascade (CRISPR-associated complex for anti-viral defence) complex has been characterized by co-immunopurification in *E. coli* K12, which consists of CasA-E. This complex cleaves pre-crRNA into mature effector crRNAs which contain the spacer sequence and a portion of the repeat on both ends. CasE is responsible for this endonucleolytic activity.

A modest amount of biochemical and structural studies have characterized the roles of some Cascade components. CasA, C, and D have no predicted domain structure or functional role. The X-ray crystal structure of the 42 kDa *T. thermophilus* CasB homologue (PDB 2ZCA) reveals a novel α-helical dimeric fold with a conserved basic patch that may be involved in nucleic acid binding (Figure 1-12a,b; Agari *et al*., 2008). Further discussion of the Cascade complex members will be presented in Appendix I.

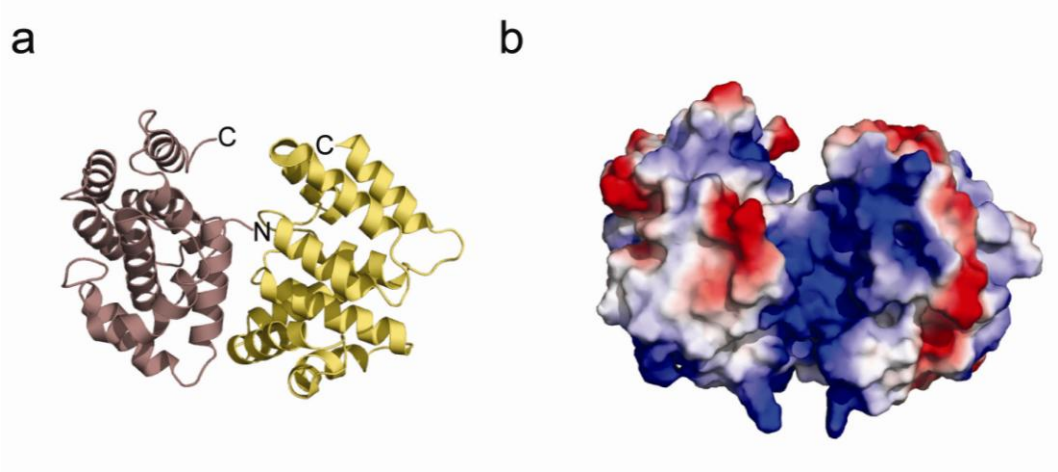

Figure 1-12. X-ray Crystal Structure of (TTHB189) CasB. (a) Ribbon diagram of *T. thermophilus* HB8 CasB homologue structure. The two monomers are shown in yellow and violet (TTHB189; PDB 2ZCA; Agari *et al*., 2008). This novel alpha-helical fold harbours a conserved basic cleft between the two monomers which may be involved in nucleic acid binding. (b) Electrostatic surface potential representation of *T. thermophilus* CasB featuring a basic cleft that may be involved in nucleic acid binding.

The *casE* gene encodes a predicted RAMP protein as it contains a conserved C-terminal glycine-rich loop and is a predicted nucleic acid binding protein (Haft *et al*., 2005). The X-ray crystal structure of the *T. thermophilus* CasE homologue features tandem ferredoxin-like domains with characteristic βαββαβ folds each consisting of a four-stranded β-sheet buttressed by two α-helices (Figure 1-13a; Ebihara *et al*, 2006). This 22 kDa protein has a PI of 9.59 and is predicted to bind nucleic acid along its basic patch (Figure 1-13b). In *E. coli*, CasE contains the enzymatic endonucleolytic activity required to process crRNAs. Interestingly, CasE has a similar tertiary structure to Cas6, which processes crRNAs in *P. furiosus*. The characterization of CasE and its comparison to Cas6 will be discussed in greater detail in Chapters 2, 3, and 4.

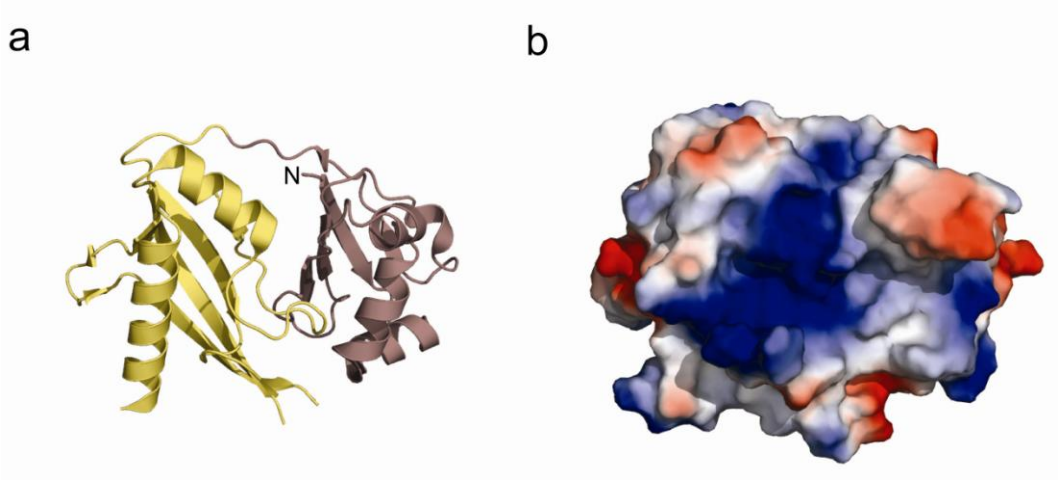

Figure 1-13. X-ray crystal structures of *T. thermophilus* CasE. (a) Ribbon diagram of TTHB192, CasE homologue from *T. thermophilus* HB8 (PDB 1WJ; Ebihara *et al.*, 2006) which contains two ferredoxin like folds featuring a basic patch on the featured surface. (b) Electrostatic surface representation of *T. thermophilus* CasE.

1-11. T. thermophilus CRISPR-Cas System

Thermus thermophilus is an extreme thermophilic bacterium with optimal growth conditions of 65-75 °C isolated from a thermal vent by Japanese investigators Oshima and Imahori (Osima & Imahori, 1974). The genome of the well characterized *T. thermophilus* HB8 strain has been completely sequenced and contains twelve CRISPR loci (Figure 1-13; Agari *et al.*, 2010; Grissa *et al.*, 2007). CRISPRs 1-10 are all found on the pTT27 mega plasmid, while 11 and 12 are on the chromosome. The pTT27 plasmid also harbours a complete set each of Ecoli, Mtube, and Cmr Cas subtypes. It carries several core *cas* genes including three copies each of Cas1 and 2; two copies of Cas3; and one copy each of Cas4 and Cas6. The Ecoli subtype genes adjacent to CRISPR 7 are ~30% identical and

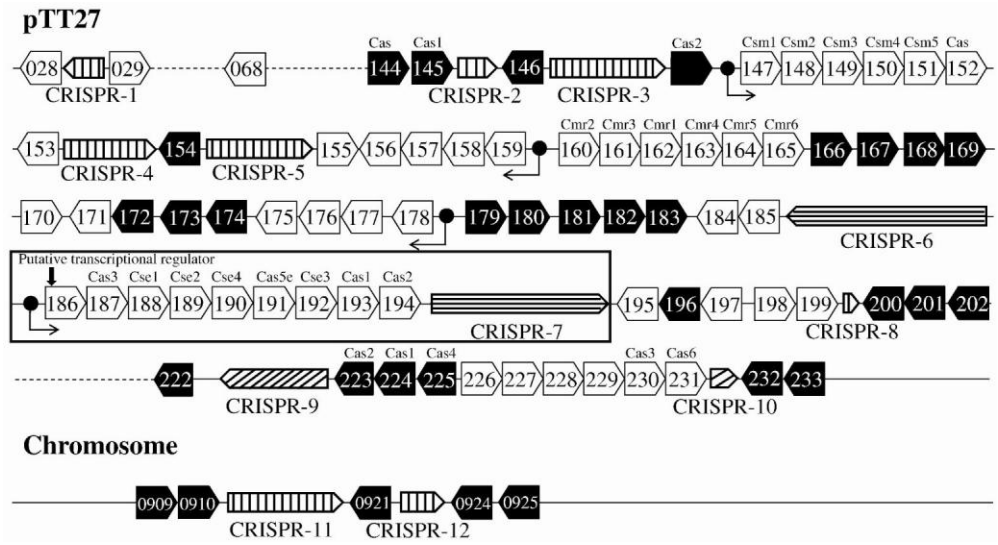

Figure 1-14. Schematic representation of CRISPRs in *T. thermophilus* HB8. The CRISPR loci and associated *cas* genes are shown from *T. thermophlius* HB8 on plasmid pTT27 and the chromosome of *T. thermophilus* HB8. *cas* gene names are shown above the arrowheads. CRISPRs are shown as patterned arrowheads, those with the same patterns having the same or similar repeat sequences. The genes encompassed by the black box are homologous to the Ecoli subtype found in *E. coli* K12. Adapted from Agari *et al*., 2010.

~60% similar to those in *E. coli* K12. The CRISPR 7 locus has a type 2 repeat sequence containing 28 nt sequences similar to those found in *E. coli* K12 which is predicted to fold into a stable RNA stem-loop (Kunin *et al*., 2007).

1-12. Discussion

The CRISPR-Cas system is a newly described genetic defence pathway in prokaryotes. This adaptable, heritable system represents a type of bacterial immune system against phage and plasmids. Integration of new spacers derived from phage or plasmid DNA into the genome protects the host from future

encounters (Marraffini & Sontheimer 2010b). The transcription and processing of small effector crRNAs (akin to siRNAs or miRNAs) is required for targeting. In *E. coli,* the essential crRNA processing step is carried out by CasE, a member of the Cascade complex (Brouns *et al*., 2008). CasE has a similar domain structure to *P. furiosus* Cas6, another identified crRNA endonuclease, making it an intriguing subject for further biochemical and structural characterization (Carte *et al*., 2008).

Due to the nature of their growth environment, thermophilic proteins are generally very stable and lend themselves well to crystallographic investigation (Jenney & Adams, 2008). We therefore endeavoured to structurally and functionally characterize Ecoli subtype Cascade members from *T. thermophilus* strain HB8 by cloning, expressing, and purifying recombinant proteins in *E. coli.* This thesis describes this work and the conclusions that can be drawn from it which have implications for all Cas subtypes.

In Chapter 2 the functional characterization of the *T. thermophilus* CasE homologue will be presented including a detailed investigation of substrate specificity and the identification of a minimal RNA substrate. In Chapter 3, the structural characterization of CasE will be described including three protein•RNA crystallographic structures, along with a comparison to other crRNA endonucleases. In Chapter 4, conclusions regarding the mechanism of the CasE endonuclease will be drawn which have implications for all crRNA endonucleases

and future directions in this work will be described. Finally, in Appendix I preliminary work to characterize the remaining *T. thermophilus* Cascade members (CasA-D) will be described.

1-13. References

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

Babu, M., Beloglazova, N., Flick, R. & other authors (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**, 484-502.

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

Beloglazova, N., Brown, G., Zimmerman, M. D. & other authors (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. & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561.

Brouns, S. J., Jore, M. M., Lundgren, M. & other authors (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**, 960-964.

Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**, 3489-3496.

Comeau, A. M. & Krisch, H. M. (2005). War is peace--dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol* **8**, 488-494.

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

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

Díez-Villaseñor, C., Almendros, C., García-Martínez, J. & Mojica, F. (2010). Diversity of CRISPR loci in Escherichia coli. *Microbiology* **156**, 1351-1361.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & 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.

Garneau, J. E., Dupuis, M. E., Villion, M. & other authors (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71.

Godde, J. S. & Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**, 718-729.

Goodier, J. L. & Kazazian, H. H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23-35.

Grissa, I., Vergnaud, G. & Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

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

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

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. & Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956.

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

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

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

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T. & Bhaya, D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **4**, e4169.

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

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

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. & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-1575.

Jenney, F. E. & Adams, M. W. (2008). The impact of extremophiles on structural genomics (and vice versa). *Extremophiles* **12**. 39-50.

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

Keeling, P. J. & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**, 605-618.

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

Kowalczykowski, S. C. (2000). Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem Sci* 25, 156-165.

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

Lillestol, R. K., Redder, P., Garrett, R. A. & Brugger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59-72.

Lillestol, R. K., Shah, S. A., Brugger, K., Redder, P., Phan, H., Christiansen, J. & Garrett, R. A. (2009). CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**, 259-272.

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

Makarova, K., Wolf, Y., van der Oost, J. & Koonin, E. (2009). Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defence against mobile genetic elements. *Biol Direct* **4**, 29.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**, 482-496.

Marraffini, L. A. & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845.

Marraffini, L. A. & Sontheimer, E. J. (2010a). CRISPR interference: RNA directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics* **11**, 181-191.

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

Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol* **17**, 85-93.

Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-182.

Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740.

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**, 760-766.

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

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901.

Oshima, T. & Imahori, K. (1974). Description of Thermus thermophilus (Yoshida and Oshima) comb. nov., a non sporulating thermophilic bacterium from a Japanese thermal spa. Int J Syst Bacteriol 24, 102-112.

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K. H., Bernardis, S., Ke, A. & Delisa, M. P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. *Mol Microbiol* **79**, 584-599.

Portillo, M. C. & Gonzalez, J. M. (2009). CRISPR elements in the Thermococcales: evidence for associated horizontal gene transfer in Pyrococcus furiosus. *J Appl Genet* **50**, 421-430.

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

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

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

Schaack, S., Gilbert, C. & Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol* **25**, 537-546.

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

Shah, S. A., Hansen, N. R. & Garrett, R. A. (2009). Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochem Soc Trans* **37**, 23-28.

Sinkuna, T., Gasjunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. Feb 22, Epub ahead of print.

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

Thomas, C. M. & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721.

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

van der Ploeg, J. R. (2009). Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. *Microbiology* **155**, 1966-1976.

Viswanathan, P., Murphy, K., Julien, B., Garza, A. G. & Kroos, L. (2007). Regulation of dev, an operon that includes genes essential for Myxococcus xanthus development and CRISPR-associated genes and repeats. *J Bacteriol* **189**, 3738-3750.

Westra, E., *et al.* (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Mol Microbiol*.

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

Wozniak, R. A. & Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol* **8**, 552-563.

Chapter 2[(1)]

Biochemical Characterization of *T. thermophilus* CasE

[1] Adapted from Gesner *et al*., (2011). *NSMB.* Accepted

2-1. Production of Effector crRNAs.

The CRISPR-Cas system has been described as a prokaryotic immune system in which bacteria and archaea can resist genetic invasion to previously encountered phage and plasmid DNA (Barrangou *et al.*, 2007; Sorek *et al.*, 2008). The adaptable, heritable component of the system stems from the incorporation of spacers derived from mobile genetic elements into the CRISPR loci of the bacterial genome which is thought to involve the core proteins Cas1 and Cas2 (Brouns *et al.*, 2008). Transcription and endonucleolytic processing of the CRISPR loci produces mature crRNAs which subsequently target invasive DNA for degradation (Brouns *et al.*, 2008; Karginov & Hannon, 2010; Marraffini & Sontheimer, 2010). The efficient and specific production of effector crRNAs by CRISPR-associated endoribonucleases is critical for the downstream targeting likely carried out by the core Cas3 protein (Brouns *et al*., 2008).

The molecular details of crRNA production are only beginning to come to light. Indeed in most subtypes, the enzymes involved in this process have not been identified. Despite their general lack of sequence similarity, the crRNA endonucleases identified to date are RAMP domain containing proteins, *E. coli* CasE, *P. furiosus* Cas6, and *P. aeruginosa* Csy4 (Brouns *et al.*, 2008; Carte *et al.*, 2008; Carte *et al.*, 2010; Haurwitz *et al.*, 2010). Most Cas systems contain at least one RAMP domain containing protein suggesting that these proteins carry out this function in other subtypes. *P. furiosus* Cas6, *P. aeruginosa* Csy4, and *E. coli*

CasE specifically recognize and cleave crRNA repeat sequences in the CRISPR transcript 8 nucleotides upstream of the next spacer sequence producing ~ 60 nt mature crRNAs (Brouns *et al.*, 2008; Carte *et al.*, 2008; Carte *et al.*, 2010; Haurwitz *et al.*, 2010). Following divalent metal independent cleavage, the effector proteins remain bound to the 5' cleavage products suggesting that they may be involved in downstream processes (Brouns *et al.*, 2008; Carte *et al.*, 2008; Carte *et al.*, 2010; Haurwitz *et al.*, 2010). For example, in *E. coli*, CasE is a member of the multi-protein Cascade complex (CRISPR-associated complex for antiviral defence) consisting of CasA-E. The roles of the CasA-D components is unknown, but it has been speculated that the complex may somehow be involved in the subsequent targeting step (Brouns *et al*., 2008).

The *P. furiosus* RAMP protein, Cas6, produces effector crRNAs by binding to the 5' end of the repeat RNA and cleaving after A22 in the 30 nucleotide repeat sequence (Figure 2-1a; Carte *et al.*, 2008; Carte *et al.*, 2010). The *P. furiosus* repeat is single-stranded (cluster 6), distinct from the *P. aeruginosa* (cluster 4) and *E. coli* (cluster 2) repeats which form stable stem-loops. The *P. aeruginosa* RAMP protein, Csy4, produces effector crRNAs by cleaving near the base of the stem-loop structure in the 28 nucleotide repeat (Figure 2-1b; Haurwitz *et al.*, 2010). *E. coli* K12's CasE, like Csy4, cleaves near the base of the stem-loop to produce effector crRNAs (Brouns *et al.*, 2008).

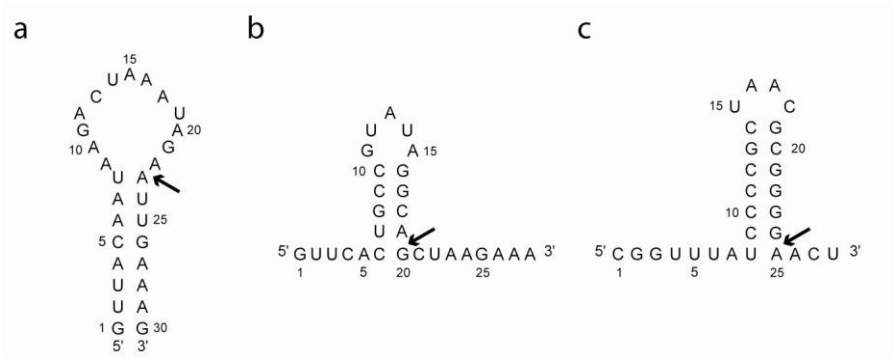

Figure 2-1. Specific endonucleolytic cleavage of CRISPR repeat RNAs. Shown are the predicted secondary structure representations of repeat RNAs, cleavage site indicated by arrow. (a) The *P. furiosus* CRISPR repeat RNA. The folding of this secondary structure is not favourable with a dG of 0.6 kcal/mol (Carte *et al.*, 2008). (b) The *P. aeruginosa* CRISPR repeat RNA. The folding of this secondary structure is favourable with a dG of -8.4 kcal/mol (Haurwitz *et al.*, 2010). (c) The *E. coli* K12 CRISPR repeat. The folding of this secondary structure is favourable with a dG of -16.5 kcal/mol and a Tm of 82.5 ºC under physiological conditions. RNA folding was computed using mfold (Brouns *et al*., 2008; Jaeger *et al*., 1989).

2-2 The Cascade complex

The *E. coli* K12 genome has two CRISPR loci and one set of Ecoli subtype *cas* genes comprised of *cas1-3* and *casA-E* (Díez-Villaseñor *et al.*, 2010; Haft *et al.*, 2005; Makarova *et al.*, 2006). Work by Brouns and colleagues identified the stable, multiprotein Cascade complex (CRISPR-associated complex for antiviral defence) comprised of CasA-E by co-immunoprecipitation (Brouns *et al.*, 2008). The long pre-crRNA transcript is processed by Cascade to yield a ~ 57 nucleotide RNA product (Figure 2-2a). This complex is specific for Ecoli subtype CRISPR transcript processing as it cannot cleave the Ypest subtype CRISPR transcript from *E. coli* UT189 (Brouns *et al.*, 2008).

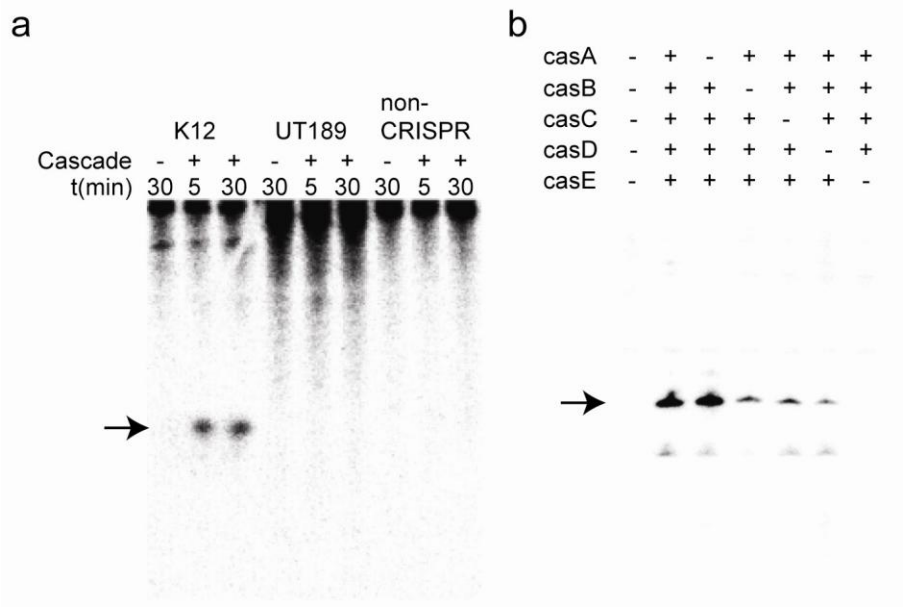

Figure 2-2. Cascade cleaves CRISPR RNA precursors into small crRNAs of ~57 nucleotides. (a) Cleavage activity assays with purified Cascade complex using *in vitro* transcribed á-^{32}P–UTP–labeled pre-crRNA from *E. coli* K12, *E. coli* UTI89, and non-crRNA as substrates. (b) Northern analysis of total RNA of BL21 (DE3) expressing the *E. coli* K12 pre-crRNA and either the complete or incomplete Cascade complex (Adapted from Brouns *et al*, 2008).

By eliminating each individual Cascade component successively, Brouns and colleagues were able to show that the enzymatic activity required for crRNA processing resides in the RAMP protein, CasE, a metal-independent riboendonuclease that cleaves at the base of the stem-loop (Fig. 2-2b; Brouns *et al.*, 2008; Pougach *et al.*, 2010). Cloning and sequencing of mature crRNAs revealed crRNA products containing a complete spacer sequence buttressed by portions of the repeat RNA. Although the 5' ends of these crRNAs consistently contained eight nucleotides of repeat sequence, the 3' termini were heterogeneous ranging from 9-19 nucleotides (Brouns *et al.*, 2008). This is difficult to rationalize as the Northern blot analysis of these RNAs presents a clean single

band. Many of the 3' ends of these cloned RNAs terminate at the top of the loop which is likely a result of non-specific RNase activity during the cloning process (Brouns *et al*., 2008). This suggests that *E. coli* CasE likely cleaves at a single position at the base of the predicted stem-loop structure similarly to Csy4.

To understand the molecular details of crRNA recognition and cleavage in the Ecoli subtype, the key component of the Cascade complex, CasE must be examined. The mode of repeat recognition in the crRNA transcript by CasE may be based on sequence, secondary structure, or a combination of the two. In this chapter, the biochemical underpinnings of *T. thermophilus* CasE endonucleolytic cleavage essential to CRISPR-based immunity will be examined.

2-3. Characterization of *T. thermophilus* CasE.

To describe the basis of pre-crRNA recognition and processing the *T. thermophilus* CasE homologue (TTHB192) has been functionally characterized. This gene is present on the pTT27 mega plasmid in an operon containing CasA-E and Cas1-3 in the identical gene order as found in *E. coli* (Cas3, CasA-E, Cas1, Cas2) (Haft *et al*., 2005). The *T. thermophilus casE* gene encodes a predicted 22 kDa protein with a theoretical PI of 9.59 which is conducive to nucleic acid binding. Due to its homology to *E. coli* CasE, it is predicted to similarly bind and process crRNA. TTHB192 was cloned into a

vector containing an amino-terminal histidine affinity tag and expressed in *E. coli* Rosetta cells. The recombinant monomeric protein was purified by successive steps of nickel bead, size exclusion and cation exchange chromatography.

2-3.1. Characterization of Cleavage Products

The purified CasE protein, consistent with the work by Brouns and colleagues, specifically binds and cleaves a 28 nucleotide RNA modeling the *T. thermophilus* repeat (Brouns *et al.*, 2008). This cleavage yielded two RNA products consistent with cleavage after G21 based on RNase T1 digestion and base hydrolysis mapping (Figure 2-3a,b). The presence of free 5' hydroxyl termini on the shorter product was confirmed by T4 kinase phosphorylation in the presence of P^{32}-ATP (Figure 2-3b). The generation of modified 2' or 3' termini on the longer product was confirmed by resistance to sodium periodate/base elimination treatment (Figure 2-3c). This evidence implies a cleavage mechanism involving in-line attack of the 2' hydroxyl of G21 on the scissile phosphate as observed in both protein and RNA catalyzed RNA cleavage (Calvin *et al.*, 2008; Xue *et al.*, 2006). Also consistent with this mechanism, substitution of a 2'-deoxy residue at the G21 position abolished cleavage in the presence of the enzyme (Figure 2-3d,e).

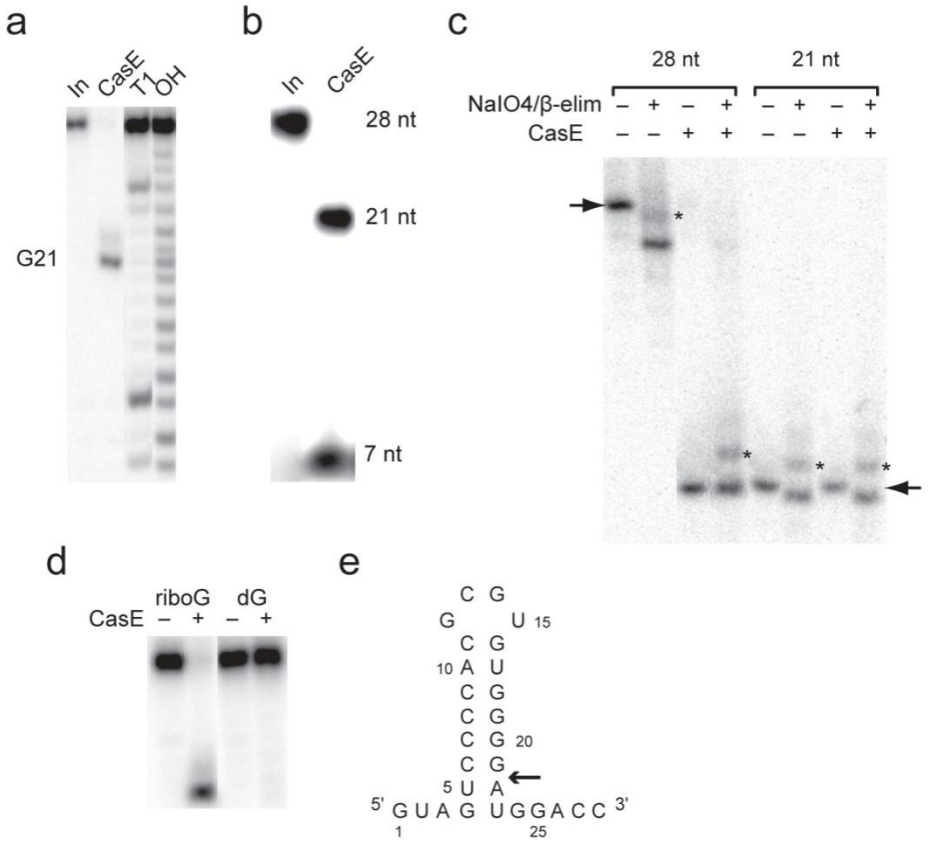

Figure 2-3. Processing of pre-crRNA by CasE. The products resulting from CasE RNA cleavage are consistent with an in-line displacement mechanism yielding a 5' product with cyclic 2',3' phosphodiester and 3' product with free 5' hydroxyl. (a) Site of cleavage by CasE was mapped to G21 by performing RNase T1 digestion and production of base hydrolysis ladder. (b) Characterization of 3' cleavage product. Unlabeled 28 nucleotide *T. thermophilus* CRISPR repeat RNA was treated with CasE , ^{32}P-5'-end-labeled with T4 polynucleotide kinase, and analyzed by 20% denaturing PAGE demonstrating that the seven nucleotide 3' product of cleavage contains a free 5' hydroxyl. (c) Characterization of 5' cleavage product. ^{32}P-5'-end-labeled RNAs were treated with periodate, followed by base elimination either in the absence of or following incubation with CasE. RNA oligonucleotides were treated as shown and resolved by 20% denaturing PAGE. Shown are comparisons between the 28 nucleotide model of the *T. thermophilus* CRISPR repeat RNA (lanes 1,2), the 5' CasE cleavage product (21 nucleotides, lanes 3,4), and the 21 nucleotide RNA corresponding to the 5' cleavage product but with free 2',3' hydroxyl terminus (lanes 5, 6). The cleavage product does not change mobility on periodate/base-elimination (arrow) showing that the 2',3' position(s) are modified. * denotes uncharacterized product of oxidation/elimination. Shown are comparisons between the 28 nucleotide and the 21 nucleotide models of the *T. thermophilus* CRISPR repeat RNA oligonucleotides. (d) Cleavage by CasE was abolished when a deoxyG residue was substituted at position 21 of the crRNA repeat. (e) Secondary structure representation of the *T. thermophilus* CRISPR repeat.

2-3.2. Optimization of Reaction Conditions.

To further characterize TTHB192 we sought to optimize the reaction conditions of CasE RNA cleavage. Rates were calculated for reactions incubated at temperatures ranging from 0-80 °C. Activity increased with increasing temperature dropping off sharply between 70 °C and 80 °C (Figure 2-4a). This is consistent with the optimal growth conditions of *T. thermophilus* which ranges from 65-75 °C. The presence of divalent metal cations was not required and reaction rates were not affected by the presence of 1 mM EDTA, a divalent metal ion chelator consistent with the work by Brouns and colleagues (Figure 2-4b; Brouns *et al.*, 2008). Cleavage occurs maximally at a salt concentration of 100 mM NaCl and altering the salt from NaCl to KCl had no effect on cleavage (Figure 2-4c). The optimal pH range of CasE is between pH 7.5-8. The pH profile displayed a broad bell shaped curve indicating the possible presence of two ionizable groups consistent with an acid-base catalysis mechanism similar to RNaseA or the tRNA splicing endonuclease (Figure 2-4d). Indeed, a universally conserved histidine mutated to alanine in the *E. coli* CasE homologue was inactive despite being incorporated into the Cascade complex (Brouns *et al.*, 2008). The imidazole sidechain of histidine has a pKa of ~6.0 which may represent the first titratable group in the bell curve. In the *T. thermophilus* CasE structure a conserved tyrosine (pKa of ~10.0) is in close proximity to the conserved histidine and may represent the second titratable group (Ebihara *et al.*, 2008).

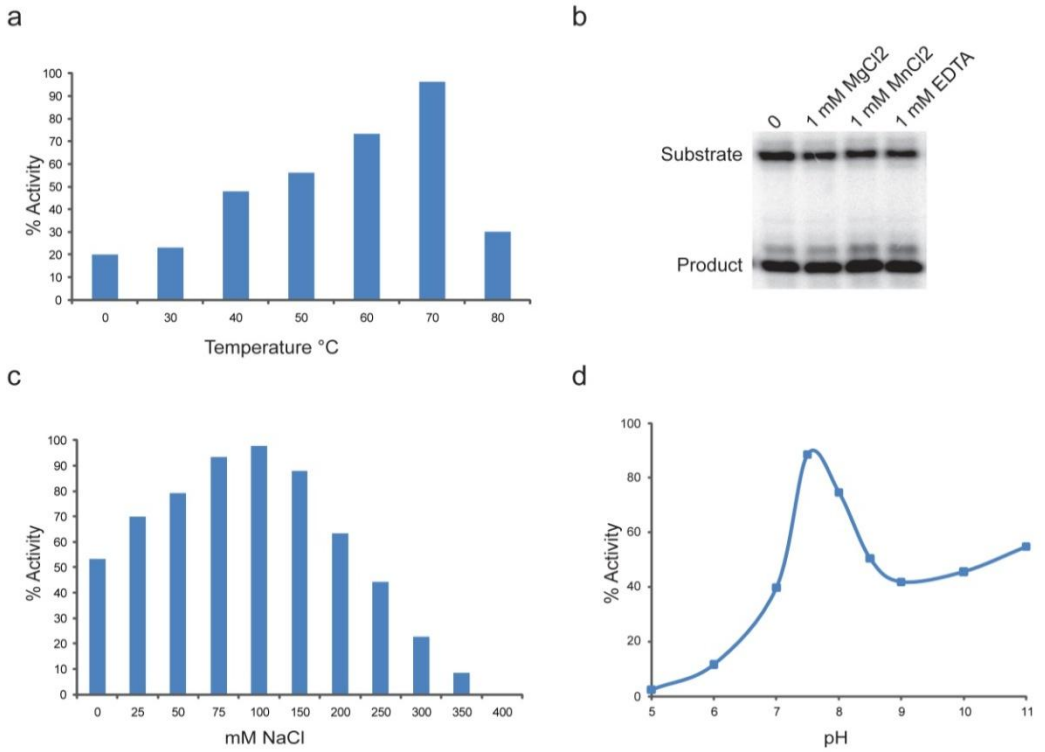

Figure 2-4. Optimization of Reaction Conditions. (a) Graphic representation of activity assay. CasE cleavage rates calculated for reactions incubated at temperatures ranging from 0-80 °C at pH 8.0. (b) Effect of cleavage activity in the presence of divalent metal cations, or a divalent metal cations chelator (EDTA), shown here is a denaturing gel of ^{32}P-5'-end-labeled 28 nt RNA substrate incubated with 10 nM wild-type CasE for 30 seconds. (c) Graphic representation of activity assay. CasE cleavage rates calculated for reactions incubated at salt concentrations ranging from 0-400 mM NaCl. (d) Graphic representation of activity assay. CasE cleavage rates calculated for reactions incubated at pH ranging from 5-11. The above activity assays were carried out at 22 °C to ensure accurate rate measurement.

2-3.3. Characterization of Substrate Specificity.

With *T. thermophilus* CasE repeat processing activity established, the mode of RNA recognition must be addressed by examining substrate preference. CasE binds (Kd ~1-5 nM) and cleaves the *T. thermophilus* repeat crRNA sequence based on denaturing PAGE and gel electrophoretic mobility shift assay.

Several species are apparent in the native gel including a crRNA substrate dimer (that is not folded into a hairpin, rather a dimer is formed), the CasE•Substrate complex, the CasE•Product complex and a CasE•Product dimer which is a super-shifted band. The repeat RNA hairpin has a salt adjusted melting temperature of 82.5 ºC, which is well above the growth environment of *T. thermophilus*, whereas the melting temperature of the RNA•RNA dimer is significantly lower at ~65 ºC most likely due to the presence of two G:U mismatches in the imperfect duplex. Although the dimer complex consisting of two CasE molecules bound to an RNA•RNA duplex is present in most of the gel shifts presented in this work, it is a minor species *in vitro*, based on EMSA and size exclusion chromatography. The major species consists of CasE bound to a single RNA hairpin. CasE cannot bind or cleave a DNA oligonucleotide homologous to the repeat RNA, an unrelated RNA hairpin (an 80 nucleotide imperfect hairpin which is a substrate for RNA editing), or a 21 nucleotide single-stranded RNA (based on human miRNA-16) (Figure 2-5a,b). Therefore, like the *E. coli* homologue, *T. thermophilus* CasE's activity is specific for its cognate repeat RNA (Brouns *et al*., 2008).

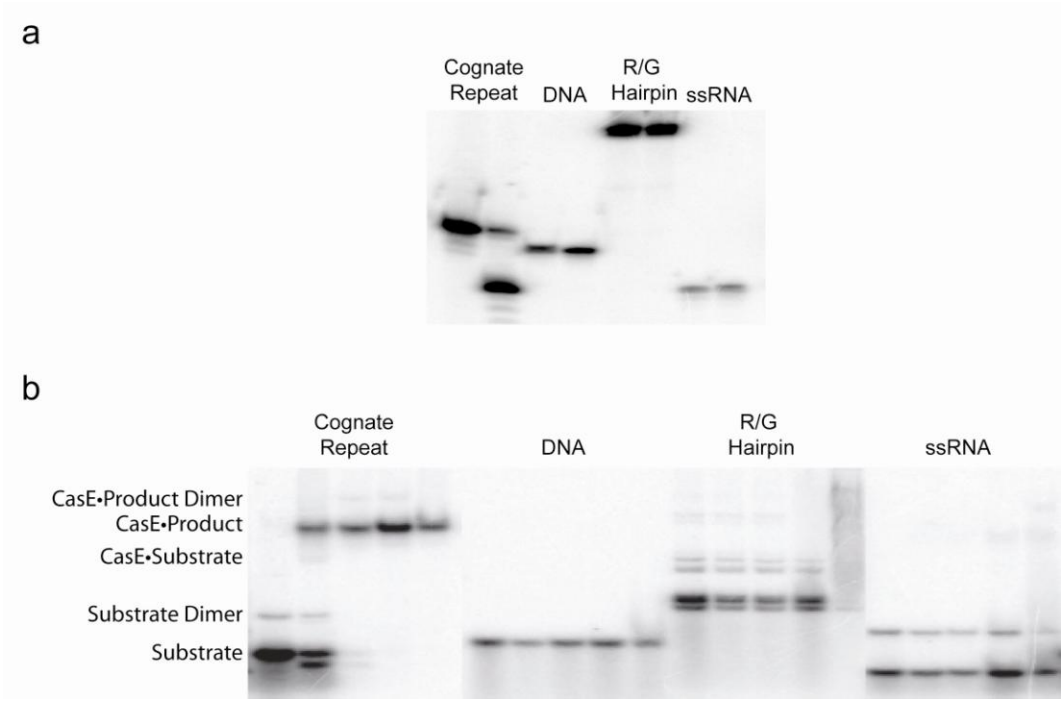

Figure 2-5. CasE specifically cleaves crRNA. (a) Denaturing PAGE analysis of ^{32}P-5'-end-labeled oligonucleotides in the absence and presence of 1 ìM CasE. CasE can only cleave the cognate repeat RNA and cannot cleave a DNA with the same sequence as the crRNA, a unrelated RNA hairpin based on the R/G substrate, or an unrelated single stranded 21 nt RNA based on human miR16. (b) Gel electrophoretic mobility shift assay of the same oligonucleotide substrates as above with 0, 1, 10, 100, 1000 nM of CasE. CasE can only shift the cognante CRISPR RNA.

To further elucidate substrate recognition, we incubated the enzyme with RNA oligonucleotides that were similar to the cognate repeat sequence with either of the two base-pairs at the base of the stem inverted. In the first mutant RNA, A21 and U22 were changed to U and G respectively, with mutations made to G4 and U5 to maintain base-pairing (Figure 2-6a). In the second mutant RNA, G20 and A21 were changed to C and U respectively, with mutations made to U5 and

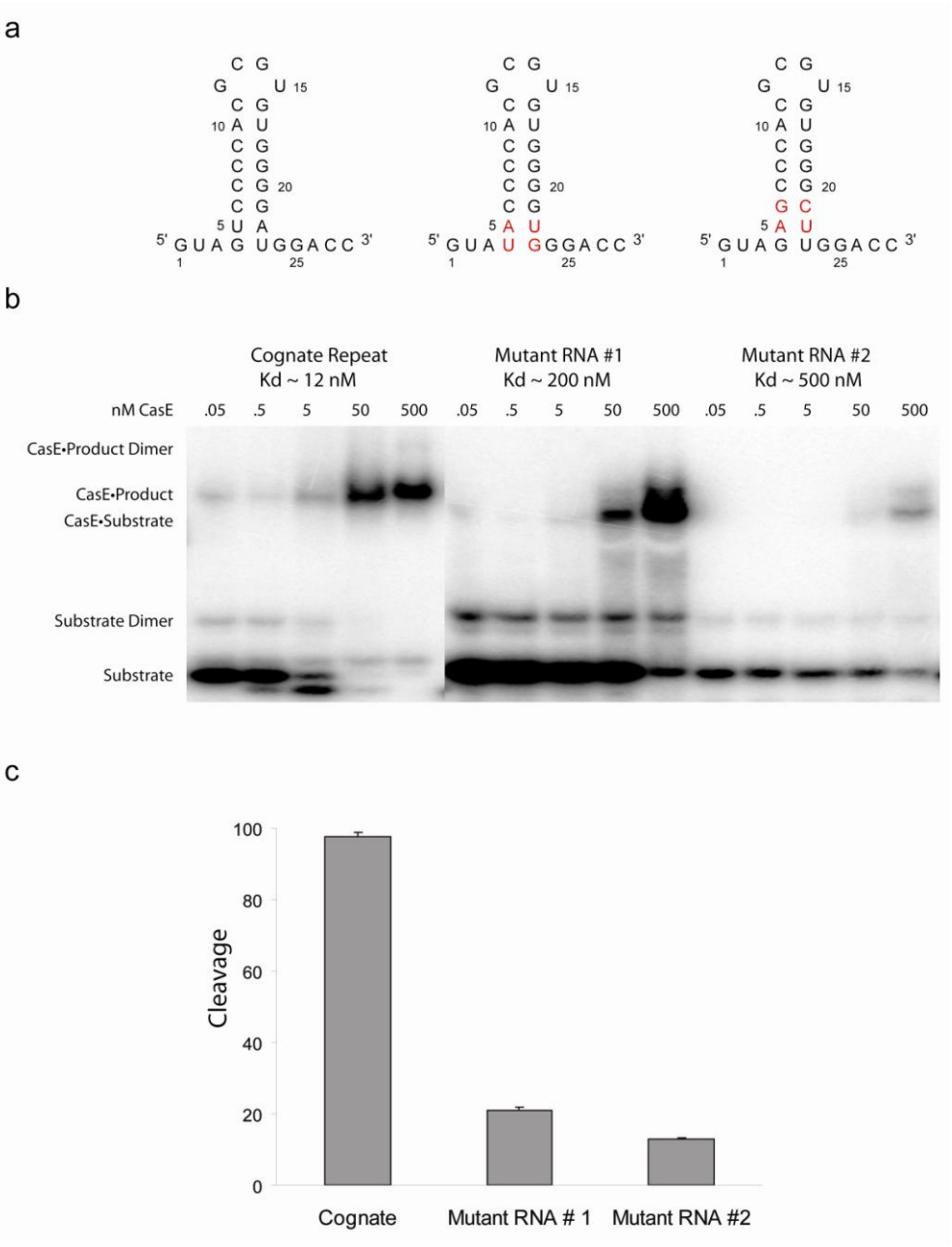

Figure 2-6. Sequence specific recognition of CRISPR RNA. (a) Gel electrophoretic mobility shift assay using ^{32}P-5'-end-labeled RNA and increasing amounts of CasE protein: .05, .5, 5, 50, 500 nM. (b) Relative cleavage activity of CasE with the cognate RNA compared to the two mutant RNAs. Cleavage rates were determined by calculating the relative % of product to substrate corrected for background of reactions comprised of CasE with the respective RNA substrates from 0 s to 5 min using Imagequant software. (c) Shown is the secondary structure representation of the cognate, mutant #1, and mutant #2 crRNAs. Mutated RNAs are shown in red.

C6 to maintain base-pairing (Figure 2-6a). These RNAs would have the same secondary structure as the cognate RNA, however their sequences near the scissile phosphate are changed (scissile phosphate is between G21 and A22). Both of these changes in sequence, but not in secondary structure caused a decrease in affinity and activity (affinity by ~20 and 40 fold, activity by ~5 and 7 fold respectively) (Figure 2-6b). This indicates that this region of the cognate repeat sequence is specifically recognized by CasE. Perturbations in this sequence result in weaker affinity and activity suggesting that secondary structure alone is not sufficient for substrate recognition.

2-3.4. Defining a Minimal Substrate.

CasE can bind an RNA mimicking the 5' cleavage product (that is an RNA that ends at G21), and an RNA that contains deoxyG at G21 with the same affinity as the cleavable substrate (~1-10 nM) (Figure 2-7). The protein has no affinity for the 3' cleavage product (Figure 2-7). Consistent with the work by Brouns and colleagues, CasE remains bound to the 5' portion of its substrate following cleavage based on gel electrophoretic mobility shift assays and gel filtration (Brouns *et al.,* 2008).

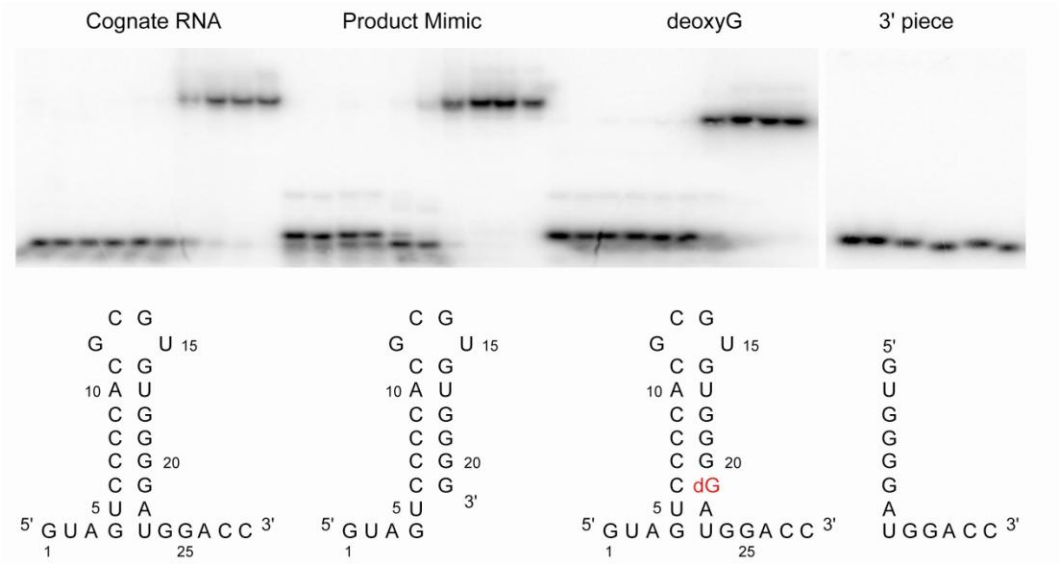

Figure 2-7. Affinity of CasE for RNAs. CasE binds to the 5' cleavage product mimic and a deoxy G21 containing CRISPR repeat RNA with the same affinity as the cognate full length RNA, but does not bind the 3' piece. Shown above is a gel electrophoretic mobility shift assay using ^{32}P-5'-end-labeled RNA and increasing amounts of protein: 0, 10 *f*M, 100 *f*M, 1 pM, 10 pM, 100 pM, 1 nM, 10 nM, 100 nM, and 1 ìM CasE protein. Secondary structure representation of oligonucleotides used are shown below.

To characterize the RNA bases necessary for substrate recognition a series of oligonucleotides were designed containing truncations from the 5' end of the repeat RNA. Removal of one base has essentially no effect on affinity; removal of two causes a ten-fold decrease in affinity; removal of three causes a twenty-fold decrease in affinity (Figure 2-8a,b). Removal of four bases from the 5' end disrupts the first base-pair at the base of the stem and causes a ten thousand fold decrease in affinity (Figure 2-8a,b). We can infer from this that the bases at the 5'

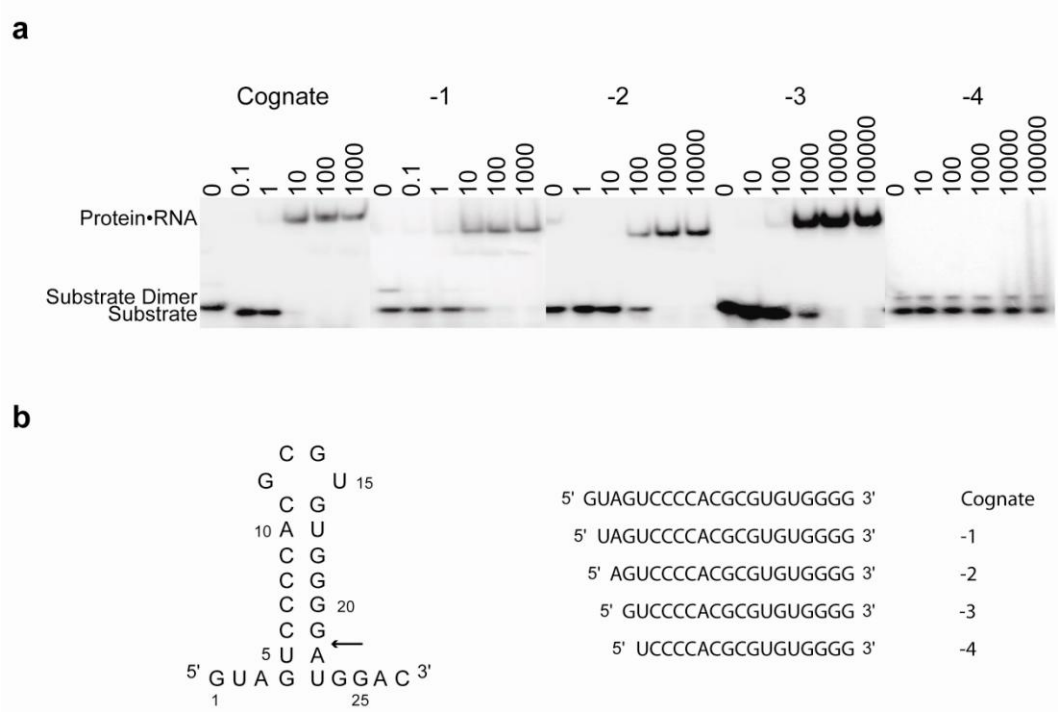

Figure 2-8. 5' end Requirements for Substrate Recognition. (a) Gel electrophoretic mobility shift assays of ^{32}P-5'-end-labeled RNA substrates containing truncations from the 5' end. The concentration of CasE in nM is shown above the lanes. (b) Shown is the predicted secondary structure of *T. thermophilus* CRISPR repeat on the left with the sequences used in the 5' truncations on the right.

end of the repeat are involved in substrate recognition and that removal of the first base-pair of the stem is detrimental to recognition.

Oligonucleotides containing 3' truncations of the CRISPR repeat were designed and all bound with the same affinity as the cognate RNA (~2 nM). However a change in reaction rate was observed when the third nucleotide after the cleavage site (G24) was removed. This had a ~ five-fold effect on cleavage

rate. Removal of U23 was inhibited by ~ ten-fold (Figure 2-9a,b). It is possible that these nucleotides are somehow involved in RNA recognition. Taken together, the CasE minimal substrate for binding and cleavage is a 21 nt RNA comprised of the cognate CRISPR repeat sequence with 3 nt removed from the 5' end and 4 nt from the 3' end.

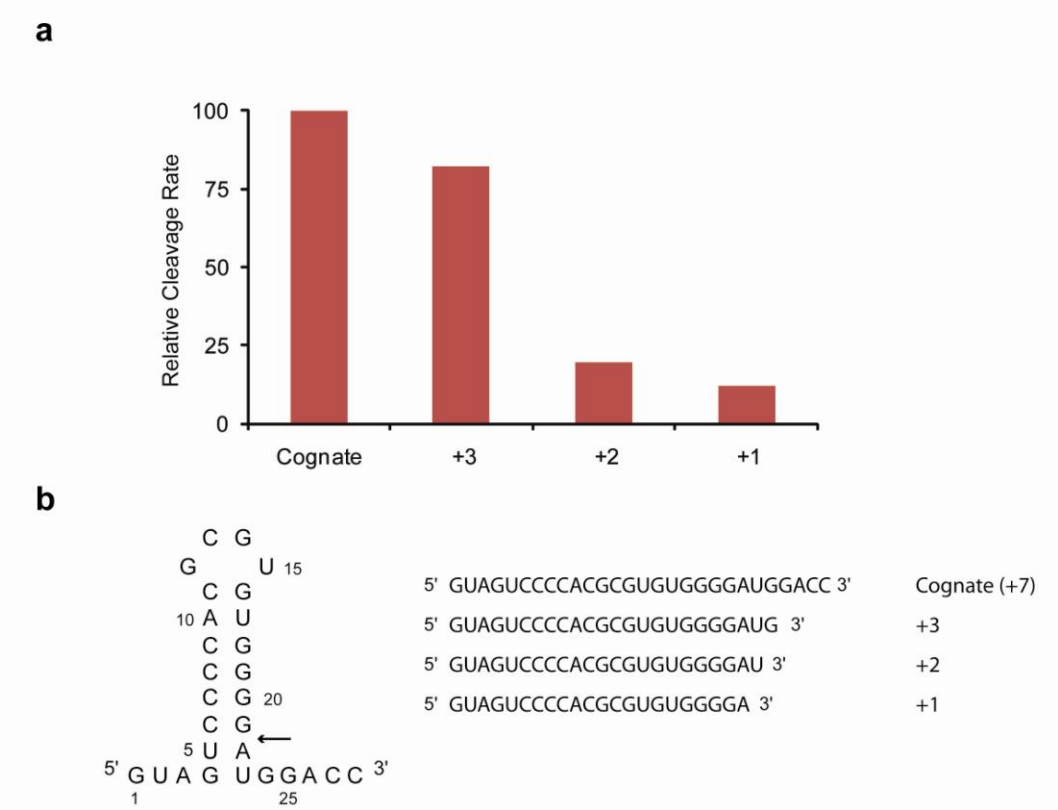

Figure 2-9. 3' end Requirements for Substrate Recognition. (a) Graphic representation of relative cleavage rates of ^{32}P-5'-end-labeled RNA substrates containing truncations from the 3' end. (b) Shown is the predicted secondary structure of *T. thermophilus* CRISPR repeat on the left with the sequences used in the 3' truncations on the right.

2-3.5. Strand Separation is Necessary for Activity

To determine whether the strength of base-pairing affected cleavage activity an oligonucleotide was designed containing the CRISPR repeat sequence, but with an extra two base-pairs at the base of the stem (Figure 2-10d). This strengthened base-pairing resulted in significantly decreased CasE cleavage activity at 65 °C and no activity at 22 °C (Figure 2-10a,b). CasE cleaved a control RNA designed with the same perturbations in the 5' sequence without the extra two base-pairs with equivalent activity to the cognate repeat. CasE bound both of these RNAs with essentially equivalent affinities to the cognate RNA based on gel electrophoretic mobility shift assay (Figure 2-10c). This indicates that the base of the stem must somehow be melted for cleavage to occur and that this unwinding step is separate from substrate binding.

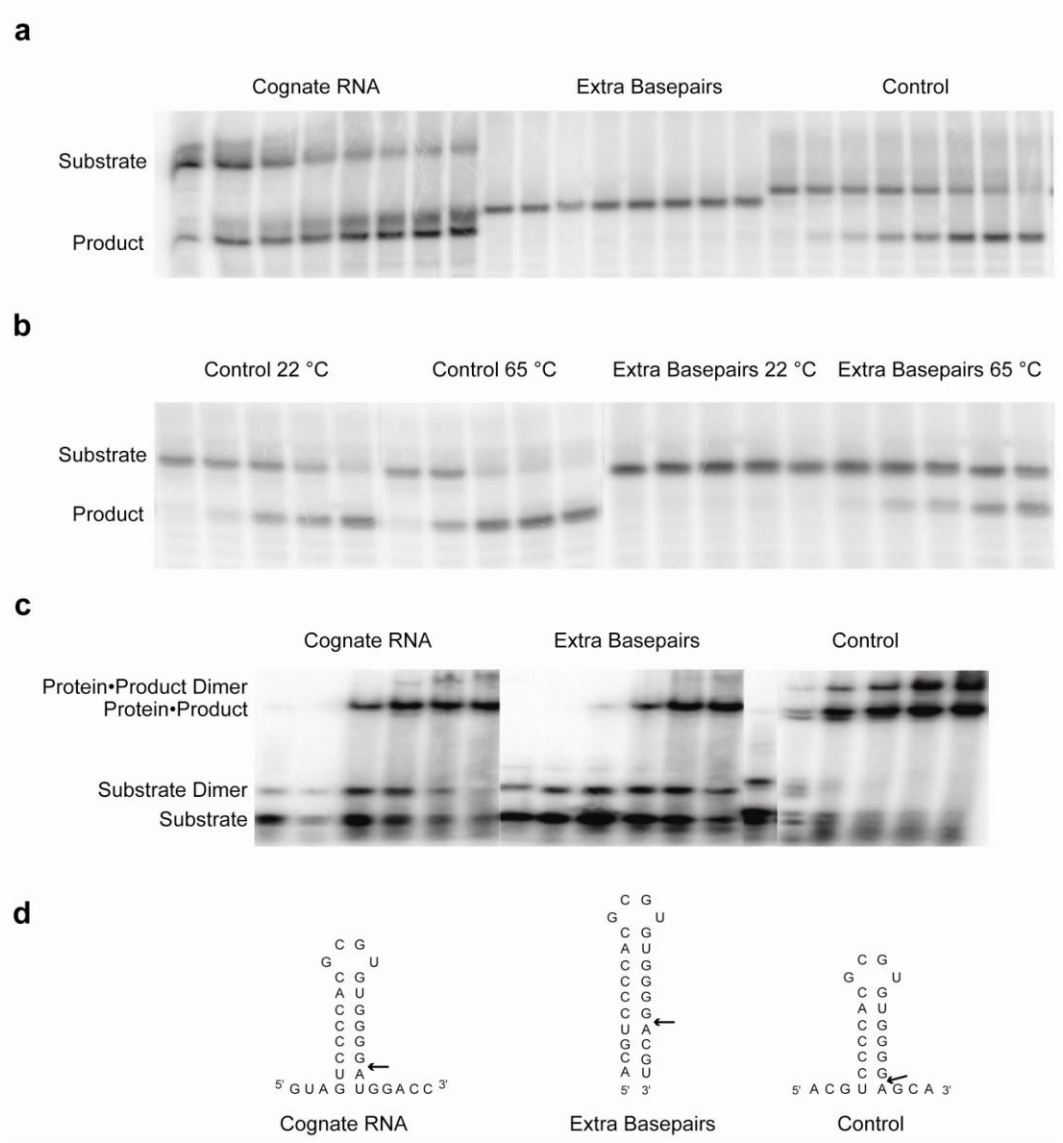

Figure 2-10. Separation of base-pairs at the base of the stem is critical for activity. (a) Time course of CasE cleavage of ^{32}P-5'-end-labeled RNA substrates at 22 °C (10 s, 30 s, 60 s, 2 min, 5 min, 15 min, 30 min, 60 min). (b) Cleavage activity assay of ^{32}P-5'-end-labeled RNA substrates over time at both 65 °C and 22 °C at pH 8.0 (10 s, 60 s, 5 min, 10 min, 20 min, 60 min). (c) Gel electrophoretic mobility shift assay of ^{32}P-5'-end-labeled RNA substrates with an increasing amount of CasE protein (3.1 nM, 6.3 nM, 12.5 nM, 25 nM, 50 nM, 100 nM). Four species are present on this gel the RNA substrate, the RNA substrate dimer, the Protein•Product complex and the protein•product dimer. (d) Secondary structure representation of the Cognate RNA, the RNA with extra base-pairs, and the control RNA.

2-3.6. Cleavage Rate Analysis

Characterization of the kinetics of CasE cleavage was carried out by performing essentially a Michaelis-Menten analysis. By keeping the concentration of enzyme constant we determined cleavage rates at ten different substrate concentrations ranging from 3X-15X the concentration of enzyme (100 nM). The V_{max} was estimated to be 47.47 *f*mol s^{-1} and the K_M to be 390 nM. The K_M is reasonably consistent with the K_d range of ~1-10 nM. This kinetic analysis was performed at 22 °C, much lower than the 65-75 °C growth environment of *T. thermophilus*. This experiment was performed in this manner to ensure accurate calculation of rates. At 65 °C, product is formed at too fast of a rate to get accurate measurements (< 15 seconds). The V_{max} and K_M would be expected to be higher and lower respectively at 65 °C.

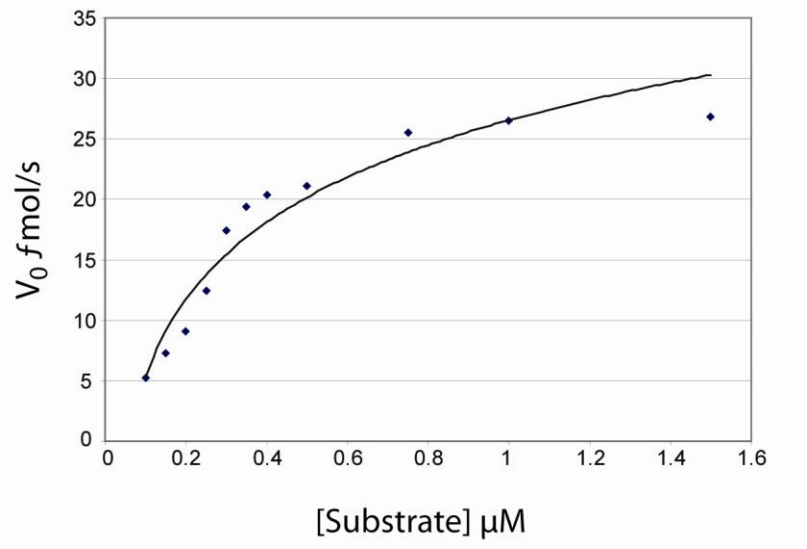

Figure 2-11. Michaelis-Menten Kinetics of *T. thermophilus* CasE cleavage. Shown here is the graphic representation of V_0 at various substrate concentrations at 22 °C, pH 8.0. Extrapolation of this graph using Graphpad Prism software, yields the V_{max} which is estimated at 47.46 *f*mol s^{-1}. The K_M is the substrate concentration at half of V_{max} which is 390 nM.

2-3.7. Analysis of *E. coli* CasE Activity

For comparison, the *E. coli* K12 CasE homologue was cloned expressed and purified from genomic DNA. Remarkably, the *E. coli* homologue has a similar affinity and endonucleolytic activity for the *T. thermophilus* repeat RNA as the *T. thermophilus* CasE homologue (2-12a,b). The *E. coli* homologue recognizes the *T. thermophilus* CRISPR repeat RNA because sequence and secondary structure between these two repeats is very conserved (2-12c; Kunin *et al*., 2007). It would therefore be expected that the *T. thermophilus* CasE would be able to bind and cleave the *E. coli* repeat sequence with a similar activity. The secondary structure of Ecoli subtype CRISPR repeat RNAs is conserved and the consensus sequence consists of a seven nucleotide stem consisting mostly of G:C base-pairs and a four nucleotide loop sequence (2-12c,d; Kunin *et al.*, 2007). The repeat RNA for *T. thermophilus* contains an extra G:U wobble pair at the base of the stem. This extension of base-pairing may be necessary in this thermophilic species due to its high temperature growth environment.

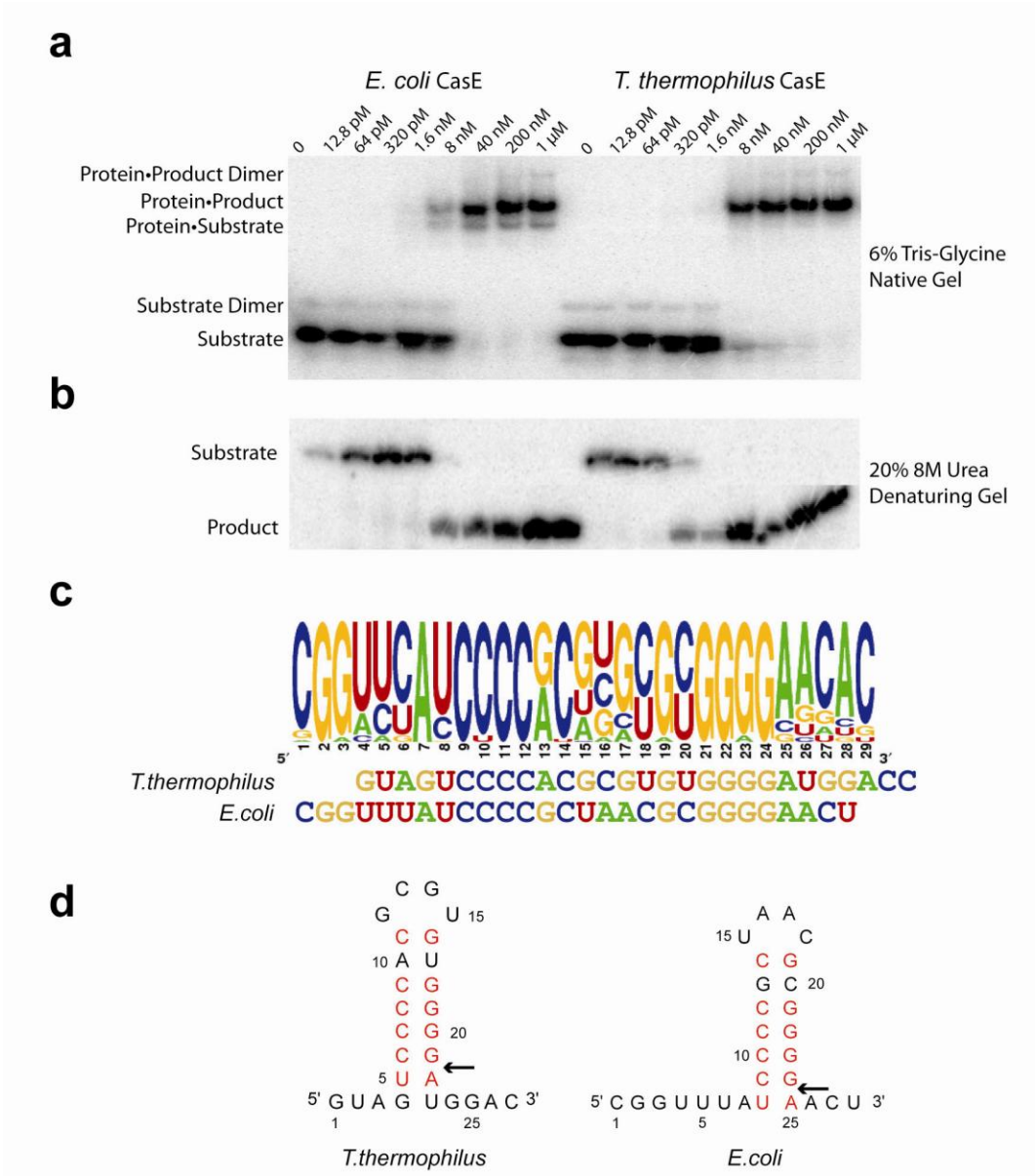

Figure 2-12. ***E. coli*** **CasE binds and cleaves** ***T. thermophilus*** **repeat RNA**. (a) Gel electrophoretic mobility shift assay of a 28 nt ^{32}P-5'-end-labeled RNA repeat and an increasing amount of protein, either *E. coli* or *T. thermophilus* CasE. Substrate, substrate dimer, CasE•Substrate, CasE•Product, and CasE•Product dimer are indicated. (b) Cleavage asssay with a 28 nt ^{32}P-5'-end-labeled RNA repeat and an increasing amount of protein, either *E. coli* or *T. thermophilus* CasE. Substrate and product are indicated. (c) Sequence conservation of Ecoli subtype CRISPR repeat RNA. Below the *T. thermophilus* and *E. coli* K12 CRISPR repeat sequences are shown (Kunin *et al.*, 2007). (d) Secondary structure representation of the *T. thermophilus* and *E. coli* CRISPR repeat RNA sequences. Cleavage site is indicated by the arrow.

2-4. Discussion

The CRISPR-Cas system is a recently discovered prokaryotic genetic defence system which allows bacteria and archaea to interfere with the expression of invasive elements. A key step in this pathway is the processing of the long precursor CRISPR RNA into short effector crRNAs. In *E. coli* this processing step is carried out by the Cascade complex consisting of CasA-E (Brouns *et al*., 2008). More recently a core ternary complex consisting of CasC,D, and E has been identified which also can process pre-crRNAs (Perez-Rodriguez *et al*., 2011). However it is known that CasE, a RAMP domain containing protein, is required for efficient and specific production of effector crRNAs. Two other RAMP domain containing proteins have also been identified as crRNA endonucleases from different subtypes. *P. furiosus* Cas6 and *P. aeruginosa* Csy4 cleave their CRISPR repeat RNA specifically in a metal-independent manner and remain bound to the 5' cleavage product suggesting that despite their dissimilarities, they may share a common mechanism (Carte *et al*., 2008; Haurwitz *et al*., 2010).

In this chapter the characterization of the *T. thermophilus* CasE homologue's specific crRNA processing activity has been described. This enzyme specifically binds to the *T. thermophilus* crRNA repeat sequence, cleaves after G21 through inline attack by the 2'OH and remains bound to the 5' cleavage product containing a cyclic 2',3' phosphate. The 3' cleavage product containing a

free hydroxyl dissociates away from the CasE·Product complex. This cleavage activity occurs optimally at 70 ºC in a buffer containing 100 mM NaCl/KCl at pH 7.5. The pH profile of activity which displays a bell shaped curve is consistent with the presence of two ionizable groups in a classic acid-base catalysis mechanism (Calvin *et al.*, 2008; Xue *et al.*, 2006). The presence of a conserved histidine and tyrosine near the basic cleft of the protein suggests that these two residues may be involved in catalysis. However simple pH profile analysis is not sufficient to identify catalytic residues and thus further mutational and structural studies are required to determine the mechanism of catalysis which will be presented in Chapter 3.

T. thermophilus CasE is specific for its crRNA substrate and does not bind or cleave unrelated oligonucleotides such as ssDNA, dsRNA or ssRNA. Perturbations of the sequence of the RNA hairpin, but not the secondary structure results in impaired cleavage rates suggesting that CasE recognizes its substrate not only based on secondary structure, but also on sequence. Interestingly, the *E. coli* CasE homologue can also bind and cleave this same substrate, due to conservation in repeat sequence amongst the Ecoli subtype.

A minimal substrate for crRNA binding and cleavage was determined through gel electrophoretic mobility shift assays and cleavage rate analysis. A minimal substrate for RNA binding affinity is comprised of an oligonucleotide

Figure 2-13. CasE minimal substrate definition. From 5' and 3' truncation experiments the minimal substrate for efficient crRNA cleavage was defined as the central 21 nts shown above.

with three nucleotides removed from the 5' end of the repeat. Removal of nucleotides past the cleavage site did not have a significant effect on affinity. A minimal substrate for RNA cleavage is comprised of an oligonucleotide with 3' bases past the site of cleavage. The removal of these nucleotides had no effect on binding affinity. Taken together these results indicated that a minimal RNA substrate for binding and cleavage consists of an RNA that contains the central twenty-one nucleotides of the repeat sequence (Figure 2-13).

Based on experiments comparing an oligonucleotide which has strengthened base-pairing at the base of the stem, it appears that separation of the strands at the base of the helix is important for cleavage activity but not binding. This suggests that a conformational change may occur which is not necessary for efficient binding but may somehow facilitate catalysis. This change in conformation will be further explored in Chapter 3 which presents structural data that is consistent with this requirement for RNA strand separation.

The *in vitro* characterization experiments presented in this chapter contribute useful information about the mechanism of crRNA processing by Ecoli subtype CRISPR-Cas systems and complement the work carried out in *E. coli* K12 (Brouns *et al*., 2008). However, questions remain regarding CasE's mode of RNA recognition and catalysis. The basic cleft of the protein containing several is likely the site of binding, however the exact orientation of the RNA is unknown. It is unclear which amino acids are important for the sequence specific substrate recognition and which are involved in catalysis. The universally conserved histidine (His26 in *T. thermophilus*) has been proposed to be involved in catalysis based on mutational analysis of a recombinant *E. coli* CasE homologue (Brouns *et al*., 2008). However, it is possible that this well conserved histidine may not be involved in catalysis, rather it could be involved in RNA recognition or the maintenance of structural stability. In order to address these issues the *T. thermophilus* CasE enzyme was characterized structurally through the solution of three protein•RNA X-ray crystal structures, the results of which will be presented in Chapter 3.

2-5. Materials and Methods

2-5.1. Cloning, expression and purification of *Thermus thermophilus* CasE.

Full-length *T. thermophilus* CasE was PCR amplified from genomic DNA (ATCC 27634D-5) using oligonucleotide primers containing EcoRI and BamHI restriction sites and cloned into the pET-30a(+) vector (WT, Y23F). The CasE expression plasmid was transformed into *E. coli* Rosetta cells which were grown to an OD_{600} of ~0.8 before induction of protein expression with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 12 h at 24 ºC. Cells were pelleted, frozen and lysed at 4 °C for 30 min (100 mM NaCl, 20 mM Tris-HCl pH 9.0, 1 mM 2-mercaptoethanol, 20 mM imidazole, 1 μg/ml lysozyme, 1 mM PMSF) followed by sonication. The lysate was cleared by centrifugation at 15,000 RMP for 30 min. Cleared lysate was bound to a Ni Sepharose 6 Fast Flow column (GE) and eluted with a buffer containing 200 mM imidazole. Protein was concentrated and further purified by heating to 65 °C for 10 minutes to remove excess contaminating proteins followed by centrifugation. The resultant His_6-tagged CasE fusion proteins were purified by Superdex 75 and cation exchange chromatography. Protein was dialyzed overnight into a buffer containing 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 15% glycerol, 0.5 mM 2-mercaptoethanol, 0.5 mM EDTA. Protein used for activity/binding assays was aliquoted and stored at -20 °C. At each step, protein purification was monitored by resolution on an SDS-PAGE gel followed by coomassie staining.

2-5.2. RNA preparation.

RNAs were designed to model a full or partial CRISPR repeat and were purchased from Integrated DNA Technologies (Skokie, Il), and used without further purification.

2-5.3. Identification of Cleavage Site.

For RNaseT1 digestion, ^{32}P-5'-end-labeled RNA was incubated in a buffer containing 20 mM sodium citrate, 7 M urea, 1 mM EDTA, 80 mM HCl and 0.2 µg/µl yeast tRNA, successively for 2 min at 95 ºC, at 4 ºC for 2 min, and at 55 ºC for 2 min. 60 U of RNase T_1 was added to reactions and time points were taken between 1-3 min by quenching with 400 µl of 300 mM Sodium Acetate and 100 µl phenol. Following phenol-chloroform extraction, reactions were precipitated at -20 ºC in 70% ethanol for 20 min and pelleted by centrifugation at 15000 rpm at 4 ºC for 20 min. Base hydrolysis was carried out in buffer D (20 mM HEPES, pH 7.9, 0.1 M KCl, 0.5 mM dithiothreitol) in the presence of 5'-^{32}P-radiolabeled RNA and 0.2 µg/ul. 1.3 µl of 1M Sodium Hydroxide was added to a 13.7 µl reaction, reactions were quenched at 10 s, 30 s, and 1 min by addition of 400 µl of 300 mM Sodium Acetate and 100 µl phenol. Reactions were phenol/chloroform extracted and ethanol precipitated as described above and resolved on a 20% 8 M urea 29:1 polyacrylamide sequencing gel.

2-5.4. Characterization of CasE-dependent 3' RNA cleavage product.

The 28 nucleotide repeat RNA (~1 pmol) was incubated with 30 ng of purified recombinant CasE protein (100 mM NaCl, 10 mM Tris-HCl pH 9.0, 100 µM EDTA, and 500 µM 2-mercaptoethanol) for 30 min at 55 ºC followed by phenol/chloroform extraction and ethanol precipitation. RNA was 5' radio-labeled using γ-^{32}P-ATP (Perkin Elmer) using T4 kinase (Invitrogen) according to the manufacturer's instructions and analyzed on a 20% 29:1 8M urea sequencing gel. Gels were exposed to a phosphor screen (Molecular Dynamics) and scanned with a Storm 840 Phosphorimager (Molecular Dynamics).

2-5.5. Characterization of CasE-dependent 5' RNA cleavage product.

Oxidation/elimination was performed essentially as described (Igloi & Kossel, 1985). The 5'-^{32}P-radiolabeled 28 nucleotide repeat RNA or 21 nucleotide repeat RNAs were incubated for 30 min at 55 ºC both in the presence and absence of CasE followed by phenol/chloroform extraction and ethanol precipitation. The samples were resuspended in 37.5 µl of water, 50 µl 0.12 M borate/boric acid buffer pH 8.6, 12.5 µl 200 mM sodium periodate and incubated at 0 ºC in the dark for 60 min. Reactions were quenched with 10 µl of glycerol before phenol chloroform extraction and ethanol precipitation. Reactions were resuspended in 20 µl of water and 20 µl of 2 M lysine-HCl pH 9.3 and incubated at 45 ºC for 90 min followed by phenol chloroform extraction and ethanol precipitation. Reactions were analyzed on a 20% 29:1, 8 M urea sequencing gel. Gels were

exposed to a phosphor screen (Molecular Dynamics) and scanned with a Storm 840 Phosphorimager (Molecular Dynamics).

2-5.6. Gel mobility shift and cleavage assays.

5'-^{32}P-radiolabeled synthetic RNA substrates (50 - 100 × 10^3 cpm) were pre-heated to 55 ºC before incubation in 10 µl reactions containing 0-10 µM CasE protein (100 mM NaCl, 10 mM Tris-HCl pH 9.0, 100 µM EDTA, 500 µM 2-mercaptoethanol and 20 µg/µl yeast tRNA). Reactions were incubated for 30 min at 55 ºC before being immediately loaded onto a 6 % tris-glycine polyacrylamide gel and run at 150 volts for 1.5 hours. Dried gels were exposed to a phosphor screen (Molecular Dynamics) and scanned with a Storm 840 Phosphorimager (Molecular Dynamics). For cleavage assays, reactions were carried out with an excess of RNA (5-fold) at various temperatures and quenched at various time points in an equal volume of loading dye containing 2% SDS, 7 M urea before run on a 20% 29:1 8 M urea sequencing PAGE. Gels were exposed without drying to a phosphor screen (Molecular Dynamics) and scanned with a Storm 840 Phosphorimager (Molecular Dynamics). Initial cleavage rates were determined in triplicate using a 5 fold excess of ^{32}P 5' radio-labeled 28 nucleotide repeat RNA at 22 °C. Samples were taken at time points ranging from 30 s to 1 h, resolved on a sequencing gel and the data was analyzed using the Imagequant software. Comparison of substrates rates were carried out using a saturating amount of protein (1 µM) and 100 *f*M of 5'-^{32}P-radiolabeled RNA substrate at 65 °C. Time

points were quenched by adding an equal volume of loading dye containing 2% SDS and 8 M Urea and run on a 20% 8M Urea 29:1 polyacrylamide sequencing gel.

2-5.7. Oligonucleotides Used in this Work.

Oligonucleotide	5' → 3'
Cognate crRNA	GUA GUC CCC ACG CGU GUG GGG AUG GAC C
Deoxy G crRNA	GUA GUC CCC ACG CGU GUG GGdG AUG GAC C
5' Cleavage Product	GUA GUC CCC ACG CGU GUG GGG
3' Cleavage Product	AUG GAC C
DNA	GTA GTC CCC ACG CGT GTG GGG ATG GAC C
Hairpin RNA (R/G)	GGG UCC UCA UUA AGG UGG GUG GGA AUA GUA UAA CAA UAU GCU CAA UGU UGU UAU AGU AUC CCA CCU ACC CUG AUG UGU C
ssRNA (miR16)	UAG CAG CAC GUA AAU AUU GGC
Mutant #1	GUA UAC CCC ACG CGU GUG GGG UGG GAC C
Mutant #2	GUA GAG CCC ACG CGU GUG GGC UUG GAC C
5' -1	UA GUC CCC ACG CGU GUG GGG AUG GAC C
5' -2	A GUC CCC ACG CGU GUG GGG AUG GAC C
5' -3	A GUC CCC ACG CGU GUG GGG AUG GAC C
5' -4	GUC CCC ACG CGU GUG GGG AUG GAC C
5' -3, 3' -4	GUC CCC ACG CGU GUG GGG AUG
5' -3, 3' -5	GUC CCC ACG CGU GUG GGG AU
5' -3, 3' -6	GUC CCC ACG CGU GUG GGG A

5' -3, 3' -4 U/C	GCC CCC ACG CGU GUG GGG AUG
5' -3, 3' -6 U/C	GCC CCC ACG CGU GUG GGG A

2.6. References

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

Brouns, S. J., Jore, M. M., Lundgren, M. & other authors (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**, 960-964.

Calvin, K., Xue, S., Ellis, C., Mitchell, M. & Li, H. (2008). Probing the catalytic triad of an archaeal RNA splicing endonuclease. *Biochemistry* **47**, 13659-13665.

Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**, 3489-3496.

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

Díez-Villaseñor, C., Almendros, C., García-Martínez, J. & Mojica, F. (2010). Diversity of CRISPR loci in Escherichia coli. *Microbiology* **156**, 1351-1361.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & 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.

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., & MacMillan, A.M. (2011). Recognition and Maturation of Effector RNAs in a CRISPR Interference Pathway. *NSMB*, Accepted.

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

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

Igloi, G.L. & Kossel, H. (1985). Affinity electrophoresis for monitoring terminal phosphorylation and the presence of queuosine in RNA. Application of polyacrylamide containing a covalently bound boronic acid. *Nucleic Acids Res.* **13,** 6881-6898.

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

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

Makarova, K., Grishin, N., Shabalina, S., Wolf, Y. & Koonin, E. (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.

Marraffini, L. A. & Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181-190.

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

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

Xue, S., Calvin, K. & Li, H. (2006). RNA recognition and cleavage by a splicing endonuclease. *Science* **312**, 906-910.

Chapter 3[1]

Structural Characterization of *T. thermophilus* CasE[2]

[1] Adapted from Gesner *et al*., (2011). *NSMB.* Accepted

[2] Matthew J. Schellenberg collected the data sets and solved the X-ray crystal structures presented in this chapter.

3-1. crRNA Endonucleases.

The prokaryotic CRISPR-Cas system is a genetic defence mechanism in which bacteria and archaea are able to acquire heritable resistance to phage and plasmids (Karginov & Hannon, 2010; Marraffini & Sontheimer, 2010c). An essential process of this system is the production of mature effector CRISPR RNAs (crRNAs) from a long precursor crRNA (pre-crRNA). This is carried out by metal-independent crRNA endonucleases, four of which have been identified and characterized (Brouns *et al*., 2008; Carte *et al*., 2008, Haurwitz *et al*., 2010). Although these proteins share little sequence similarity, structurally and functionally they are analagous. The *P. furiosus*, Cas6 protein specifically cleaves the CRISPR transcript eight nucleotides from the 3' end of a single-stranded repeat retaining the mature product (Carte *et al.*, 2008; Carte *et al.*, 2010). The *P. aeruginosa* Csy4 protein possesses a similar metal-independent endonuclease activity towards a stable RNA stem-loop (Haurwitz *et al.*, 2010). In *E. coli* K12, CasE processes crRNA as part of the multi-protein Cascade complex (CRISPR-Associated Complex for Antiviral Defence) (Brouns *et al.*, 2008). In Chapter 2 , the basis of specific pre-crRNA recognition and processing by the *T. thermophilus* CasE homologue was characterized. In this Chapter, the structural characterization of *T. thermophilus* CasE•RNA complexes will be described.

3-2. T. thermophilus CasE.

Although molecular and biochemical techniques are useful in the characterization of an enzyme, structural characterization can provide a wealth of mechanistic knowledge that is otherwise not easily obtainable. Together, these two disciplines can determine answers to pertinent questions with respect to substrate recognition and catalytic mechanism. Thermophilic proteins are useful targets for crystallization as they can be very stable compared to mesophilic homologues (Jenney & Adams, 2008). As such, the *T. thermophilus* CasE homologue makes an excellent target for crystallographic studies.

3-2.1. CasE X-ray Crystal Structure.

The *T. thermophilus* CasE gene on the pTT27 mega plasmid encodes a ~22 kDa protein which is 30% identical to the *E. coli* homologue. Its basic theoretical PI of 9.59 is consistent with its observed role as a nucleic acid binding protein. The CasE structure has been reported and features tandem ferredoxin-like domains with characteristic βαββαβ folds each consisting of a four-stranded anti-parallel β-sheet with the two α-helices on one side (Ebihara *et al.,* 2006). The two β-sheets pack together to form an extended β-platform flanked by the two α-helices from each domain. The X-ray crystal structure of *T. thermophilus* CasE was solved at 1.85 Å resolution by molecular replacement using PDB entry 1WJ9

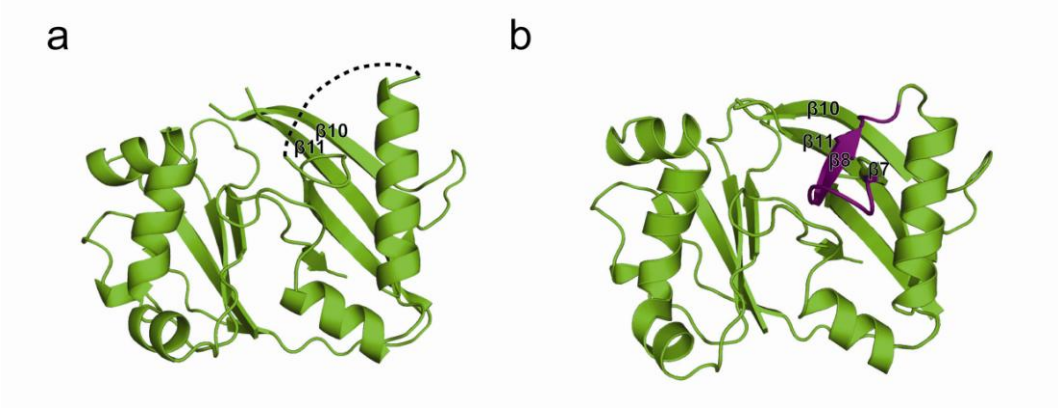

Figure 3-1. Comparison of CasE Structures. (a) Ribbon diagram of the published structure of *T. thermophilus* CasE, shown by dashed line is the disordered region which makes up the β7-β8 hairpin (PDB 1WJ9; Ebihara et al., 2008). (b) Ribbon diagram of the structure of CasE solved by molecular replacement to 1.85Å resolution. Shown in purple is the region that is disordered in PDB 1WJ9 which forms the β7-β8 hairpin. In both structures part of β10-β11 hairpin is disordered.

as a search model[1]. The structure is essentially the same as that published by Ebihara and colleagues, but is slightly more complete (Figure 3-1). Amino acids 108-115, disordered in the reported structure, form the β7 and β8 hairpin in this structure. Notable in both structures is a disordered loop extending from the adjacent β-10/β-11 strands.

3-2.2. Comparison with RRM

A protein structure comparison search to the CasE structure was performed using the Dali server in which the 3D structures of two RNA recognition motifs (RRM) containing proteins, the Sex-lethal protein (Sxl) and

[1] Matthew J. Schellenberg collected the data set and solved the CasE X-ray crystal structure

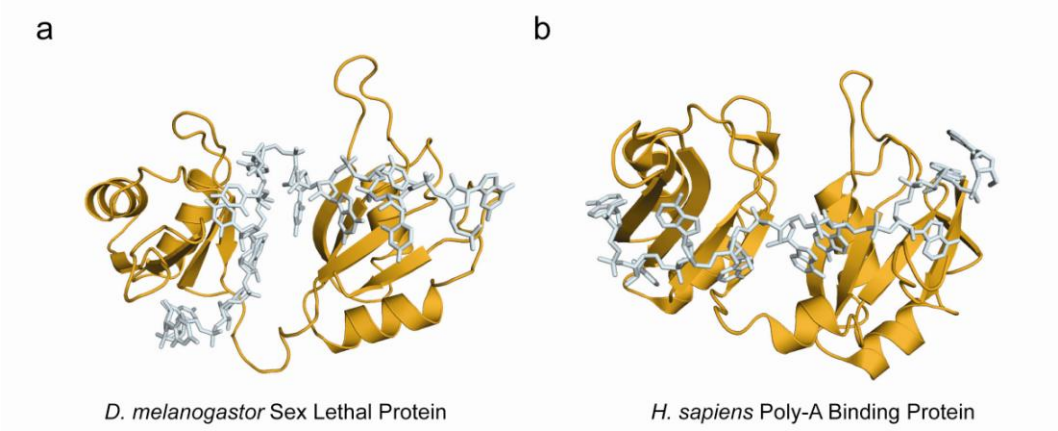

Figure 3-2. RNA Binding mode of RRM domains. (a) Ribbon diagram of *D. melanogaster* Sxl protein bound to RNA (PDB 1B7F; Handa *et al.*, 1999). (b) Ribbon diagram of *H. sapiens* PABP bound to poly-A RNA (PDB 1CVJ; Deo *et al*., 1999). Both proteins bind RNA across the β-platform formed between the two RRM domains.

polyadenylate binding protein (PABP) were identified (Holm & Rosenström, 2010). RNA-binding by Sxl and PABP, each containing tandem RRMs, which is a subclass of the ferredoxin-like fold, is mediated across an extended β-surface (Figure 3-2; Deo *et al.*, 1999; Handa *et al.*, 1999). It has been suggested that the analogous surface of CasE might play a role in RNA binding (Ebihara *et al.*, 2006). The electrostatic surface potential representation of this protein suggests that the side opposite the β-platform is the RNA interaction surface as it is more basic. This surface contains a basic groove which is ~ 25 Å wide and 28 Å high consistent with the binding of a 7 bp dsRNA stem, found in crRNA (Figure 3-3).

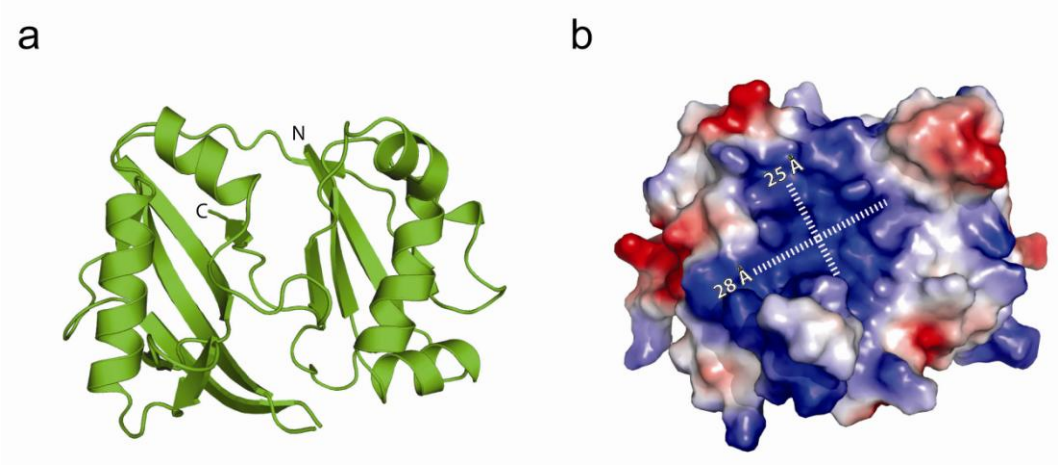

Figure 3-3. ***T. thermophilus*** **CasE.** (a) Ribbon diagram of 1.85 Å X-ray crystal structure of *T. thermophilus* CasE. (b) Electrostatic surface potential representation of CasE. Indicated are the dimensions of the basic groove.

3-2.3. Comparison with other crRNA Endonucleases

Two structures identified by the Dali structural comparison were other known crRNA endonucleases. *P. furiosus* Cas6, a functional homologue of CasE, harbours a similar domain structure although their primary sequences are dissimilar (12% identity, RMSD of 3.2) (Carte *et al.*, 2008). Cas6 contains two tandem ferredoxin-like folds very much like CasE. The recently reported structure of *P. aeruginosa* Csy4 bears some similarity to CasE (10% identity, RMSD of 4.2) (Haurwitz *et al.*, 2010). This structure features a single N-terminal ferredoxin-like fold (compared to the two ferredoxin-like folds of CasE and Cas6) and a separate C-terminal domain that includes two α-helices joined to the main body by extended linker sequences (Haurwitz *et al.*, 2010). These crRNA endonucleases have dissimilar primary sequences and appear to have evolved separately to perform a common functional role.

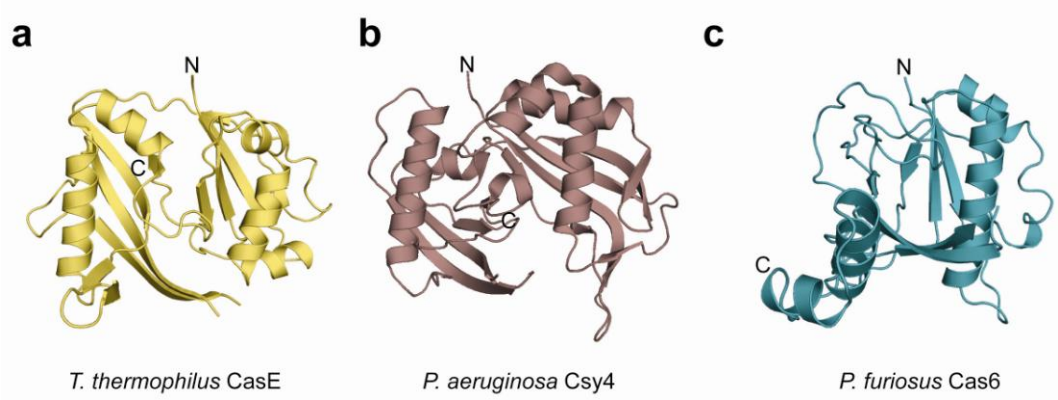

Figure 3-4. X-ray crystal structures of CasE and its functional homologues. (a) Ribbon diagram of *T. thermophilus* CasE which contains two ferredoxin-like folds. (b) Ribbon diagram of *P. furiosus* Cas6 which has a similar fold to CasE (PDB 3I4H; Carte *et al.*, 2008). (c) Ribbon diagram of *P. aeruginosa* Csy4 which contains a single ferredoxin-like fold with a C-terminal alpha helical domain joined by a linker region (PDB 2XLI; Haurwitz *et al.*, 2010).

3-3. Protein•RNA Crystal Structures.

Although the CasE protein structure provided a modest amount of new structural information, elucidating the mode of RNA binding and catalysis through the solution of protein•RNA co-crystal structures would be crucial. Crystal structures of *T. thermophilus* CasE complexed to three RNA substrates were solved through molecular replacement[2]. The structure of CasE bound to a non-cleavable 19 nucleotide RNA modelling a minimal CRISPR repeat containing deoxyG at position 21 was solved to 3.2 Å resolution (deoxyG Structure; Figure 3-5a)[2]. The X-ray crystal structure of CasE bound to an 18 nucleotide RNA mimicking the 5' cleavage product but lacking the cyclic 2',3' phosphodiester was solved to 2.3 Å (Product Mimic Structure, Figure 3-5b)[2].

[2] Matthew J. Schellenberg collected the data sets and solved the three CasE•RNA X-ray crystal structures

Finally, the 3.1 Å X-ray structure of CasE bound to a true 18 nucleotide 5' cleavage product containing a cyclic 2',3' phosphodiester was solved (Product Structure; Figure 3-5c)[2]. These complexes were formed by incubating the purified recombinant protein with the RNA substrate at 55 ºC for thirty minutes followed by separation of complexes through size exclusion chromatography. All three protein•RNA assemblies purified as 1:1 complexes, however strand exchange on crystallization yielded a dimeric structure in both the deoxyG and the product structures. The product mimic structure crystallized as a monomer but lacks two of the loop nucleotides (C13 and G14) presumably due to adventitious RNase activity. All structures nevertheless include the base-pairing of the predicted stem loop of the CRISPR repeat and thus represent the cognate complex. Each of these structures represents snapshots along the pathway of crRNA endonucleolytic cleavage. The deoxyG and product mimic structures taken together represents the substrate bound to the active site of CasE, in a conformation primed for catalysis. CasE remains bound to the product following cleavage and thus the product structure represents the post-cleavage conformation.

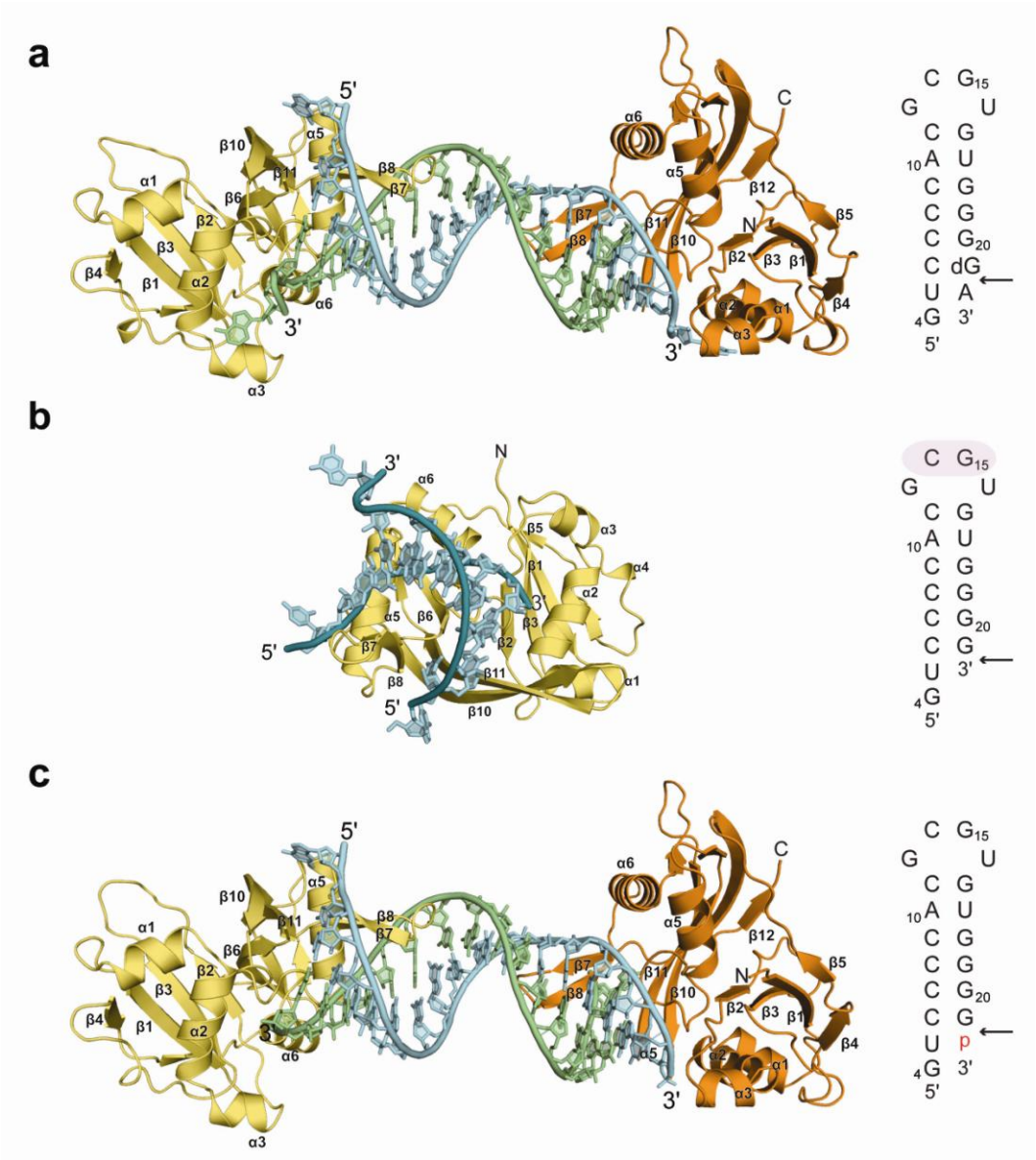

Figure 3-5. Structural basis for RNA recognition by CasE. (a) Ribbon diagram of the 3.2 Å X-ray crystal structure of *T. thermophilus* CasE bound to a 19 nucleotide mimic of an RNA repeat containing a deoxyG at position 21. Strand exchange during crystallization yielded a 2:2 dimeric structure related by a two-fold axis of symmetry. (b) Ribbon diagram of the 2.35 Å X-ray crystal structure of *T. thermophilus* CasE bound to an 18 nucleotide model of the 5' product of cleavage containing a 3' OH after G21. The nucleotides that are not present due to RNase activity are highlighted in violet. (c) Ribbon diagram of the 3.1 Å X-ray crystal structure of *T. thermophilus* CasE bound to a true 18 nucleotide RNA cleavage product containing a 2',3' cyclic phosphate at position 21. Strand exchange during crystallization yielded a 2:2 dimeric structure related by a two-fold axis of symmetry. The secondary structure of RNAs used for complex formation and crystallization of these structures are shown on the right.

3-3.1. RNA Recognition.

The overall architecture of all three complexes is very similar specifically with respect to the details of protein-RNA recognition as these three structures are superimposable. The essentially A-form RNA double-stranded helix is mounted on the protein opposite the extended β-surface corresponding to the positive face based on the electrostatic surface representation (Figure 3-6).

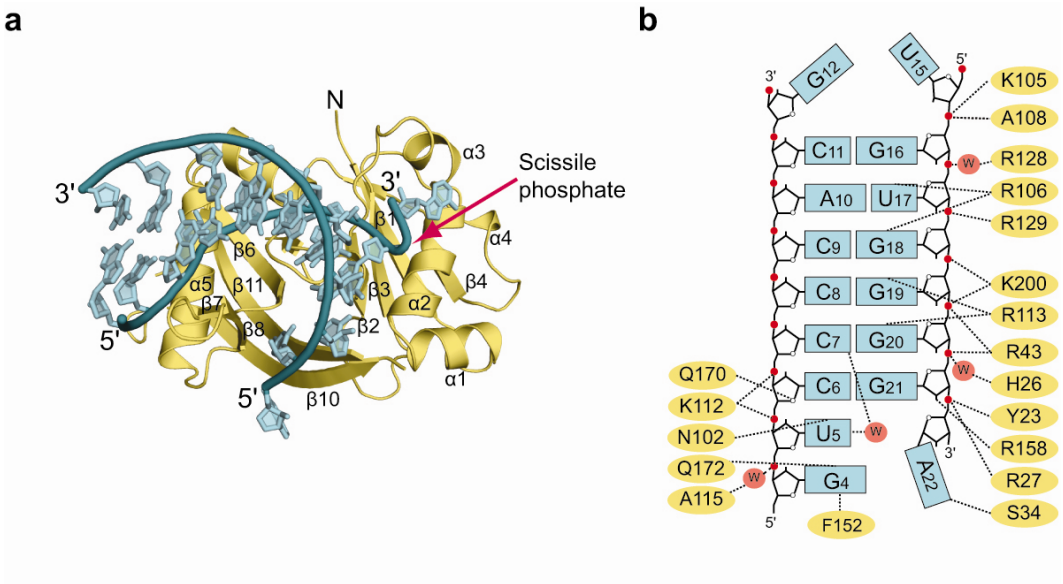

Figure 3-6. CasE•RNA Interactions. (a) Overview of the crystal structure of CasE bound to a deoxyG containing RNA. CasE is coloured yellow and the RNA is coloured cyan. (b) Summary of CasE•RNA interactions. Represented in this composite of interactions observed in the three CasE•RNA structures are sequence-specific base interactions, contacts with ribose functionalities, and both direct and water mediated contacts to the phosphodiester backbone (red).

The negatively charged phosphodiester backbone is stabilized through electrostatic interactions with positively charged amino acids including: Arg43, Lys105, Lys112, Arg128, Arg129, and Lys200 (Figure 3-6b; Figure 3-7; Figure 3-8). These amino acids are generally conserved amongst CasE homologues suggesting a common mode of RNA binding (Figure 3-7).

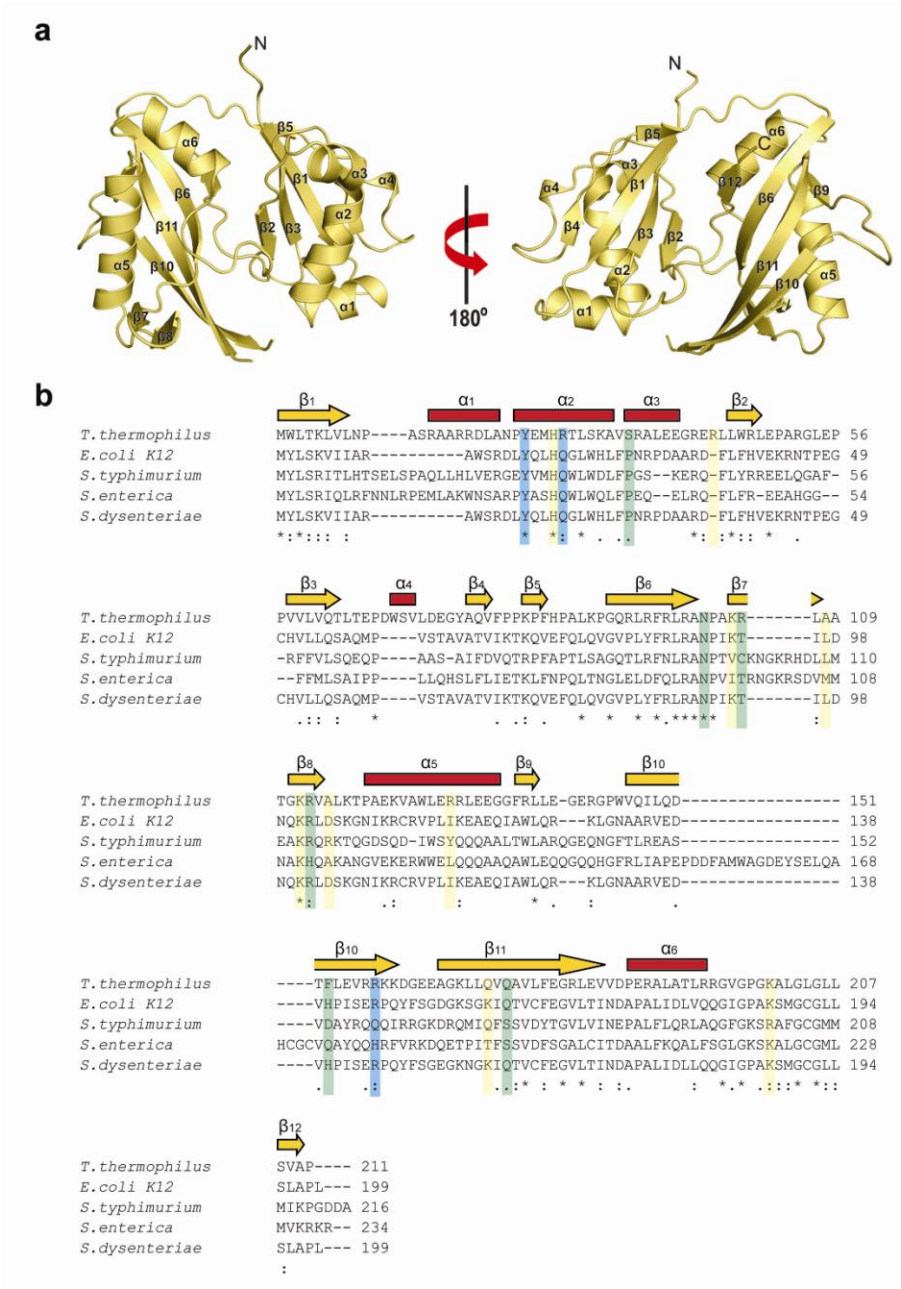

Figure 3-7. Structure of *T. thermophilus* CasE. (a) 1.85 Å resolution crystal structure of *T. thermophilus* CasE. Ribbon diagram showing front and back views. (b) Secondary-structure diagram depicting α-helices (red) and β-strands (yellow) of *T. thermophilus* Cse3. Also shown is an alignment with a subset of predicted CasE orthologs from *E. coli K12*, *S. dysenteriae* *S. typhimurium*, and *S. enterica* indicating sequence identity (*), conservation (:), and partial conservation (.). Proposed active site residues are highlighted in blue. Residues contacting the RNA bases are shown in green and residues contacting the phosphodiester backbone are shown in yellow.

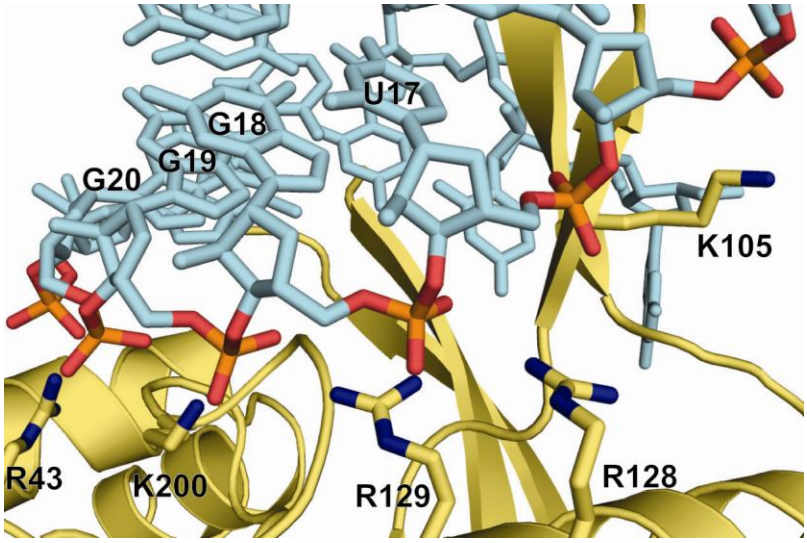

Figure 3-8. Interactions with phosphodiester backbone. Semi-conserved residues which form hydrogen bonds with the phosphodiester backbone along the basic cleft of the protein: R43, K105, R128, R129, and K200 are shown here in the 2.3 Å product mimic structure. These interactions stabilize the helix and help hold the RNA in position for cleavage.

In all three structures, a series of sequence-specific interactions exist between side chains of the short arginine-rich β7–β8 hairpin (aa 102-115) and the major groove of the dsRNA (Figure 3-9). Arg106 recognizes the bases of G16 and U17 through hydrogen bonds with the C6 carbonyl of G16 and the C4 carbonyl of U17 (Figure 3-9). Arg113 forms similar hydrogen bonding interactions with the C6 carbonyls of G19 and G20 (Figure 3-9). These interactions are reminiscent of the Tat RNA binding by the bovine immunodeficiency virus (BIV) Tar protein in which an arginine rich β-hairpin makes major groove contacts to the RNA stem-loop present near the 5′ terminus of all retroviral mRNAs (Puglisi *et al*., 1995).

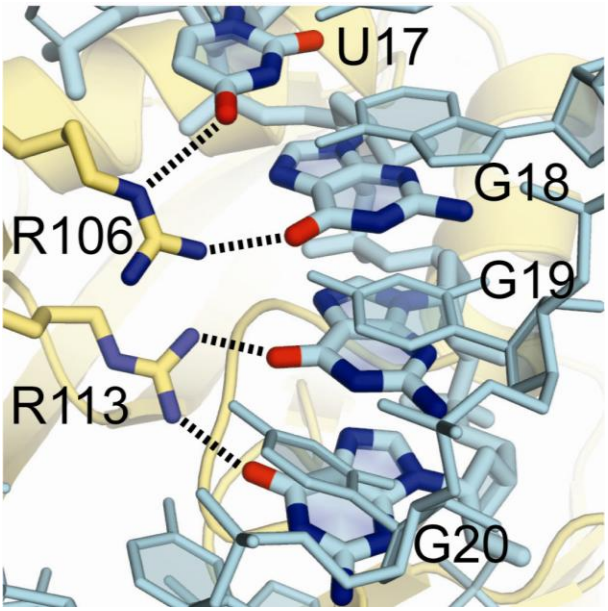

Figure 3-9. Details of CasE Major Groove Interactions. Sequence-specific recognition of the major groove base-pairs by arginine side-chains on the β7-β8 hairpin in the product mimic structure. Arginine 106 and 113 make base-specific interactions in the major groove of the A-form helix contacting the carbonyls of U17-G20.

At the 5' end of the repeat, a base-triple interaction is formed in which the exocylic amine of C7 that participates in the C7-G20 base-pair interacts with the O2 carbonyl of U5 through a water molecule (Figure 3-10). This base triple may help stabilize the separation of the U5•A22 base-pair. This also suggests a further level of sequence discrimination which may play a part in an induced fit binding mechanism.

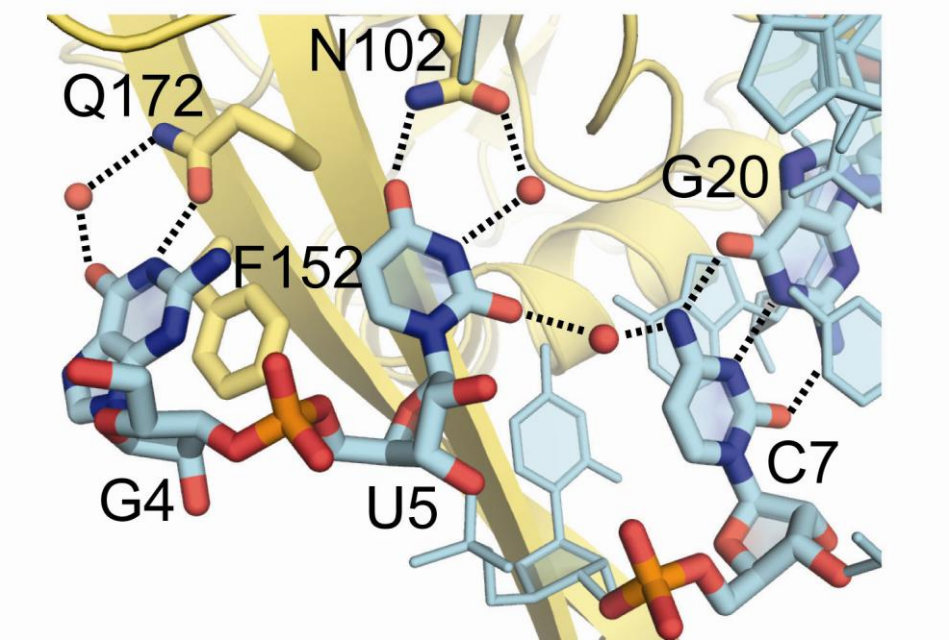

Figure 3-10. Details of CasE •RNA interactions. Sequence-specific recognition of unpaired 5' region of repeat RNA in the product mimic structure. Phe152 forms a stacking interaction with the purine base of G4 which is recognized by two hydrogen bonds to Gln172. U5 is recognized by hydrogen bonding to Asn102 and forms a water-mediated base triple with the C6-G21 base-pair.

3-3.2. Splaying of the Base of the Stem by the β10-β11 hairpin.

In the deoxyG structure, although the C6-G21 base-pair is contained within an A-form duplex, a sharp turn in the backbone following the deoxyG residue disrupts the predicted U5-A22 base-pair (Figure 3-11). The base of the RNA stem is bisected by the insertion of the β10-β11 hairpin (amino acids 146-176). This divergence from A-form is stabilized through several interactions including some from the β10-β11 hairpin (Figure 3-11). On the 5' end of the stem, G4 is recognised by Q172 which forms hydrogen bonds with the N7 amino

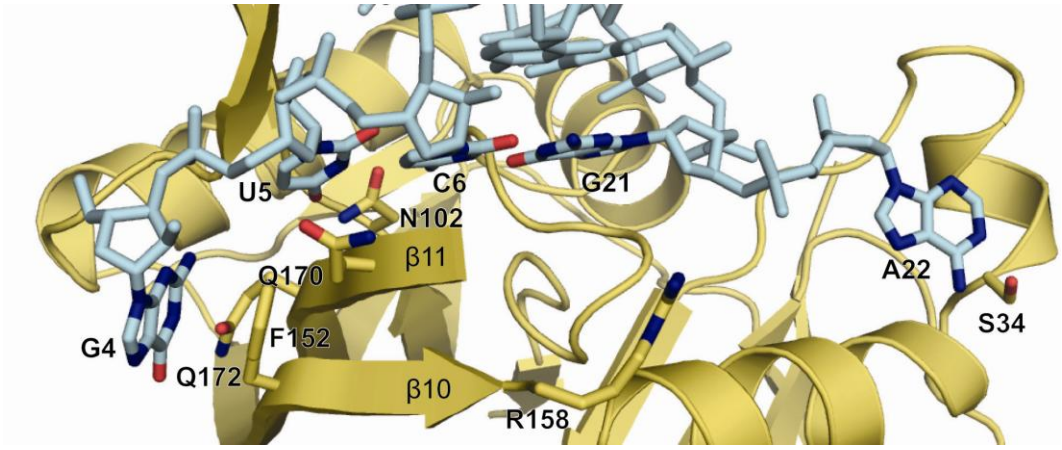

Figure 3-11. Splaying of the base of the stem. View of the base of the RNA stem in the deoxyG structure. Although C6 and G21 are base-paired and part of the A-form helix, this base-pairing is disrupted between U5 and A22 by the insertion of the β10-β11 hairpin. This conformation is stabilized by several interactions including: π stacking interactions between G4 and Phe152, hydrogen bonds between G4 and Gln172, Asn102 and U5, Gln170 and the ribose of C6, Ser34 and A22, and Arg 158 and the 2' OH of G21.

group and the O4 carbonyl (through water) and by Phe152 through π-stacking interactions (Figure 3-10, Figure 3-11). The C4 carbonyl and the N3 amino group (through water) of U5 are recognized by the highly conserved Asn102 residue (Figure 3-10, Figure 3-11). C6 is recognized by Gln170 which forms a hydrogen bond with the cyclic ribose oxygen (Figure 3-11). On the 3' end of the stem, A22 is held in position through a hydrogen bond between the Watson-Crick face and the side-chain of Ser34 (Figure 3-11). The position occupied by Ser 30, below the plane of the base is a tryptophan in most CasE homologues and thus may further stabilize this conformation of the RNA by π stacking in these homologues. The β10-β11 hairpin also contains Arg158 which forms a hydrogen bond to the 2' OH of G21 (Figure 3-11).

This bisection of the RNA stem by the β10-β11 hairpin may explain a curious observation with regards to CasE•RNA complex mobility. There is a clear change in mobility between CasE•RNA complexes that contained cleaved RNA versus uncleaved RNA by size exclusion chromatography. Protein•Product complexes and Protein•Product Mimic complexes eluted at ~172 ml while Protein•deoxyG RNA complexes eluted at ~176 ml (Figure 3-12a). This change in mobility was also observed by gel electrophoretic mobility shift assays. Complexes bound to deoxyG21 containing RNAs migrate more quickly through native gels than those bound to cleaved RNA or the 5' product mimic. As well mutant proteins that disrupt catalytic rates, but not binding affinity display two shifted bands (Figure 3-12b). For example, an electrophoretic mobility shift assay with CasE Y23F, a catalytically impaired mutant which will have cleaved ~50% of the repeat RNA after thirty minutes of incubation yields two shifted bands in which the lower band represents the CasE•Substrate complex and the upper the CasE•Product complex. This was confirmed by gel excision and denaturing gel electrophoresis. This result suggests that the CasE•Product complex is more compact than the CasE•Substrate complex. The two observed conformations may represent the fully base-paired RNA, which may be the predominant form in solution, and the splayed apart RNA bisected by the β10-β11 hairpin. In section

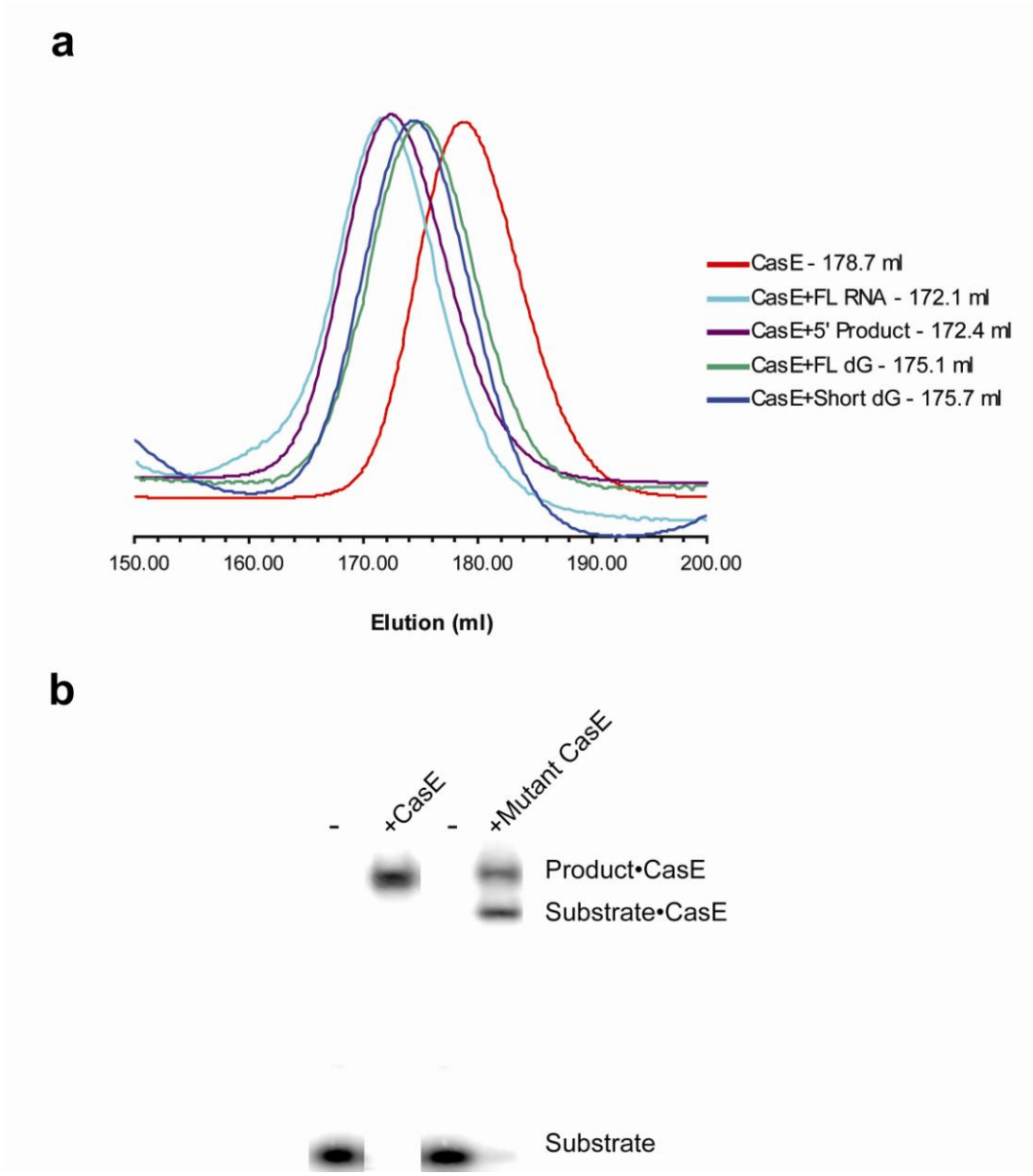

Figure 3-11. Distinct Mobility of Product•CasE and Substrate•CasE Complexes. (a) Overlay of elution curves from Superdex 75 size exclusion column. In red – the elution of CasE protein alone, in cyan – the elution of CasE and the full length RNA substrate which was confirmed to be cleaved, in purple – the elution of CasE and the 5' cleavage product model, in green – the elution of CasE and the full length deoxyG RNA, and in blue – the elution of CasE and a deoxyG RNA that extends only a single base past the cleavage site. (b) Electrophoretic mobility shift assay of the cognate 5'-^{32}P-labelled substrate with wild-type CasE and with a catalytically impaired point mutant (Y23F). The two shifted bands in lane four represent the product and the substrate complexed with the protein respectively.

2-3.5 it was demonstrated that the separation of the stem is necessary for activity as determined by experiments using RNA with extended base-pairing at the base of the stem. These structures confirm that this unwinding is necessary to position the RNA in a catalytically favourable conformation.

A consequence of the observed kinking of the backbone and splaying of the dsRNA helix seen in the deoxyG structure is the positioning of the scissile phosphodiester bond in a conformation consistent with in-line displacement of the A22 5'-oxygen leaving group by the G21 2'-hydroxyl. This means that the deoxyG complex is in a conformation primed for catalysis which is critical in establishing the mechanism of catalysis. A consequence of the separation of the U5-A22 base-pair by the β10-β11 hairpin is an interaction between R158 and the 2'OH of G21. Unlike the CasE•Product mimic structure the β10-β11 hairpin was not completely ordered in the CasE•deoxyG structure and the CasE•Product structure most likely due to crystal packing interactions. Nonetheless, this suggests that this conformational change serves to position the RNA in a catalytically favourable conformation and may also be a physical signal that the RNA is a bonified substrate that has been cleaved. Because processing of pre-crRNA by CasE takes place within the context of the multi-protein Cascade complex, the RNA-dependent ordering and extension of the β10-β11 hairpin may facilitate downstream events in the CRISPR pathway.

3-3.3. Active Site Components.

The organization of the CasE•RNA interface at the 3' end of the repeat has clear implications for the catalytic mechanism of RNA cleavage by the enzyme. The backbone conformation at the scissile phosphate suggests that the structure of the complex models an assembled active site primed for catalysis in the deoxyG structure. In the CasE active site of all three structures, the conserved Tyr23 is positioned proximal to the scissile phosphate to stabilize the charged intermediate (Figure 3-13a,b,c). The side chain of Arg27 is positioned to both stabilize the transition-state/intermediate and departing leaving group (Figure 3-13a). In the product mimic and the product structure, the side chain of Arg158 is positioned to interact with the 2' oxygen of G21; this side chain is less ordered in the deoxyG containing deoxyG structure presumably because of the lack of a substituent at the G21 2' position (Figure 3-13a,b,c).

A composite model based on the three CasE•RNA structures describes an assembled active site wherein Arg158 and Arg27 stabilize the nucleophile and leaving group respectively and Tyr23 interacts with the backbone non-bridging oxygens of the transition state intermediate (Figure 3-14).

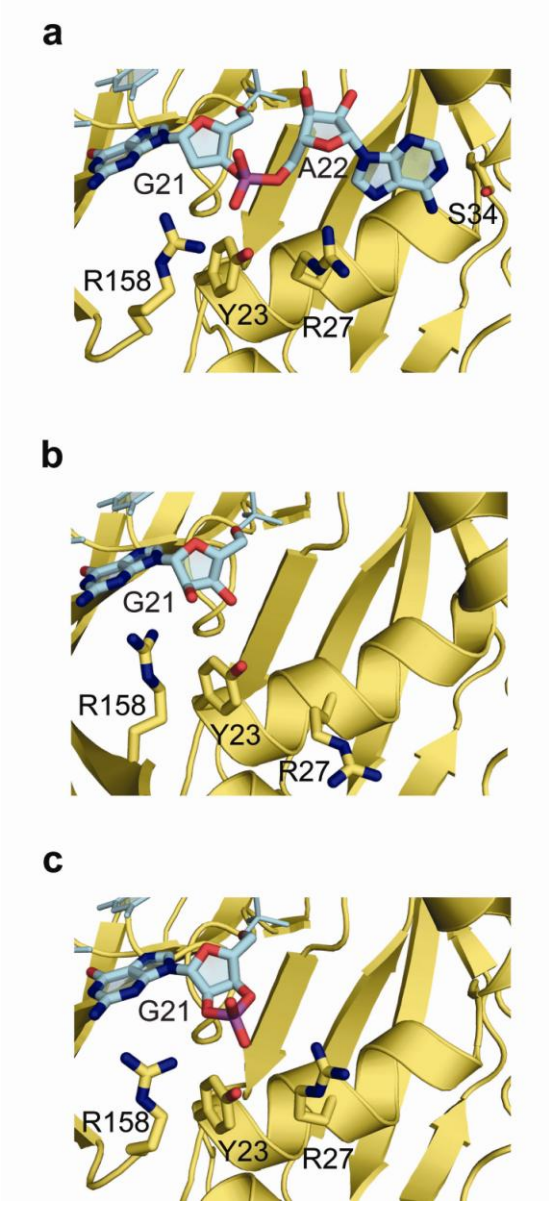

Figure 3-13. Structural basis for cleavage of pre-crRNA by CasE. (a) Detail of scissile phosphate environment in structure of CasE bound to deoxyG containing RNA. The conformation of the G21-A22 backbone is consistent with an in-line displacement mechanism by the 2' hydroxyl of G21 which is absent in this substrate. Tyr23 is positioned to stabilize the charged transition state/intermediate through interaction with non-bridging phosphate oxygens. The side-chain of Arg27 is positioned to stabilize the A22 5'OH leaving group. The scissile phosphate is coloured magenta. (b) Active site environment in structure of CasE bound to RNA mimicking the 5' cleavage product. The side-chain of Arg158 is ordered and positioned to interact with the 2' hydroxyl of G21. (c) Active site environment in structure of CasE bound to RNA product. The side-chain of Arg158 is ordered and positioned to interact with the 2' oxygen of G21. The side-chain of Tyr23 is positioned to stabilize the transition state/intermediate.

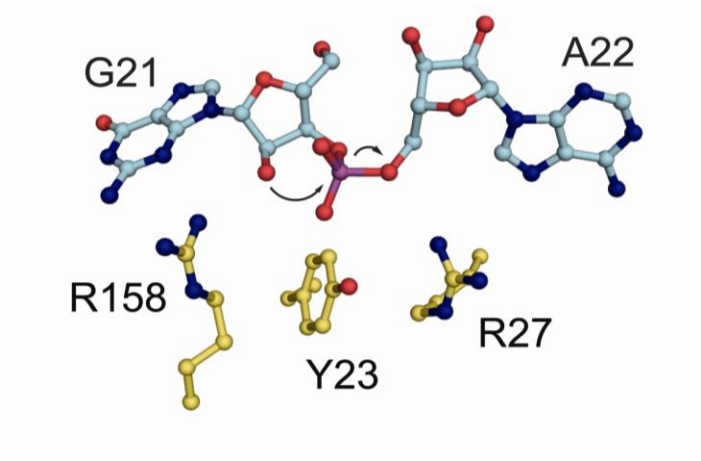

Figure 3-14. Catalytic model. The mechanism of RNA cleavage by CasE based on a composite of the three CasE•RNA structures reported here. R158 interacts with the 2'-hydroxy nucleophile while R27 stabilizes the 5' leaving group. The side chain of Y23 interacts with the non-bridging oxygens of the transition-state intermediate which may also be stabilized by R27.

We confirmed the importance of the proposed active site residues by functional analysis of the Y23F, R27A, and R158A CasE point mutants. While none of these is seriously compromised in binding the CRISPR repeat, the observed rate of cleavage of the R27A, Y23F, and R158A mutants were ~100 fold decreased at 22 °C, while the mutation of Arg157, involved in stabilization, of the β10-β11 hairpin, to alanine results in a modest 7 fold decrease in rate (Figure 3-15). Intriguingly, these mutations appeared less severe when rates were calculated from reactions incubated at 65 °C. The activity of the Y23F, R27A and R158A mutants were decreased by 15, 2.8, and 2.5 fold compared to wild-type at 65 °C. This disparity may be due to the ease of overcoming energetic barriers at increased temperatures. Perhaps at these increased temperatures, not all of the catalytic residues are required for cleavage. RNA in and of itself is an unstable

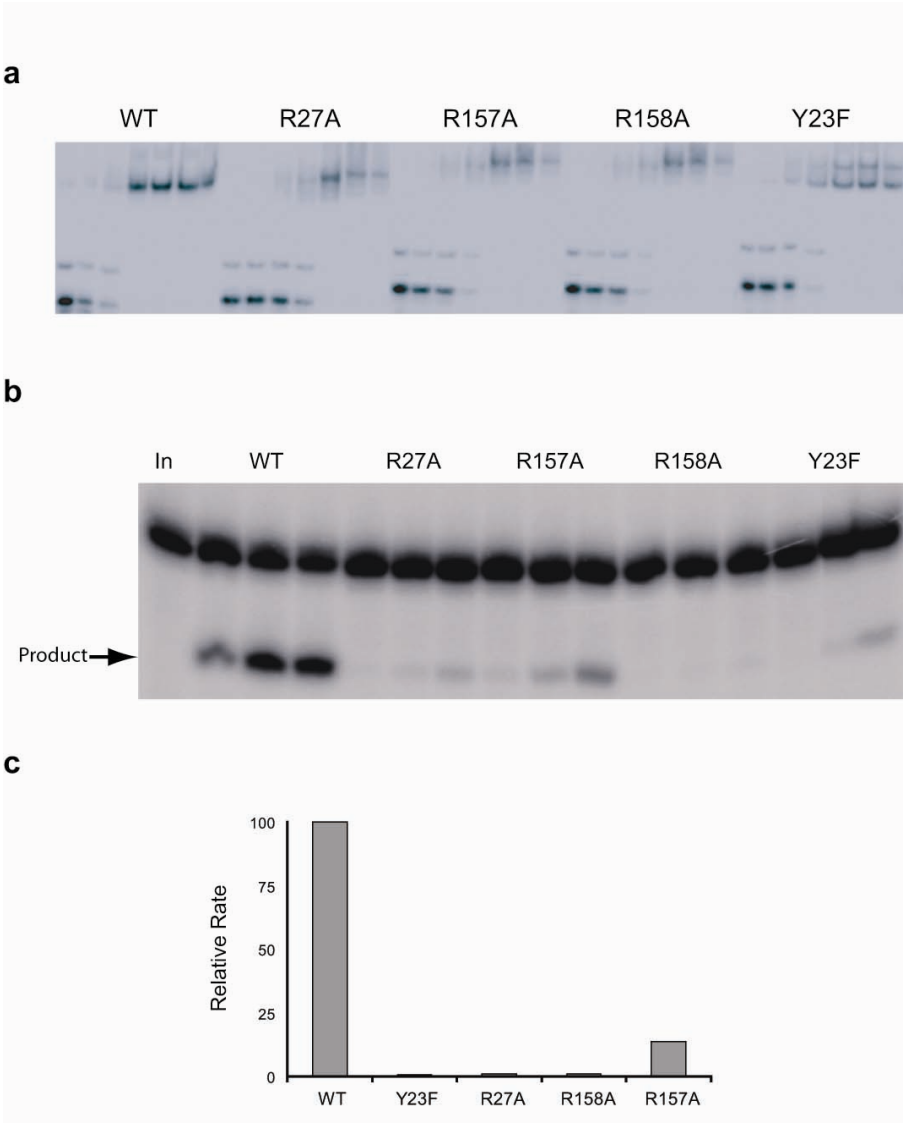

Figure 3-15. Characterization of point mutants. (a) RNA binding by *T. thermophilus* CasE point mutants. Gel mobility-shift assays for 5' ^{32}P end-labeled 28 nucleotide repeat RNA binding to CasE showing the affinity of wild type and four mutants with increasing concentrations (0.1-1000 nM) of protein. Only a modest change in affinity is observed between the mutants and wild-type (3-5 fold). (b) RNA cleavage by WT and four mutants of *T. thermophilus* CasE. Denaturing PAGE analysis for a cleavage assay of 5' ^{32}P end-labeled 28 nucleotide repeat RNA by CasE showing the activity of wild type and four mutants with increasing time (10 s, 300 s, 1800 s). (c) Comparison of cleavage rates for wild-type CasE and four point mutants. Quantification of initial rates indicates rate of cleavage of Y23A, R27A, R158A is reduced ~100 fold, and R157A ~7 fold compared to wild-type.

molecule hydrolytically due to the proximity of a potential nucleophile, the 2'OH, to the phosphodiester bond. It may be that proper positioning of the RNA through binding to CasE, is a very significant part of the "catalysis". Considering that only Tyr23 is well conserved amongst CasE homologues, this is a likely explanation.

3-3.4. Role of the Conserved His26

The absolutely conserved His26 has been proposed by others to be an active site residue as its' mutation to alanine causes catalytic impairment is not in position to play a direct role in catalysis in any of these structures (Brouns *et al*., 2008). In the 2.3 Å product mimic structure His26 interacts through a bound water with the phosphate one nucleotide 5' to the cleavage site (Figure 3-16). This is reminiscent of the direct interaction of an invariant histidine in the active site of Topoisomerase II with the DNA backbone (Schmidt *et al*., 2010). This water is indistinguishable in the electron density map of the product or the deoxyG structures. In all except the product structure, His26 makes a hydrogen bond with a Trp46 near the core of the protein (Figure 3-16). This hydrogen bond involves the lone pair of the histidine, rendering it unavailable to act as a base. This interaction is absent in the product structure possibly due to the low pH of the crystallization conditions (pH 6). In section 2-3.2, pH profile experiments showed that the enzyme loses activity at pH 5-6, suggesting that this interaction is

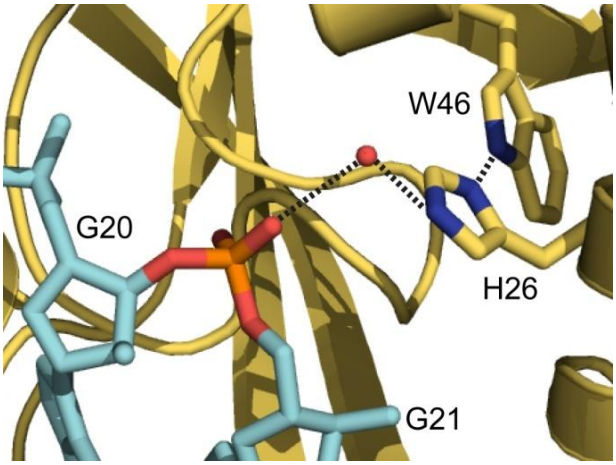

Figure 3-16. Position of Invariant Histidine. Close-up view of the conserved His 26 which forms a hydrogen bond with Trp 46 in all three protein•RNA structures. In the product mimic structure a hydrogen bonding interaction with the phosphate between G20 and G21 can also be modelled suggesting that this residue is not available or in position to contribute to catalysis.

important for the proper functioning of the enzyme. It also argues against His26 acting as an acid, as the pKa of histidine side chains is approximately 6.0.

To further investigate the role of His26 in catalysis, point mutants of this residue were cloned, expressed and purified. Mutation of the invariant His26 which is not in a position to contribute catalytically in any of the three structures caused the largest decrease in activity at 65 °C. Mutation of His26 to Ala, Gln, Trp, and Asn caused 37, 30, 15, and 5 fold decreases in activity compared to the wild-type (Figure 3-17). It is possible that this histidine which forms a hydrogen bond to the phosphate 5' to the scissile phosphate and to a tryptophan residue near the core of the protein may be important to the stabilize the structure of the protein. This is consistent with the observed instability of the point mutants

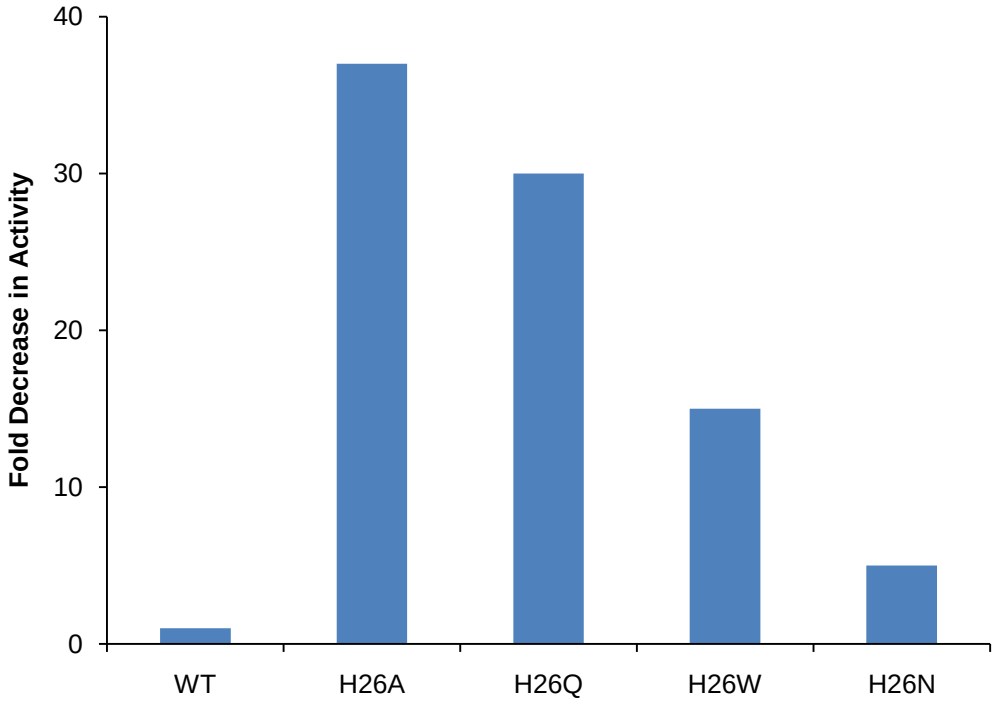

Figure 3-17. Activity of His26 Mutants. Shown here is the graphic representation of the fold decrease in CasE cleavage activity as a result of mutation of His26 to Ala, Gln, Trp, or Asn. Rates were determined by calculating the ratio of product to substrate in experiments performed at 65 ºC, pH 8.0 in triplicate using Imagequant software.

during protein preparation. Compared to wild-type and other mutants, the His26 mutants had a ~100 fold reduction in protein expression levels. Most of the protein expressed remained in the pellet of lysate following centrifugation suggesting that it is improperly folded. During size exclusion chromatography, only 10% of the His26 mutant proteins purified as monomers, the rest was aggregated. In comparison, the wild-type protein purified as ~90% monomeric. Also consistent with this result is that the H26N mutation which cannot perform the acid-base chemistry, but can form the hydrogen bond with the tryptophan is least impaired.

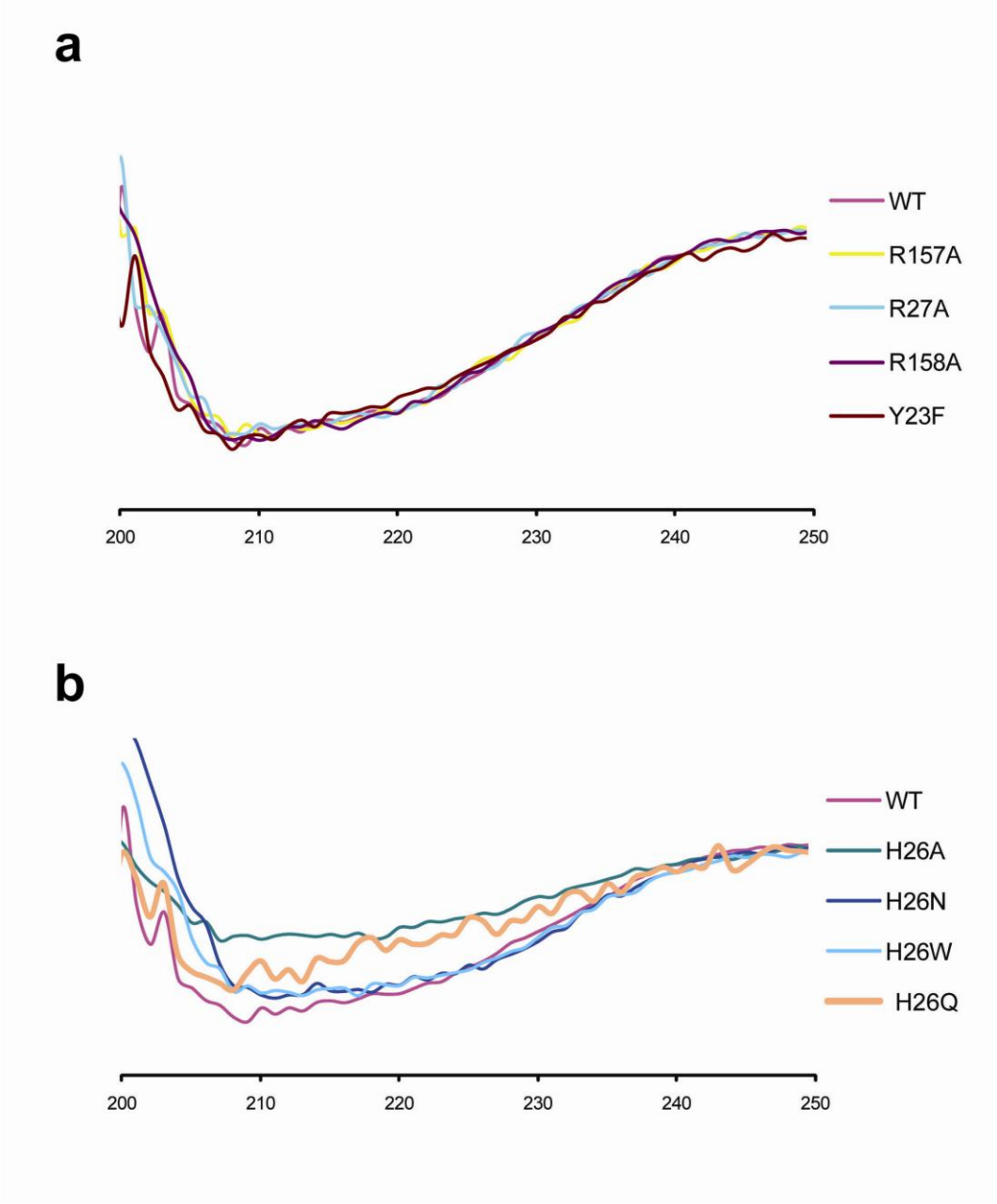

Figure 3-18. CD Spectroscopic Analysis of Point Mutants. (a) CD spectra of WT CasE compared with R157A, R24A, R158A, and Y23F. Absorbance is plotted on the Y-axis and the wavelength of light in nM is plotted on the X-axis. The WT displays similar spectra to all of the active site mutants. (b) CD spectra of WT CasE compared to H26 point mutants. All of the His26 mutants display a shallower curve than the wild-type suggesting a change in structure has occurred.

To further test this hypothesis, we used circular dichroism (CD) spectroscopy to make a rough comparison of the overall secondary structural features of the wild-type compared to the point mutants. Analysis of the changes in CD spectra allows a comparison of relative protein stability. The CD spectrum of the wild-type protein suggests a mixed α-helical/β-sheet structure, which is consistent with the crystal structures. While the R157A, R24A, R158A, and Y23F mutants had nearly identical spectra, all of the His26 mutants displayed a shallower curve with a slightly different shape indicating that a change in structure has taken place (Figure 3-18).

3-4. Discussion.

The model of the CasE active site derived from our structural analysis places two Arg residues, Arg158 and Arg27, in the position expected to be occupied by a general base/acid. The side chain of Arg27 may act as a general acid in the protonation of the leaving group. With respect to the role of Arg158, although there are some examples of Arg residues implicated as general bases in enzyme active sites, (Guillen Schlippe & Hedstrom, 2005) the high pKa of the Arg side-chain and the solvent exposure of Arg158 argues against this mechanism in the current case. An attractive alternative is that possibly both residues function as metal surrogates with a role similar to that of divalent metal ions in the classic "two-metal mechanism" common to many protein and RNA-based phosphoryl transfers (Steitz & Steitz, 1993). Arg158 may not necessarily

deprotonate the 2' OH, but rather increases the nucleophilicity of the 2' OH through electrostatic interactions. Likewise, Arg27 electrostatically stabilizes the 5' leaving group.

T. thermophilus CasE, a crRNA endonuclease, specifically binds and cleaves its pre-crRNA substrate into mature crRNAs. Electrostatic Interactions between the phosphodiester backbone and the conserved basic residues in the cleft of the protein make binding favourable. The sequence specific contacts are largely mediated by major groove contacts with the β7-β8 hairpin. The β10-β11 hairpin splays apart the base of the helix, supported by several hydrogen bonding interactions with the bases and phosphodiester backbone of G4, U5, and A22. The scissile phosphate is thus in a conformation which facilitates inline displacement by the 2' OH. The β10-β11 hairpin contributes a catalytic residue, Arg158, which may increase the nucleophilicity of the 2' OH. Another arginine, Arg27, is in position to possibly protonate the leaving group. A conserved tyrosine residue directly interacts with the non-bridging oxygens to stabilize the transition state intermediate. Although their structures and RNA substrates are distinct, similarities can be drawn between CasE's mode of RNA binding and catalysis and those of other known crRNA endonucleases.

3-4.1. Comparison to *P. aeruginosa* Csy4.

The X-ray crystal structure of *P. aeruginosa* Csy4 bound to a model of its RNA substrate containing a deoxyG residue at the site of cleavage (the base downstream of the scissile phosphate being disordered) has been solved by Jennifer Doudna's laboratory (Figure 3-19; Haurwitz et al., 2010). This structure depicts the Ypest subtype crRNA endonuclease bound to its' stable RNA hairpin substrate which bears striking similarities to the CasE·RNA structures despite the proteins and repeats being very much distinct. In *T. thermophilus* CasE and *P. aeruginosa* Csy4, highly specific RNA binding is largely mediated through major groove recognition by distinct modules, a short β-hairpin and an α-helix respectively, that extend from within the C-terminal regions of the proteins (Haurwitz *et al.*, 2010). In both cases, the arginine rich region makes specific major groove interactions with the stem of the hairpin (Haurwitz *et al.*, 2010). However, unlike CasE whose extensive interactions are focused primarily around the 3' strand, Csy4's interactions are focused on the 5' strand side of the duplex (Figure 3-19; Haurwitz *et al.*, 2010). Specifically, in Csy4 Arg115 recognizes G11 through hydrogen bond interactions, and Arg 114, 115, 118, and 119 make electrostatic interactions with the phosphodiester backbone of U7-G12 (Figure 3-19b; Haurwitz *et al.*, 2010). These interactions likely play a role in the positioning of the RNA helix with respect to the N-terminal ferredoxin-like fold and the C-terminal RNA recognition domain (Haurwitz *et al.*, 2010). A similar

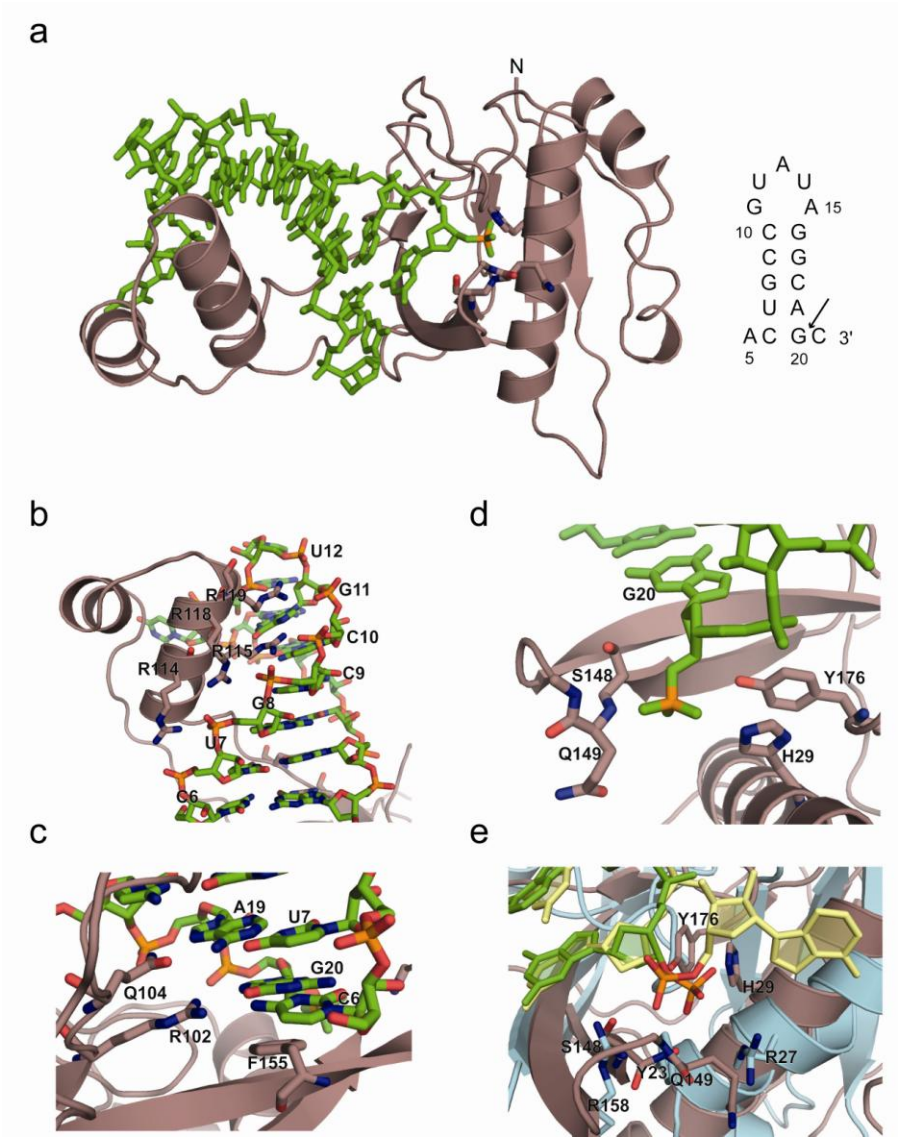

Figure 3-19. Csy4•RNA co-crystal structure. (a) Shown is the ribbon diagram of Csy4 crystal structure bound to crRNA mimic (PDB 2XLI-K; Haurwitz *et al.*, 2010)). Active site residues are shown in sticks. The RNA that was crystallized is on the right. (b) Arginine-rich α-helix/linker domain of Csy4 inserts into the major groove of the RNA helix forming hydrogen bonds with the bases and the phosphodiester backbone. (c) Specific recognition of A19 and G20 by Q104 and R102 respectively. F155 forms π-stacking interactions with C6. (d) Close up view of the Csy4 active site. Predicted catalytic residues are shown in stick format. (e) Superimposition of Csy4•RNA structure (protein violet, RNA green) with CasE•dG structure (protein cyan, RNA yellow). The position of Csy4's Gln149 suggests that it may function as CasE's Tyr23 in transition state stabilization.

π-stacking interaction to that of G4 and Phe152 in the CasE structure is seen between C6 and Phe155 in the Csy4 structure (Figure 3-19c; Haurwitz *et al.*, 2010). Like CasE's β10-β11 hairpin, Csy4's β6-β7 hairpin bisects the base of the stem and makes specific contacts with the RNA (Haurwitz *et al.*, 2010). This hairpin, like CasE's Arg158, provides one of the predicted catalytic amino acids, Ser148.

The active site cannot be clearly defined in the Csy4•RNA complex as the base and the sugar of C21 are disordered (Haurwitz *et al.*, 2010). The proposed members of the active site include: His29, Tyr176, Ser148, and Gln149 (Figure 3-19d). Alignment with the CasE•deoxyRNA structure supports the proposed interaction of the Ser148 side-chain with the 2' hydroxyl nucleophile and suggests stabilization of the leaving group by the conserved His29 (Haurwitz *et al.*, 2010). Indeed, mutation of these residues to alanine and cysteine respectively caused serious defects in activity (Haurwitz *et al.*, 2010). The conserved Tyr176 which is in close proximity to the scissile phosphate showed insignificant loss of activity when mutated to a phenylalanine (Haurwitz *et al.*, 2010). Superimposition with CasE indicates that in Csy4, stabilization of the cleavage transition state/intermediate, mediated in CasE by Tyr23, is likely a function of the main-chain N-H of Gln149 (Figure 3-19d).

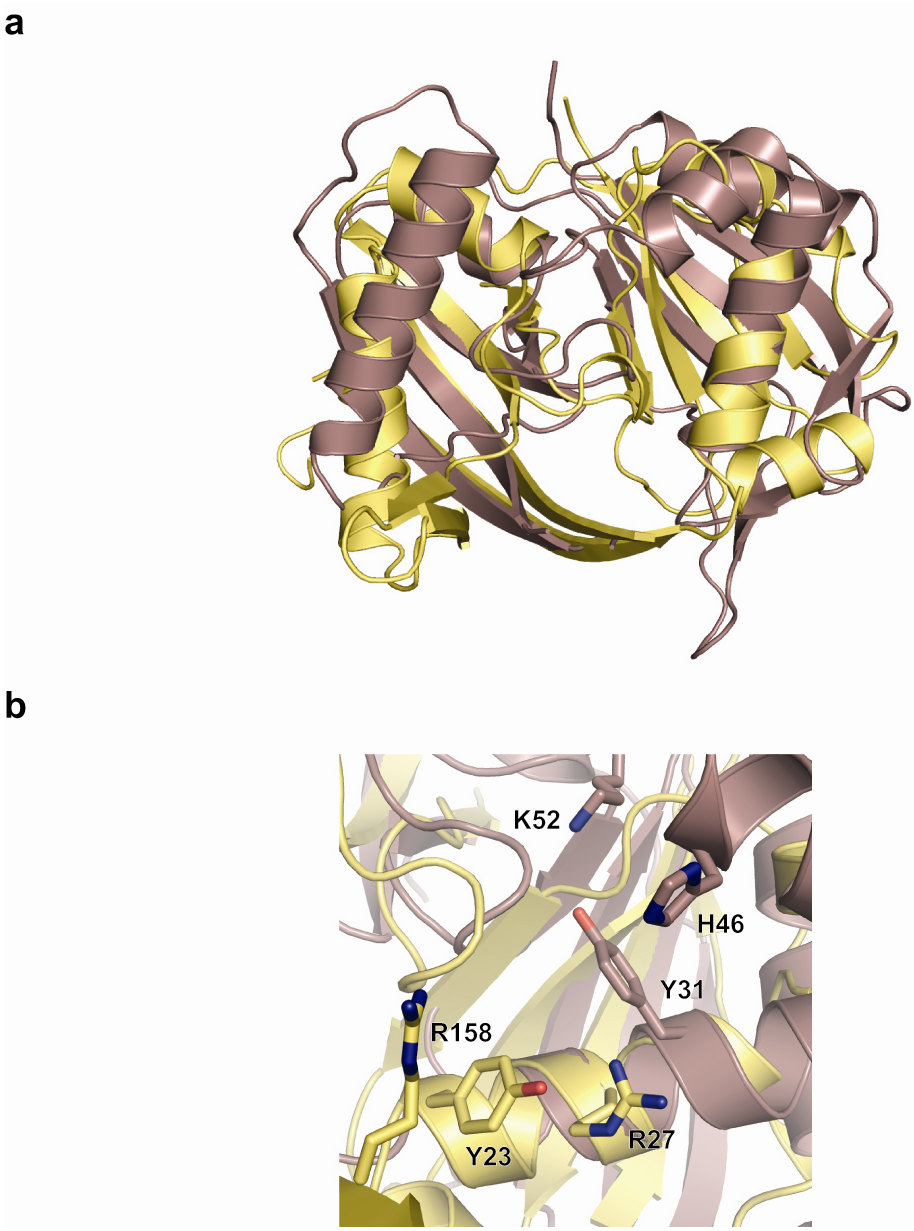

Figure 3-20. Comparison of Cas6 and CasE structures. (a) Superposition of *P. furiosus* Cas6 (violet, pdb entry 3I4H) with *T. thermophilus* CasE protein (yellow) from the CasE deoxy G RNA structure. (b) Comparison of the CasE active site with that proposed for Cas6 (Carte *et al*., 2008). Superposition of Cas6 showing proposed catalytic triad consisting of amino acids Lys52, His46, and Tyr31 with CasE highlighting active site residues Tyr23, Arg158 and Arg27.

3-4.2. Comparison to *P. furiosus* Cas6.

The *P. furiosus* Cas6 crRNA endonuclease shares a similar domain structure with CasE of two tandem ferredoxin-like folds (Figure 3-20a). Alignment of the CasE and Cas6 structures suggests that disordered regions of Cas6 may correspond to the major groove interacting β7-β8 hairpin as well as the helix bisecting β10-β11 hairpin (Figure 3-20b). However, these two enzymes cleave RNAs with different secondary structure; CasE recognizes a stable RNA stem-loop and Cas6 a ssRNA suggesting that their mode of binding may differ. Indeed, a recent structure of Cas6 bound to a crRNA mimic shows that the 5' end of the repeat sequence is specifically recognized by residues in a cleft between the two β-sheets (PDB 3PKM; Wang *et al.*, 2011). These interactions are on the opposite face to that corresponding to the CasE RNA binding face. The Terns group extrapolate that the RNA extends beyond the groove, wrapping around the protein to the active site.

In their analysis of the homologous Cas6 protein structure, Terns and co-workers suggested a catalytic triad of Tyr, Lys and His residues might function in catalysis of repeat cleavage analogous to the mechanism of the t-RNA splicing endonuclease (Calvin *et al.*, 2008; Carte *et al.*, 2008; Carte *et al.*, 2010; Xue *et al.*, 2006). The X-ray structure of the latter enzyme bound to RNA suggests that Tyr and His side-chains act as general base and acid in the deprotonation of the 2'-hydroxyl nucleophile and protonation of the 5' leaving group; the Lys side chain

is positioned to stabilize the transition-state/intermediate in the cleavage reaction (Calvin *et al.*, 2008). Alignment of the CasE and Cas6 structures suggests that the catalytic triad proposed for the latter may indeed function to promote RNA cleavage but with a disposition and specific function of catalytic residues distinct to that reported here for CasE (Figure 3-20a,b).

3-4.3. Implications.

Genetic and structural evidence argue for an evolutionary relationship between the pre-crRNA processing activity in distinct CRISPR systems (Carte *et al.*, 2008; Haft *et al.*, 2005; Haurwitz *et al.*, 2010; Kunin *et al.*, 2007). Superposition of the CasE structures reported here with the high-resolution structures of both the Csy4•RNA complex and Cas6 endonuclease suggest the evolution of distinct RNA binding mechanisms form a common platform, provide a basis for modelling the active site in the latter proteins, and, furthermore, indicate a coupling of specific RNA recognition to catalysis of RNA cleavage.

3-5. Materials and Methods.

3-5.1. Cloning, expression and purification of *T. thermophilus* CasE.

The *T. thermophilus* CasE gene TTHB192 was PCR amplified from genomic DNA (ATCC 27634D-5) using primers containing EcoRI and BamHI sites and cloned into pET-30a(+) (WT, Y23F) or pACYC-duet (R24A, R157A, R158A, H26A, H26N, H26Q, H26W) (Novagen). Mutagenesis was carried out by PCR and confirmed by sequencing. *E. coli* Rosetta cells were transformed with the plasmid, grown to an OD_{600} of ~0.8 and induced with 1 mM IPTG for 12 h at 24 ºC. Cells were lysed at 4 °C for 30 min (100 mM NaCl, 20 mM Tris-HCl pH 9.0, 1 mM 2-mercaptoethanol, 20 mM imidazole, 1 µg ml^{-1} lysozyme, 1 mM PMSF) followed by sonication. Lysate was cleared by centrifugation at 40,000 g for 30 min, bound to Ni Sepharose 6 Fast Flow resin (GE), and eluted with lysis buffer containing 200 mM imidazole. The resultant His_6-tagged CasE fusion proteins were heated to 55 ºC for 10 min (or 30 min as is the case for only the His26 mutants) and purified by Superdex 75 and anion exchange chromatography. Dialyzed protein was aliquoted and stored at -20 °C in a buffer containing 15 % (v/v) glycerol, 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 0.5 mM 2-mercaptoethanol, 0.5 mM EDTA. Protein purity was monitored at each step by resolution on an SDS-PAGE gel followed by coomassie staining.

3-5.2. RNA preparation.

RNAs purchased from IDT (Skokie, Il) were designed to model a full or partial CRISPR repeat: 5' GUA GUC CCC ACG CGU GUG GGG AUG GAC C 3' (28 nucleotides modeling the full repeat), 5' GUA GUC CCC ACG CGU GUG GGdG AUG GAC C 3' (28 nucleotide non-cleavable repeat), 5' GUC CCC ACG CGU GUG GGdG A 3' (22 nucleotide non-cleavable minimal repeat), 5' GUC CCC ACG CGU GUG GGG A 3' (22 nucleotide cleavable minimal repeat), and 5' GUC CCC ACG CGU GUG GGG 3' (21 nucleotide mimic of the 5' cleavage product).

3-5.3. Purification of CasE•RNA complexes.

Purified CasE protein and RNA oligonucleotides were pre-heated to 55 ºC for 5 min before being mixed at a 1:1.2 protein:RNA ratio in 100 mM NaCl, 10 mM Tris-HCl pH 9.0, 100 µM EDTA, 500 µM 2-mercaptoethanol and incubated for 30 min at 55 ºC. The reactions were purified on a Superdex 75 column (GE), and dialyzed into storage buffer containing 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 100 µM EDTA, 100 µM 2-mercaptoethanol.

3-5.4. Crystallization.

Crystals of CasE were grown at 23 °C by hanging drop vapor diffusion by mixing 1µl of 5 mg ml^{-1} protein solution with 1 µl of precipitant. Crystallization conditions were as follows, CasE crystals (50 mM Bicine pH 9.0, 15-18 % PEG

2000); CasE •product mimic crystals (100 mM Tris pH 8.0, 10-15 % (w/v) PEG 8000, 100 mM KCl, 5 mM $MgSO_4$, 1 mM spermidine); CasE •deoxyG crystals (100 mM HEPES pH 7.5, 10-15 % PEG 3350, 100 mM KOAc, 1 mM spermidine); CasE •product crystals (50 mM sodium succinate pH 6.0, 100 mM NaCl, 1 mM spermidine). Crystals were transferred to precipitant containing 20 % (v/v) glycerol and frozen in liquid nitrogen for data collection.

3-5.5. Data collection and processing.

Data were collected at 100 K at beamline 9-2 of the Stanford Synchrotron Radiation Lightsource (Palo Alto, CA) and beamline 12.3.1 at the Advanced Light Source (Berkeley, CA). Data were collected at a single wavelength for each crystal: 0.97946 Å for the CasE •product mimic RNA complex, 0.97945 Å for the CasE •dG RNA complex, and 1.11588 Å for the CasE •product RNA complex. Data were processed and scaled using the HKL package (Otwinowski & Minor, 1997).

3-5.6. Model building and refinement.

The structures were solved by molecular replacement using the program PHASER (Read, 2001). For the CasE structure and the CasE •product mimic RNA structure, the search model was PDB entry 1WJ9. The refined model of CasE •product mimic RNA was used as the search model for the CasE •deoxyG RNA and the CasE •product RNA structure. The initial molecular replacement model

was improved by iterative cycles of manual model building using COOT and refinement using REFMAC (Emsley & Cowtan, 2004; Murshudov *et al*., 1997). Final refined models occupied the following regions of a Ramachandran plot: CasE •product mimic RNA – 99.5 % favored, 0.5 % allowed, and 0 % outliers, CasE•dG RNA – 96.0% favored, 3.5 % allowed, and 0.5 % outliers, and CasE•product RNA – 97 % favored, 3.0 % allowed, and 0 % outliers. Ramachandran plots were generated using RAMPAGE (Lovell *et al*., 2002). All structural representations were created using PyMOL (Delano Scientific).

3-5.7. CD Spectroscopy.

CD spectra were collected at room temperature on a J-720 (Jasco) spectrophotometer. Spectra were collected from protein samples (0.2 mg/mL) in buffer containing 10 mM Tris HCl pH 8.0 and 100 mM NaCl in a 50 μm path-length cuvette. For each sample, the CD spectrum between 190 and 250 nm was measured 8 times and averaged.

3-5.8. Crystallographic data collection and refinement statistics

	3'OH RNA	dG RNA	Cut RNA	CasE
Data collection				
Space group	$P2_12_12_1$	I222	I222	$P2_12_12_1$
Cell dimensions				
a, *b*, *c* (Å)	37.06, 72.41, 100.23	61.50, 68.91, 158.06	62.17, 68.75, 155.84	43.13, 60.69, 100.28
α, β, γ (°)	90, 90, 90	90,90,90	90,90,90	90,90,90
Wavelength (Å)	0.97946	0.97945	1.11588	0.97949
Resolution (Å)	50-2.35	50-3.2	50-3.1	50-1.85
R_{sym} or R_{merge}	0.112 (0.502)	0.168 (0.412)	0.176 (0.499)	0.070 (0.409)
$I / \sigma I$	9.9 (2.0)	9.8 (3.8)	9.2 (2.7)	20.1 (3.2)
Completeness (%)	95.5 (75.1)	97.6 (89.4)	99.3 (99.8)	98.4 (92.9)
Redundancy	3.5	7.8	4.2	4.9
Refinement				
Resolution (Å)	50-2.35	50-3.2	50-3.1	50-1.85
No. reflections	11135	5452	5979	21725
R_{work} / R_{free}	0.185/0.227	0.223/0.266	0.243/0.297	0.180/0.219
No. atoms	2208	2027	2017	1809
Protein	1709	1623	1631	1628
Ligand/ion	345	402	386	0
Water	154	0	0	181
B-factors (Å^2)				
Protein	23.6	26.0	29.3	24.8
Ligand/ion	25.7	38.8	34.3	36.6
Water	27.8	na	na	na
R.m.s deviations				
Bond lengths (Å)	0.013	0.006	0.005	0.017
Bond angles (°)	1.59	1.16	0.72	1.55

Values in parentheses are for the highest-resolution shell (2.43-2.35Å for the 3'OH structure 3.31 to 3.20Å for the dG RNA structure and 3.21-3.10Å for the cut RNA structure).

3-6. References.

Brouns, S. J., Jore, M. M., Lundgren, M. & other authors (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**, 960-964.

Calvin, K., Xue, S., Ellis, C., Mitchell, M. & Li, H. (2008). Probing the catalytic triad of an archaeal RNA splicing endonuclease. *Biochemistry* **47**, 13659-13665.

Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**, 3489-3496.

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

Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**, 835-845.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & 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. & Cowtan, K (2004). Coot: model-building tools for molecular graphics. *Acta Crysta. D* **60,** 2126-2132.

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., & MacMillan, A.M. (2011). Recognition and Maturation of Effector RNAs in a CRISPR Interference Pathway. *NSMB*, Accepted.

Guillen Schlippe, Y. V. & Hedstrom, L. (2005). A twisted base? The role of arginine in enzyme-catalyzed proton abstractions. *Arch Biochem Biophys* **433**, 266-278.

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

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y. & Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **398**, 579-585.

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

Holm, L. & Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**, W545-549.

Jaeger, J.A., Turner, D.N., & Zuker, M. (1989). Improved predictions of secondary structures for RNA. PNAS 86, 7706-7710.

Jenney, F. E. & Adams, M. W. (2008). The impact of extremophiles on structural genomics (and vice versa). *Extremophiles* **12**. 39-50.

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

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

Lovell, S.C. *et al*. (2002). Structure validation by Cα geometry: φ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics* **50,** 437-450.

Marraffini, L. A. & Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181-190.

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

Otwinowski, Z., & Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* 276, 307–326.

Puglisis, J. D., Chen L., Blanchard S., & Frankel A. D. (1995). Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. Science 270, 1200-1203.

Read, R.J. (2001). Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Cryst. D* **57,** 1373-1382.

Schmidt, B.H., Burgin, A.B., Deweese, J.W., Osheroff, N., & Berger, J.M. (2010). A novel and unified two-metal mechanism for DNA cleavage by type II and IA topoisomerases. *Nature* 465, 641-644.

Steitz, T. A. & Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* **90**, 6498-6502.

Xue, S., Calvin, K. & Li, H. (2006). RNA recognition and cleavage by a splicing endonuclease. *Science* **312**, 906-910.

Wang, R., Preamplume, G., Terns, M. P., Terns, R. M. & Li, H. (2011). Interaction of the Cas6 Riboendonuclease with CRISPR RNAs: Recognition and Cleavage. *Structure* **19**, 257-264.

Chapter 4

Summary and Conclusions

4.1. The CRISPR-Cas System

The CRISPR-Cas system is a mechanism by which prokaryotes can acquire immunity to phage and plasmids. Over 90% of archaea and 40% of bacteria possess genetic loci characterized by palindromic repeat sequences interspaced by plasmid and phage derived spacers (Haft *et al*., 2005; Grissa *et al.*, 2007). The core Cas1 and Cas2 proteins are implicated in the acquisition and integration of new spacer sequences into the CRISPR loci (Brouns *et al*., 2008; Han *et al*., 2009; Makarova *et al*., 2006). These spacers are always incorporated at the leader end of the locus which creates a genetic recording of past phage and plasmid encounters (Mokrousov *et al.*, 2007; Zhang *et al.*, 2010). The locus is transcribed as a long pre-crRNA which is efficiently and specifically processed by a crRNA endonuclease (Brouns *et al.*, 2008; Carte *et al.*, 2008; Carte *et al.*, 2010; Haurwitz *et al.*, 2010). In the Ecoli CRISPR-Cas subtype, the Cascade multi-protein complex is responsible for this processing (Brouns et al., 2008). The enzymatic activity required for this cleavage resides in CasE, a metal-independent endoribonuclease (Brouns *et al*., 2008). These effector RNAs then act as guides to target homologous DNA for degradation which likely involves the activity of the Cas3 core protein (Horvath & Barrangou, 2010; Karginov & Hannon, 2010; Sorek *et al.*, 2008).

4.2. *T. thermophilus* CasE

In Chapter 2, the specific binding and cleavage of the CRISPR repeat sequence by the *T. thermophilus* CasE homologue after G21 was demonstrated. Following metal-independent cleavage, CasE remains bound to the 5' product. A minimal substrate for binding and cleavage consisting of the central 21 nucleotides from G4-G24 was defined by testing truncations of the CRISPR repeat RNA. Perturbations to the sequence but not secondary structure of the repeat RNA substrate significantly impair cleavage, suggesting that CasE recognizes its substrate in a sequence specific manner. Cleavage activity is impaired, while binding affinity is unaffected when presented with a substrate that has extended base-pairing at the base of the stem, suggesting that unwinding at the base of the stem may be necessary for cleavage.

In Chapter 3, analyses of three protein•RNA X-ray crystal structures were described detailing the mechanism of RNA recognition and substrate positioning for catalysis. Conserved basic residues in the RNA binding cleft stabilize the phosphodiester backbone of the 3' side of the strand through electrostatic interactions. CasE makes base-specific contacts to the first three unpaired nucleotides on the 5' end and to the major groove of the 3' side of the strand primarily with the β7-β8 hairpin. The importance of splaying at the base of the stem by the β10-β11 hairpin supported by several stabilizing contacts with the

unpaired nucleotides is consistent with the *in vitro* experiments of the RNA with the extended base-pairing at the base of the stem.

The combined active sites of these three structures establishes a mechanism in which the conserved Tyr23 stabilizes the transition state intermediate, Arg158 activates the nucleophile, and Arg27 stabilizes the 5' leaving group. The conserved His26 is not in a position to contribute to catalysis; rather it forms a hydrogen bond with a tryptophan at the core of the protein which may be important structurally.

4-3. Bimodal Domain Structure.

CasE can be described as having a bimodal structure consisting of a C-terminal RNA recognition domain and an N-terminal catalytic domain (Figure 4-1). These domains overlap with respect to the β10-β11 hairpin which is involved in both RNA recognition and catalysis. CasE specifically recognizes the crRNA repeat hairpin through major groove interactions with the β7-β8 hairpin. A conformational change occurs in which the base of the RNA stem-loop is separated and the β10-β11 hairpin stabilizes this conformation through a series of hydrogen bonds and π-stacking interactions. The scissile phosphate is positioned such that the 2' hydroxyl is primed for in-line displacement. The β10-β11 hairpin provides Arg158 which activates the nucleophile. The conserved Tyr23 residue stabilizes the transition state/intermediate and Arg27 is in position to protonate the

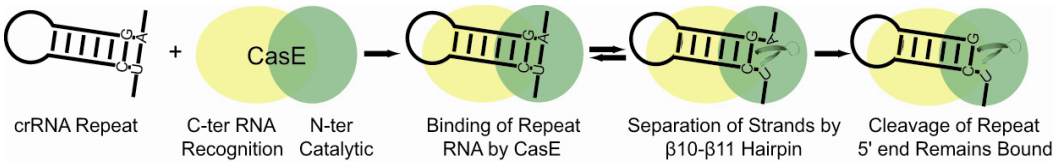

Figure 4-1. Model of CasE Binding and Cleavage. CasE is composed of a bimodal structure consisting of a C-terminal RNA recognition domain and an N-terminal Catalytic domain which overlap with respect to the β10-β11 hairpin. CasE binds to its crRNA repeat, making sequence specific contacts with the dsRNA helix. The splaying of the strands by the insertion of the β10-β11 hairpin places the scissile phosphate in the catalytically optimal position. This conformational change is presumably in dynamic equilibrium until CasE cleaves the RNA and remains bound to the 5' product of cleavage.

leaving group. The RNA is cleaved and CasE remains bound to the 5' product. The CasE•Product complex may then somehow be involved in downstream targeting events.

4.4. Implications for Functional Homologues

Comparison with the two other structurally characterized crRNA endonucleases reveals that a common domain structure exists between these seemingly unrelated proteins (Carte *et al.*, 2008; Haurwitz *et al.*, 2010). They share an N-terminal catalytic domain characterized by a ferredoxin-like fold which contributes the residues important for catalysis. In CasE, Csy4, and likely Cas6, this domain also contains a β hairpin which appears to be important for separating the base of RNA stem-loops and correctly positioning the scissile phosphate in the active site (Carte *et al.*, 2008; Haurwitz *et al.*, 2010). The

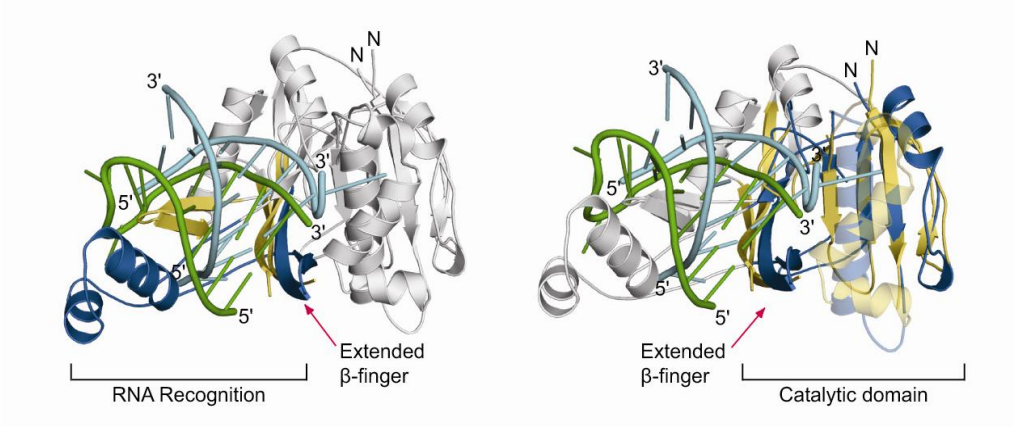

Figure 4-2. Modular organization of pre-crRNA recognition and processing revealed by comparison of CasE and Csy4 RNA complexes. Superposition of cognate *P. aeruginosa* Csy4•RNA (pdb 2XLK) and *T. thermophilus* CasE•dG RNA X-ray structures highlighting (left) RNA recognition by distinct C-terminal modules that specifically recognize the RNA major groove as well as by a common extended β-finger interacting with the base of the duplex proximal to the cleavage site and (right) the N-terminal catalytic domains. Csy4 and CasE are coloured blue and yellow respectively. The Csy4 RNA is coloured green and the CasE RNA is coloured cyan. The β-finger contributes to both RNA binding and the architecture of the active site.

C-terminal RNA recognition domain makes sequence specific interactions with the major groove. In CasE, this RNA recognition is carried out by an arginine rich β hairpin. In Csy4, an arginine rich α-helical domain and linker region carry out this role (Carte *et al.*, 2008; Haurwitz *et al.*, 2010)

The active sites of these three proteins vary considerably, suggesting that the mechanism of catalysis is not conserved between subtypes. In *T. thermophilus* CasE, two arginines and a tyrosine residue make up the active site. In Csy4, a histidine, a serine and likely the α-amino group of a glutamine are involved (Haurwitz *et al.*, 2010). Cas6's active site is comprised of a tyrosine, a

histidine and a leucine, very much like the tRNA splicing endonuclease (Calvin *et al*., 2008; Carte *et al.*, 2008). This disparity amongst crRNA endonuclease active sites suggests that these proteins may have co-evolved to carry out a similar function as most of the similarities exist between their tertiary structures and not their primary sequence.

All three proteins, have a bimodal structural organization consisting of an N-terminal ferredoxin-like folds including elements of the endonuclease active site with distinct C-terminal RNA binding modules (Carte *et al.*, 2008; Haurwitz *et al.*, 2010). The ordering of the extended β-hairpin in the CasE and Csy4 RNA complexes suggests that that the final assembly of the enzyme active site is a function of specific RNA binding and thus acts as a regulatory mechanism (Haurwitz *et al.*, 2010).

Our comparison of three diverse CRISPR endonucleases alone and complexed to RNA show that while these proteins have evolved both distinct RNA recognition modes and catalytic mechanisms, they nevertheless share key common features governing both these steps of pre-crRNA processing. The documented evolutionary relationship between CRISPR sub-types suggest that the features described here will be broadly generalizable to a large family of processing endonucleases.

4-5. Broad Implications.

Exploring the mechanism behind the CRISPR-Cas system is essential to the development of this bacterial genetic interference system as a tool in molecular biology and as a basis for new medical therapies and industrial applications. With delineation of the biochemical underpinnings of this pathway and understanding of the similarities and differences between subtypes, this pathway may be exploited both in the laboratory and in the clinic much like the eukaryotic RNA interference pathway. The CRISPR-Cas system may be used to target endogenous bacterial and archaeal genes to understand their microbiological function. Understanding this pathway may also lead to the development of new phage therapy protocols in the treatment of antibiotic resistant infections and to the improvement of phage resistant bacterial starter cultures in industries that use the fermentation process such as the dairy industry. The development of new genetic and medical tools is a probable outcome resulting from our increased knowledge of the mechanism behind this pathway.

4-6. Future Directions.

An important goal of future studies will be to determine whether crRNA processing is coupled to downstream targeting events. Because the endonucleases remain bound to the product following cleavage it is possible that they may play a role in targeted DNA degradation (Brouns *et al.*, 2008; Carte *et al.*, 2008; Haurwitz *et al.*, 2010). Although Cas3 has been implicated as the ultimate

effector protein, it is not known which other proteins may be involved in this process (Brouns *et al*., 2008). In *E. coli*, CasE is part of the multi-protein Cascade complex consisting of CasA-E (Brouns *et al*., 2008). Perez-Rodriguez and colleagues also describe a ternary complex in E. coli consisting of CasC, D, and E (Perez-Rodriguez *et al*., 2011). Establishing a role for members of these complexes is crucial to our understanding of its function in the pathway. Substantiation of the speculation that Cascade is somehow coupled to downstream targeting events is imperative. Further characterization of these proteins is necessary to elucidate the role of the Cascade complex.

The role of the invariant histidine as a structurally important amino acid must be confirmed through further studies. A thorough CD spectrographic analysis of the histidine point mutants under increasing denaturing conditions may provide some insight. The use of differential scanning fluorimetry which makes use of a dye that fluoresces when it binds to exposed hydrophobic patches during denaturation is another method that could help answer this question. It would be beneficial to explore the effect of lower pH on the histidine point mutants as the wild-type enzyme becomes inactive at pHs lower than 6.0, possibly due to loss of the His26/Trp46 hydrogen bonding interaction important for protein stability. Ultimately, the solution of X-ray crystal structures of these mutants in the presence and/or absence of RNA could confirm or disprove our hypothesis.

To further establish the bimodal crRNA endonuclease organization, it would be useful to perform further structural and functional studies with other Cas sub-types. This may validate the endoribonuclease domain model of a C-terminal RNA recognition domain and an N-terminal catalytic domain. For example, characterization of the Dvulg sub-type crRNA endonuclease, which also is associated with a stable stem-loop forming CRISPR repeat would make an excellent candidate for future structural and functional studies (Haft *et al*., 2005).

4-7. References

Brouns, S. J., Jore, M. M., Lundgren, M. & other authors (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**, 960-964.

Calvin, K., Xue, S., Ellis, C., Mitchell, M. & Li, H. (2008). Probing the catalytic triad of an archaeal RNA splicing endonuclease. *Biochemistry* **47**, 13659-13665.

Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**, 3489-3496.

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

Grissa, I., Vergnaud, G. & Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

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

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

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

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

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

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

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

Xue, S., Calvin, K. & Li, H. (2006). RNA recognition and cleavage by a splicing endonuclease. *Science* **312**, 906-910.

Appendix *I*

Characterization of Cascade components A-E.

I-1. Cascade complex

The essential processing of the CRISPR transcript into effector crRNAs in the prokaryotic CRISPR-Cas system is carried out by crRNA endoribonucleases. The known endonucleases *P. aeruginosa* Csy4, *P. furiosus* Cas6, *E. coli* CasE, and *T. thermophilus* CasE are all RAMP domain containing proteins which remain bound to their product following cleavage suggesting that these endonucleases may somehow be coupled to downstream DNA targeting (Carte *et al.*, 2008; Carte *et al.*, 2010; Haurwitz *et al.*, 2010).

In *E. coli* K12, the enzymatic activity of the Cascade complex (CRISPR-associated complex for antiviral defence) consisting of CasA-E which was purified by co-immunoprecipitation, resides in CasE (Figure I-1a; Brouns *et al.*, 2008). Recent work has shown that the Cascade complex, when reconstituted from recombinant proteins, is made up primarily of a core ternary complex consisting of CasC, D, and E (Figure I-1b; Perez-Rodriguez *et al.*, 2011). This stable complex purified at a 6:1:1 CasC:CasD:CasE ratio. Unlike the Van der Oost laboratory's Cascade complex, which contained CasA and CasB at sub-stoichiometric levels, the Delisa laboratory's ternary complex did not stably associate with CasA or CasB (Brouns *et al.*, 2008; Perez-Rodriguez *et al.*, 2011). This suggests that CasA and B may be loosely associated accessory proteins, or alternatively, that CasA and B may only associate with CasC, D and E in an RNA dependent manner, which was not tested (Perez-Rodriguez *et al.*, 2011). CasC

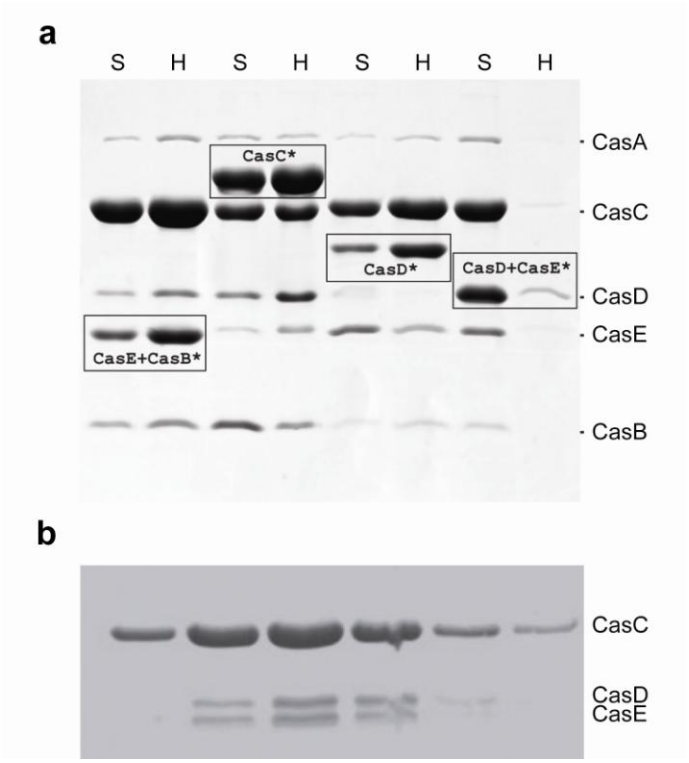

Figure I-1. Cascade complex and Ternary Complex. (a) CasB-E with N-terminal StrepII and C-terminal His affinity tags can co-immunoprecipitate CasA-E (taken from Brouns *et al.*, 2008). (b) A stable ternary complex comprised of CasC, D, and E in a 6:1:1 ratio has been purified by size exclusion chromatography (taken from Perez-Rodriguez *et al.*, 2011).

tends to oligomerize into a hexamer, while CasD is not soluble in the absence of CasC. It is likely that CasC makes up the scaffold of this complex as CasD and CasE did not interact *in vitro* (Perez-Rodriguez *et al.*, 2011). Perez-Rodriguez and colleagues have also shown that a homologous ternary complex exists in *T. thermophilus* HB8. Additionally, they found that the *T. thermophilus* CasE protein can be incorporated into the *E. coli* ternary complex (Perez-Rodriguez *et al.*, 2011). This new work is very informative as it would allow for mixing and

matching of Cas components amongst species which may be useful experimentally.

CasE has been defined as the crRNA endoribonuclease, however the role of the remaining Cascade components remains unclear (Brouns *et al.*, 2008; Pul *et al.*, 2010a; Pul *et al.*, 2010b; Westra *et al.*, 2010). They may be involved in targeting the complex to areas of active CRISPR transcription or possibly in coupling crRNA processing to downstream targeting events. CasE remains bound to its product following cleavage which has also been referred to as the 3' handle (Figure 1-2). It is possible that a Cascade protein binds to the 5' handle of a mature crRNA following cleavage (Figure 1-2). Eric Sontheimer's group demonstrated that in *S. epidermidis*, the upstream flanking regions of the proto-spacer require complementarity for self versus non-self discrimination (Marraffini & Sontheimer, 2010). If a protein were identified which bound to the 5' handle of the mature crRNA, a possible role would be to check for DNA•RNA hybridization with proto-spacers to regulate targeting (Marraffini & Sontheimer, 2010).

I.2. Cloning of *T. thermophilus* Cascade Components.

To characterize the role of the Cascade components in crRNA processing, we cloned the Ecoli subtype proteins from *T. thermophilus* HB8 as N-terminal His-tag fusion proteins and expressed them in *E. coli* cells which do not contain a

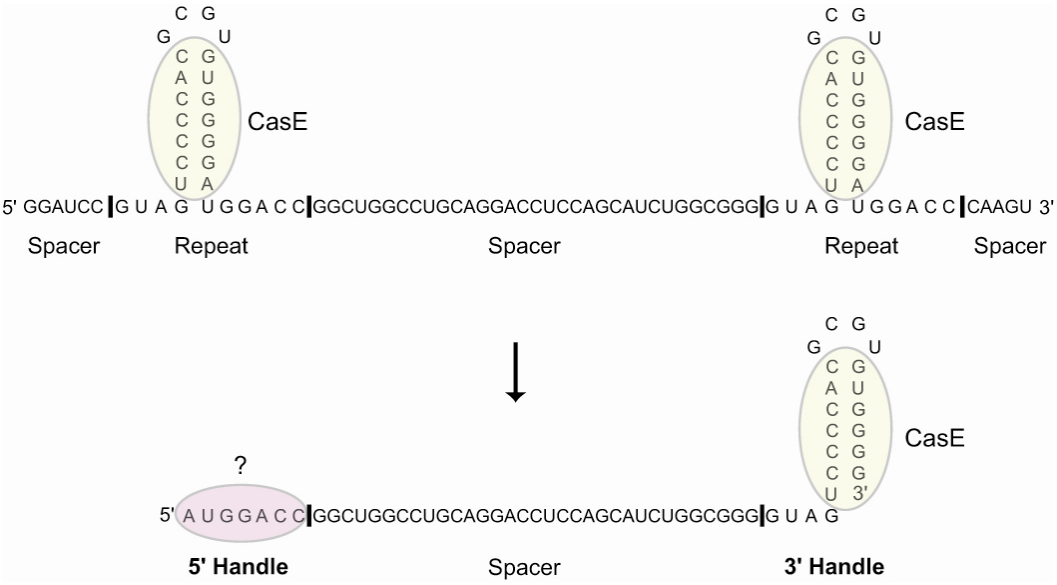

Figure I-2. Possible interactions with "5' handle". CasE cleaves the CRISPR repeat RNA and remains bound to the "3' handle" of the RNA (yellow). It is possible that the role of another member of the Cascade complex is to bind to the "5' handle" (pink) (Brouns et al., 2008; Maraffinni & Sontheimer, 2010).

functional CRISPR system (Rosetta). Although work was primarily focused on CasE, preliminary data were collected for CasA-D which is described in this appendix.

I-2.1 **CasA**.

Expression of the 55 kDa *T. thermophilus* CasA with a 5 kDa N-terminal His-tag produced a mostly monomeric protein eluting at a volume of 162 ml from

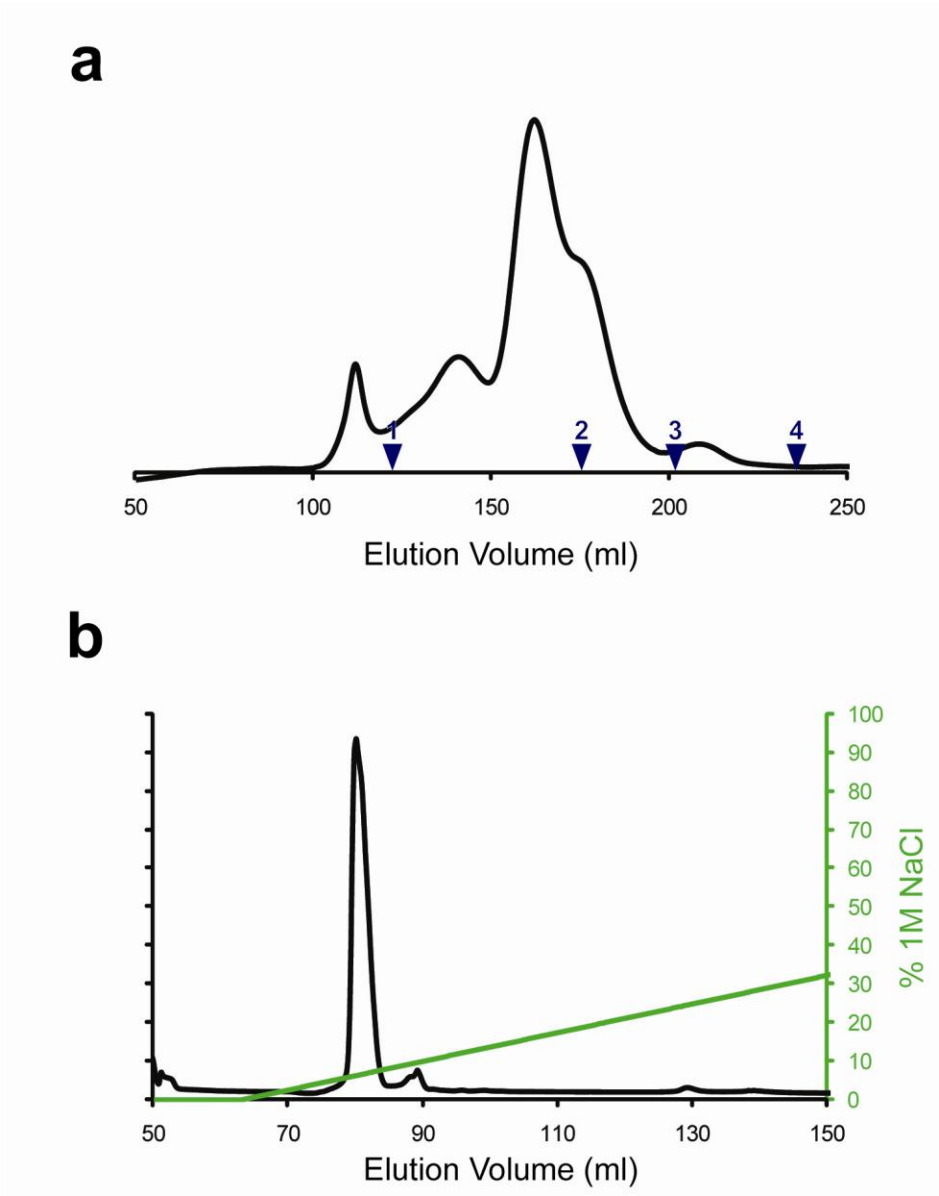

Figure I-3. Purification of *T. thermophilus* CasA. (a) The elution profile of *T. thermophilus* CasA from a Superdex 200 size exclusion column. CasA eluted at ~162 ml, consistent with a 55 kDa CasA monomer containing a 5 kDa N-terminal affinity tag. The elution peaks of molecular weight standards are also indicated (purple triangles) and numbered 1-4 and are as follows: 1. Catalase, 250 kDa, 121 ml; 2. Aldolase, 158 kDa, 173 ml; 3. Bovine serum albumin, 67 kDa, 200 ml; 4. Chymotrypsin, 25 kDa, 240 ml. (b) The elution curve of *T. thermophilus* CasA from a Q-Sepharose anion exchange column. CasA eluted at a volume corresponding to ~80 mM NaCl. mAu are shown in black and % 1M NaCl is shown in green.

the Superdex 200 size exclusion column along with several contaminating peaks (Figure I-3a). The monomeric CasA fractions were collected, concentrated, and subjected to Q-Sepharose anion exchange chromatography where it eluted at a volume corresponding to 80 mM NaCl (Figure I-3b). These fractions were collected, concentrated, and dialyzed overnight into a buffer containing 100 mM NaCl, 10 mM Tris-HCl pH 9.0, 0.5 mM 2-mercaptoethanol, 0.5 mM EDTA, and 15% glycerol. This storage buffer was used for all of the Cascade protein components which were aliquoted into 100 µl volumes and stored at -20 °C. CasA did not bind to the CRISPR repeat RNA, or unrelated, single-stranded, double-stranded, and hairpin RNAs based on gel electrophoretic mobility shift assays (data not shown).

I-2.2 CasB.

CasB was expressed as an 18 kDa fusion protein with a 5 kDa N-terminal His-tag and eluted as a monomer by size exclusion chromatography at a volume of ~243 ml on the Superdex 200 column (Figure I-4). The X-ray crystal structure of the *T. thermophilus* homologue has been published (PDB entry 2ZCA; Agari *et al*., 2010) which reveals a mostly α-helical fold with a conserved basic patch that may be involved in nucleic acid binding discussed in Chapter 1 (Figure 1-12). CasB did not bind to either the anion or cation exchange columns. CasB did not bind to unrelated, single-stranded, double-stranded, and hairpin RNAs based on

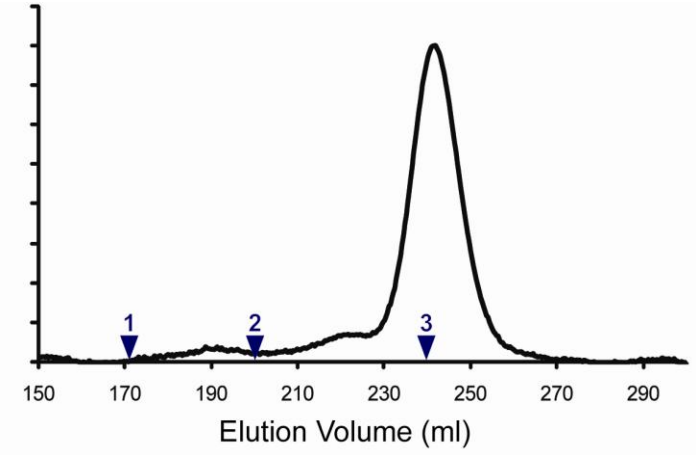

Figure I-4. Purification of *T. thermophilus* CasB. The elution profile of *T. thermophilus* CasB from a Superdex 200 size exclusion column. CasB eluted at ~243 ml, consistent with the 18 kDa CasB monomer with a 5 kDa N-terminal affinity tag. The elution peaks of molecular weight standards are also indicated (purple triangles) and numbered 1-3 and are as follows: 1. Aldolase, 158 kDa, 173 ml; 2. Bovine serum albumin, 67 kDa, 200 ml; 3. Chymotrypsin, 25 kDa, 240 ml. mAu are shown in black.

gel electrophoretic mobility shift assays (data not shown). It displayed a very limited affinity for the CRISPR repeat RNA (> 100 μM) (Figure I-8).

I-2.3. CasC.

CasC, a 40 kDa protein, was cloned and expressed as a fusion protein with a 5 kDa N-terminal His-tag resulting in a predicted 45 kDa protein. This protein eluted at a volume corresponding to a larger than expected multimer on the Superdex 200 size exclusion column, ~130 ml (Figure I-5a). This peak may represent the 270 kDa hexamer which was observed by the Delisa group, however the elution volume corresponding to protein size standards suggests that this multimer is actually ~240 kDa (Perez-Rodriguez *et al.*, 2011). We used this

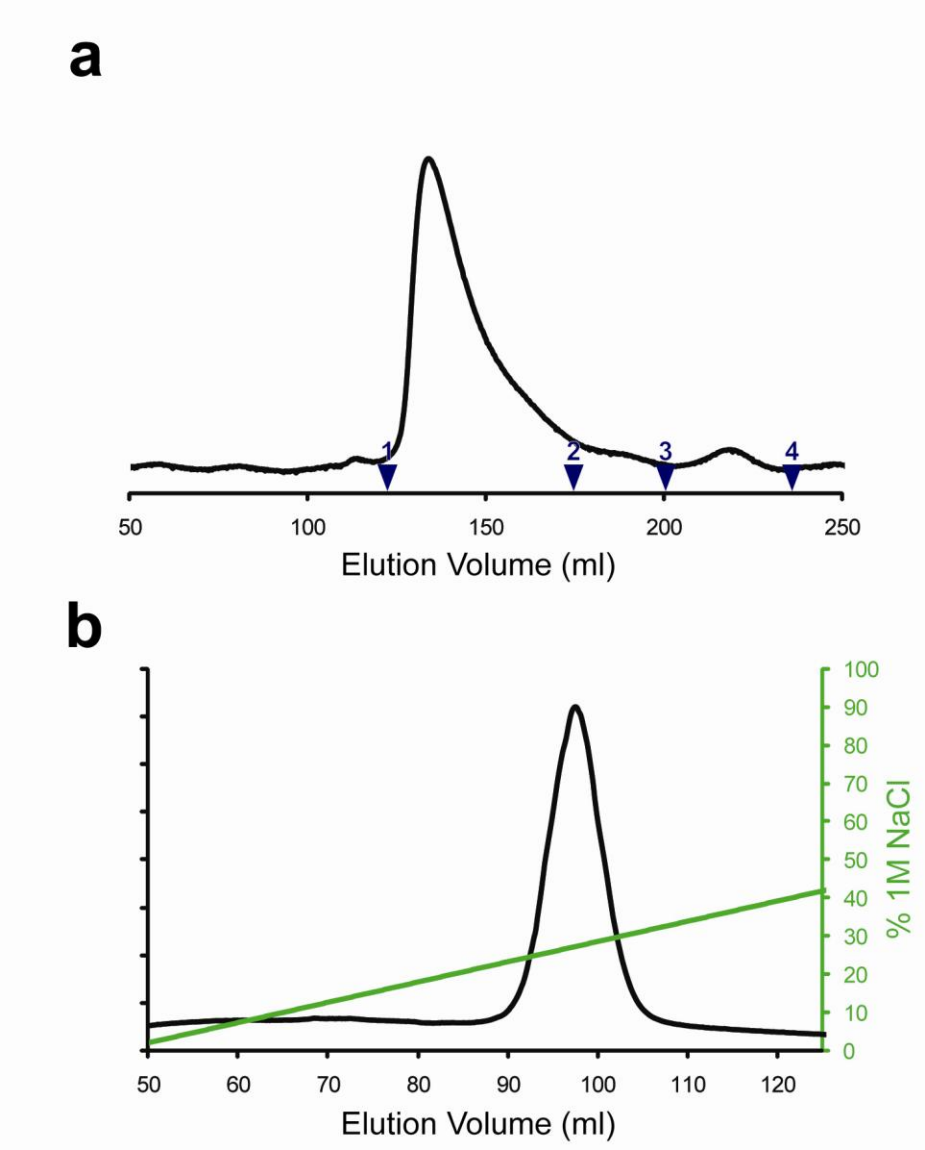

Figure I-5. Purification of *T. thermophilus* CasC. (a) The elution profile of *T. thermophilus* CasC from a Superdex 200 size exclusion column. CasC eluted at ~130 ml, which is inconsistent with a 270 kDa CasC hexamer and also inconsistent with the 45 kDa monomer. The elution peaks of molecular weight standards are also indicated (purple triangles) and numbered 1-4 and are as follows: 1. Catalase, 250 kDa, 121 ml; 2. Aldolase, 158 kDa, 173 ml; 3. Bovine serum albumin, 67 kDa, 200 ml; 4. Chymotrypsin, 25 kDa, 240 ml. (b) The elution profile of *T. thermophilus* CasC from a Q-Sepharose anion exchange column. CasC eluted at a volume corresponding to ~275 mM NaCl. mAu are shown in black and % 1M NaCl is shown in green.

multimeric peak from our preparations to perform our experiments. This CasC multimer eluted from the Q-Sepharose anion exchange column at a volume corresponding to ~275 mM NaCl (Figure I-5b). In gel electrophoretic mobility shift assays, CasC binds to the *T. thermophilus* cognate repeat RNA with a 30 nM affinity, but does not bind to unrelated single-stranded, double-stranded, or hairpin RNAs (Figure I-8, data not shown). Like CasE, CasC binds the 5' cleavage product with the same affinity as the full length repeat and does not bind to the 3' piece (data not shown). CasC's affinity for the repeat was ~30 times weaker than that of CasE's (~1 nM). CasC did not possess any endonucleolytic cleavage activity, suggesting that its interaction with the repeat may be involved in assisting CasE's processing of the long CRISPR transcript, or in downstream targeting events. Because CasC is likely a hexamer in the ternary complex, it may be a scaffolding protein which interacts with the long CRISPR transcript (Perez-Rodriguez *et al.*, 2011).

I-2.4. CasD.

CasD was as expressed as a 34 kDa fusion protein containing a 5 KDa His-tag. This protein eluted at a volume corresponding to a monomeric protein (~194 ml) by size exclusion chromatography so long as the 55 °C incubation step following affinity tag purification was carried out faithfully during the preparation (Figure I-6a). Intriguingly, if the purification was performed without heating, size exclusion chromatography yielded a combination of aggregate, a potential dimer

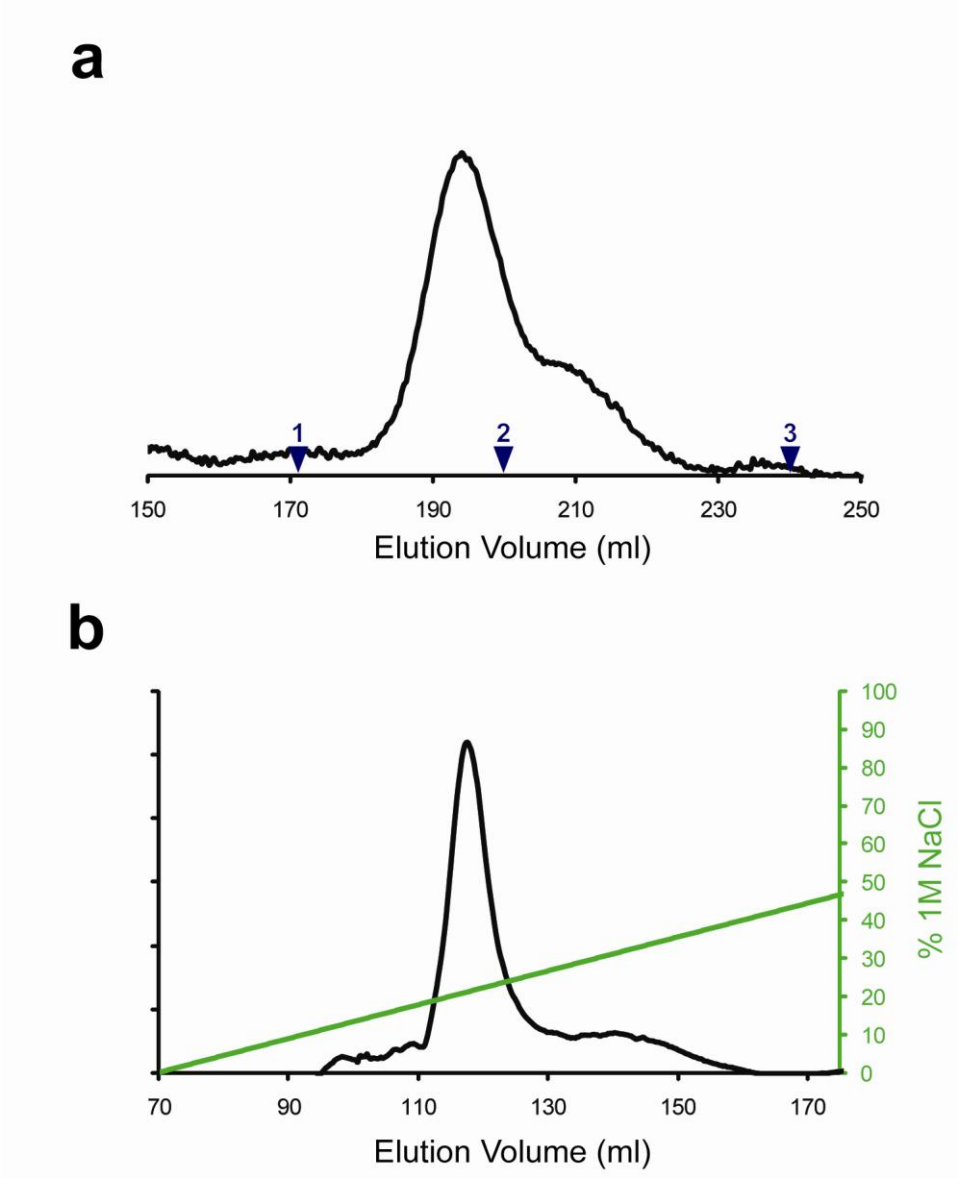

Figure I-6. Purification of *T. thermophilus* CasD. (a) The elution profile of *T. thermophilus* CasD from a Superdex 200 size exclusion column. CasD eluted at a volume of ~194 ml, consistent with a 34 kDa CasD monomer. The elution peaks of molecular weight standards are also indicated (purple triangles) and numbered 1-3 and are as follows: 1. Aldolase, 158 kDa, 173 ml; 2. Bovine serum albumin, 67 kDa, 200 ml; 3. Chymotrypsin, 25 kDa, 240 ml. (b) The elution profile of *T. thermophilus* CasD from a Q-Sepharose anion exchange chromatography column. CasD eluted at a volume corresponding to ~213 mM NaCl. mAu are shown in black and % 1M NaCl is shown in green.

peak, and a monomer peak. If we isolated the aggregate or dimer peak and heated it extensively at 55 °C, size exclusion chromatography yielded once more the monomeric peak (data not shown). This suggests that this protein tends to aggregate over time and is unstable at lower temperatures consistent with Perez-Rodriguez's observation that CasD is unstable in the absence of CasC. We used the monomeric protein peak to perform our experiments. This protein eluted from the S-Sepharose cation exchange column at a volume corresponding to 213 mM NaCl (Figure I-6b). In gel electrophoretic mobility shift assays, CasC bound to the *T. thermophilus* cognate repeat RNA at an estimated 90 nM affinity, but not unrelated dsRNA, ssRNA or RNA hairpins (Figure I-8). Like CasE, CasD bound to the 5' cleavage product with the same affinity as the full length repeat and did not bind the 3' piece. However, CasD's affinity for the repeat was ~90 times weaker than CasE's (~1 nM). Like CasC, CasD did not cleave the cognate CRISPR repeat. Like CasC, CasD did not possess any endoribonucleolytic activity. Despite CasD's observed affinity for the repeat RNA, its role in crRNA processing and targeting remains unclear.

I-2.5. CasE

T. thermophilus CasE is a ~23 kDa protein which was expressed as a 28 kDa fusion protein containing an N-terminal His-tag. CasE eluted at a volume corresponding to a monomer from the Superdex 75 size exclusion column at 178 ml (Figure I-7a). CasE eluted at a NaCl concentration of 307 mM by cation

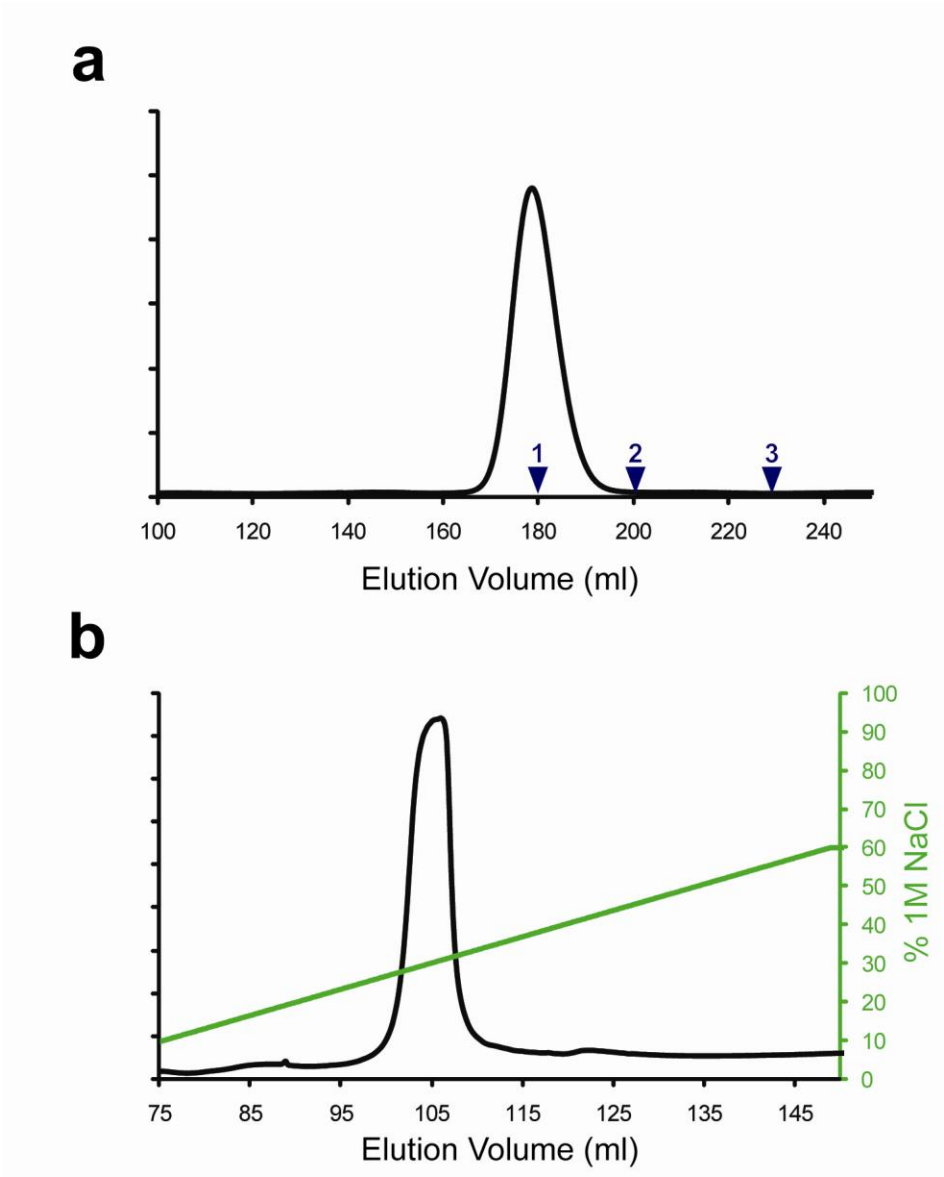

Figure I-7. Purification of *T. thermophilus* CasE. (a) The elution profile of *T. thermophilus* CasE from a Superdex 75 size exclusion column. CasE eluted at ~178 ml, consistent with a 28 kDa CasE monomer. The elution peaks of molecular weight standards are also indicated (purple triangles) and numbered 1-3 and are as follows: 1. His6-TEV protease, 34 kDa, 180 ml; 2. p14•SF3b complex, 19.5 kDa, 200 ml; 3. U1A binding protein, 11 kDa, 227 ml. (b) The elution curve of *T. thermophilus* CasE from a S-Sepharose cation exchange column. CasE eluted at ~ 307 mM NaCl. mAu are shown in black and % 1M NaCl is shown in green.

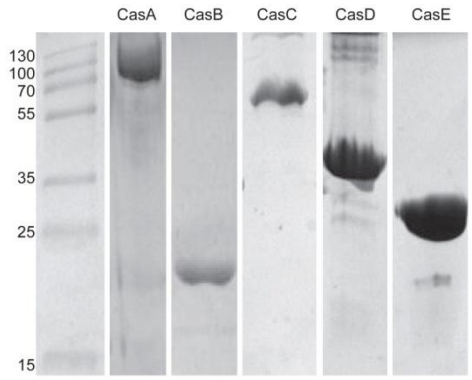

Figure I-8. SDS-PAGE analysis of *T. thermophilus* Cascade components. Shown here is purified N-terminal His-tagged recombinant *T. thermophilus* HB8 CRISPR-associated proteins CasA-E.

exchange chromatography on the S-Sepharose column (Figure I-7b). This protein bound to the *T. thermophilus* crRNA repeat with an ~ 1 nM affinity. The characterization of this protein was described in detail in Chapters 1 and 2.

I-3. A Hierarchy of crRNA Repeat Binding.

It was intriguing to find that proteins other than CasE bound to the crRNA repeat sequence. We originally hypothesized that the role of one of the Cascade components may be to bind to the 3' end of a mature crRNA (Figure I-1). However, when each of the Cascade components were tested for an affinity to this sequence, none bound to the oligonucleotide (Figure I-9). Instead, we found that CasC and CasD both bound to the full 28 nt repeat sequence with an equal affinity for the 21 nt 5' product (Figure I-9). As well, a hierarchy of repeat binding was established where CasE binds the tightest followed by CasC and then CasD. The implications for this observation are unclear, however it is interesting that these

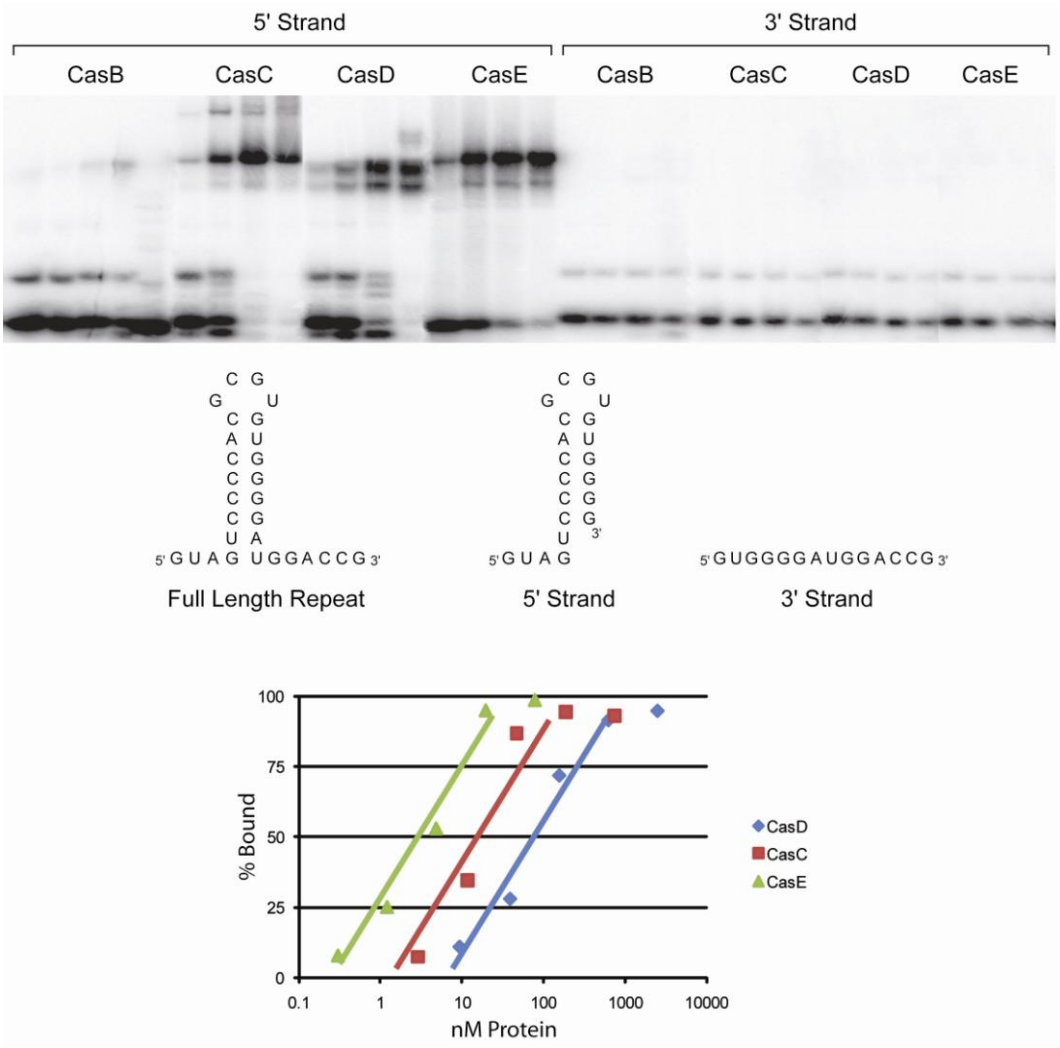

Figure I-9. Electrophoretic Mobility Shift Assays of CasB, C, D, and E. CasB incubated with the 5'-radiolabeled 21 nt product mimic RNA and the 12 nt RNA corresponding to the 3' end of the repeat at 100, 10, 1, 0.1 μM. CasB has a very low affinity for this RNA. CasC incubated with these same RNAs at 1000, 100, 10, 1 nM. CasD incubated with these same RNAs at 10, 1, 0.1, 0.01 μM. CasE incubated with these same RNAs at 100, 10, 1, 0.1 nM.

same three proteins make up the ternary complex. Perhaps CasC and CasD somehow bind to the length of the CRISPR transcript and pass a single repeat off to CasE so it can be cleaved.

I-4. Reconstitution of Ternary Cascade Complex.

We attempted to reconstitute the Cascade complex in the absence of CasA by incubating the individually purified components together at 55 °C for thirty minutes before separation by size exclusion chromatography. We observed the elution of what might have been a ternary complex from the Superdex 200 column containing CasC, D, E at a volume of 140 ml, which is estimated to be a size of ~210 kDa compared to molecular weight standards (Figure I-10). The 6:1:1 CasC:CasD:CasE complex found by Perez-Rodriguez and colleagues would have an expected molecular weight of ~330 kDa (Perez-Rodriguez *et al.*, 2011). Therefore it is unclear what the stoichiometry might be between these three proteins as it was a very crude experiment. CasB did not co-eluted with the three other proteins and instead eluted at ~240 ml as it did when purified alone. These experiments were not carried any further, but it would be a reasonable goal to try to reconstitute the Delisa laboratory's experiments.

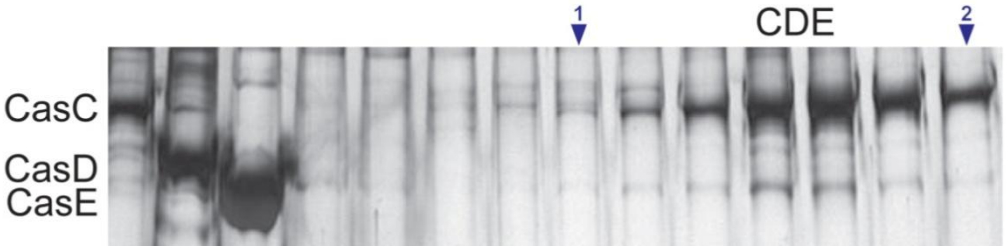

Figure I-10. Cascade Ternary Complex. SDS-PAGE gel of fractions from Superdex 200 size exclusion chromatography of Cascade components CasB-E incubated together at 55 °C for 30 minutes. Between fractions 16-18 a ternary complex consisting of CasC,D, and E eluted at ~140 ml. The elution volume of molecular weight standards are indicated (purple triangles) and numbered 1-2 and are as follows: 1. Catalase, 250 kDa, 121 ml; 2. Aldolase, 158 kDa, 173 ml.

I-5. Discussion and Future Directions.

An interesting finding was the affinity of CasC and CasD for the cognate repeat RNA. There appeared to be a hierarchy of affinities from these three Cascade components, with CasE being the tightest binder followed by CasC then CasD. This is interesting in light of the new ternary complex identified by Perez-Rodriguez and colleagues consisting of these three components. Perhaps all three proteins contribute somehow to the repeat binding. It is unclear what this could mean for downstream events.

I-6. Materials and Methods.

I-6.1. Cloning, Expression, and Purification of Cascade Components.

Full length CasA, B, C, D, and E were PCR amplified from *T. thermophilus* genomic DNA (ATCC 27634D-5) using oligonucleotide primers containing EcoRI and BamHI restriction sites which insert a TEV protease cleavage site between the His-tag and the Cas gene and cloned into the pET-30a(+) vector (Novagen). Expression plasmids were transformed into *E. coli* Rosetta cells which were grown to an OD_{600} of ~0.8 before induction of protein expression with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 12 h at 24 ºC. Cells were pelleted, frozen and lysed at 4 °C for 30 min (100 mM NaCl, 20 mM Tris-HCl pH 9.0, 1 mM 2-mercaptoethanol, 20 mM imidazole, 1 μg/ml lysozyme, 1 mM PMSF) followed by sonication. The lysate was cleared by centrifugation at 15,000 RMP for 30 min. Cleared lysate was bound to a Ni Sepharose 6 Fast Flow column (GE) and eluted with a buffer containing 200 mM imidazole. Eluted protein was concentrated and heated to 55 °C for 30 min and centrifuged at 15,000 RPM for 10 min to remove contaminating proteins. The resultant His_6-tagged fusion proteins were purified by Superdex 75 or Superdex 200 and anion or cation exchange chromatography. Protein was dialyzed overnight into a buffer containing 100 mM NaCl, 10 mM Tris-HCl pH 9.0, 0.5 mM 2-mercaptoethanol, 0.5 mM EDTA. Protein used for activity/binding assays was aliquoted and stored at -20 °C in a buffer containing 15 % glycerol.

I-6.2. RNA preparation.

RNAs were designed to model a full or partial CRISPR repeat and purchased from Integrated DNA Technologies (Skokie, Il) and used without further purification. Details of oligonucleotide sequences can be found in section I-6.5.

I-6.3. Gel mobility shift and cleavage assays.

5'-^{32}P-radiolabeled synthetic RNA substrates (50 - 100 × 10^3 cpm) were pre-heated to 55 ºC before incubation in 10 µl reactions containing 0-100 µM Cas protein (100 mM NaCl, 10 mM Tris-HCl pH 9.0, 100 µM EDTA, 500 µM 2-mercaptoethanol and 20 µg/µl yeast tRNA). Reactions were incubated for 30 min at 55 ºC before being immediately loaded onto a 6 % tris-glycine polyacrylamide gel and run at 150 volts for 1.5 hours. Dried gels were exposed to a phosphor screen (Molecular Dynamics) and scanned with a Storm 840 Phosphorimager (Molecular Dynamics).

I-6.4. Reconstitution of Cascade ternary complex.

Purified *T. thermophilus* CasB, C, D, and E proteins were pre-heated to 55 ºC before being mixed in equal parts with a final buffer concentration of 100 mM NaCl, 10 mM Tris-HCl pH 9.0, 100 µM EDTA, 500 µM 2-mercaptoethanol and incubated for 30 minutes at 55 ºC to form the complex. The complex reaction was subjected to Superdex 200 size exclusion chromatography. Samples were

taken from every two fractions and added to an equal volume of SDS loading dye before being resolved by SDS-PAGE gel. Gels were stained with Coomasie dye.

I-6.5. Oligonucleotides Used in this Work.

Oligonucleotide	5' → 3'
Cognate crRNA	GUA GUC CCC ACG CGU GUG GGG AUG GAC C
5' Strand	GUA GUC CCC ACG CGU GUG GGG
3' Strand	UGG GGA UGG ACC
Hairpin RNA (R/G)	GGG UCC UCA UUA AGG UGG GUG GGA AUA GUA UAA CAA UAU GCU CAA UGU UGU UAU AGU AUC CCA CCU ACC CUG AUG UGU C
ssRNA (miR16)	UAG CAG CAC GUA AAU AUU GGC

I-12. References.

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

Brouns, S. J., Jore, M. M., Lundgren, M. & other authors (2008). Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**, 960-964.

Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**, 3489-3496.

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

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

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

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K. H., Bernardis, S., Ke, A. & Delisa, M. P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. *Mol Microbiol* **79**, 584-599.

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

Westra, E., Pul, U., Heidrich, N. & other authors (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Mol Microbiol*. 75, 1513-1522.