CHARACTERIZATION OF CRISPR RNA BIOGENESIS PATHWAYS IN *STREPTOCOCCUS THERMOPHILUS*

by

JASON MICHAEL CARTE

(Under the Direction of Michael P. Terns)

ABSTRACT

The CRISPR-Cas (CC) immune systems protect many prokaryotes from invasive genetic elements. These immune systems are composed of small CRISPR (cr)RNAs which guide effector Cas (CRISPR-associated) proteins to silence foreign DNA or RNA. crRNAs are derived from CRISPR arrays which are characterized by a series of short direct repeats separated by similarly sized invader-derived sequence elements known as spacers. Processing of crRNA precursors to the mature forms is a critical step in CC-mediated immunity. We have investigated the crRNA biogenesis pathways in the lactic acid bacterium *Streptococcus thermophilus*. *S. thermophilus* contains four CC systems. Each CC module is composed of a distinct CRISPR array and an adjacently encoded cluster of *cas* genes. We found that each CRISPR array gives rise to distinct crRNAs and that biogenesis of crRNAs from each CC system in *S. thermophilus* is mediated by Cas protein factors associated with that locus, independently of the other three CC modules.

INDEX WORDS: CRISPR, Cas, crRNA, RNA, RNA processing, viral defense, *Streptococcus thermophilus*

CHARACTERIZATION OF CRISPR RNA BIOGENESIS PATHWAYS IN *STREPTOCOCCUS THERMOPHILUS*

by

JASON MICHAEL CARTE

B.S., Georgia Southern University, 2006

M.S., The University of Georgia, 2009

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

Doctor of Philosophy

ATHENS, GEORGIA

2013

CHARACTERIZATION OF CRISPR RNA BIOGENESIS PATHWAYS IN

STREPTOCOCCUS THERMOPHILUS

by

JASON MICHAEL CARTE

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

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

ACKNOWLEDGEMENTS

I would first like to express my tremendous gratitude to my major professors Dr. Michael Terns and Dr. Rebecca Terns. I am very grateful for your guidance, support, encouragement and the exemplary example you both set for the members of your lab. It has been a true honor to have worked in the Terns' lab. I would like to thank my committee members, Dr. Stephen Hajduk and Dr. Sidney Kushner for their support and advice. I would like to thank all the members of the Terns' lab past and present with whom I had the privilege of working. In particular, I would like to thank Dr. Caryn Hale for support and advice. I would also like to thank Ross Christopher, Duane Jurma, and Justin Smith for their invaluable assistance. Lastly, I would like to thank my family, especially my parents, Alan and Cindy Carte. Without your love and support I would not be where I am today.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

2 MULTIPLE CRISPR-CAS SYSTEMS FUNCTION INDEPENDENTLY IN STREPTOCOCCUS THERMOPHILUS 59

3 DISCUSSION 112

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Invasive genetic elements pose a potentially lethal threat to all cellular organisms. Prokaryotes have therefore developed a number of strategies to combat theses molecular parasites. Most bacteria and archaea encode robust restriction-modification systems that are composed of DNA endonucleases that cleave foreign DNA (for reviews see (Pingoud *et al*., 2005; Tock and Dryden, 2005). In addition, many bacteria are able to block phage adsorption or entry of invader DNA into the cell (reviewed by (Labrie *et al*., 2010)). An added line of defense found in some bacterial species are the abortive infection (Abi) systems, which limit phage infections within a population by inducing suicide in an infected host (reviewed by (Chopin *et al*., 2005; Labrie *et al*., 2010)). The recently discovered CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated) RNA-directed immune systems represent novel defense pathways against genome invaders. Contrary to the aforementioned defense mechanisms, CRISPR-Cas (CC hereafter) systems provide microbes with a highly adaptive and heritable immune system that protects against extrachromasomal elements. Analogously to the well studied eukaryotic RNAi pathways (for reviews see (Agrawal *et al*., 2003; Aravin *et al*., 2007; Tomari and Zamore, 2005)), small host-encoded CRISPR (cr)RNAs direct effector Cas proteins to silence invader DNA or RNA.

Initially identified in *Escherichia coli*, CC immune systems have been found in many other bacterial, and nearly all archaeal and genomes (Ishino *et al*., 1987; Makarova *et al*., 2006). CRISPR arrays are composed of an AT-rich leader sequence followed by a series of short direct

repeats separated by similarly sized unique sequences known as spacers (Jansen *et al*., 2002; Makarova *et al*., 2006; Mojica *et al*., 2005). Several *in silico* analyses of spacers sequences demonstrated that many map to known extrachromasomal elements such as viruses and plasmids (Bolotin *et al*., 2005; Mojica *et al*., 2005; Pourcel *et al*., 2005). In addition, genes encoding putative nucleases, helicases, and nucleic acid binding proteins (*cas* genes) were identified associated with CRISPR arrays (Haft *et al*., 2005; Jansen *et al*., 2002; Makarova *et al*., 2006). These findings led to the prediction that CC modules function as a prokaryotic adaptive immune systems (Lillestol *et al*., 2006; Makarova *et al*., 2006). Experimental evidence for CC-mediated immunity was provided when it was shown that *Streptococcus thermophilus* was capable of becoming refractory to infection by a lytic phage by incorporating phage-derived sequences into its CRISPR array (Barrangou *et al*., 2007). Since this initial study, CC systems in many other organisms have been shown function as adaptive immune systems as well (Cady *et al*., 2012; Erdmann and Garrett, 2012; Millen *et al*., 2012; Nozawa *et al*., 2011; Yosef *et al*., 2012).

Classification and distribution of CC systems

CC systems are highly variable in terms of *cas* gene content and CRISPR repeat sequences. CC modules were initially classified into nine sub-types on the basis of *cas* gene content (Haft *et al*., 2005). A set of six "core" *cas* genes (*cas*1-6) that are widely distributed between CC systems were identified (Haft *et al*., 2005). Genes encoding Cas1 and Cas2 are common to all CC modules, whereas the other core *cas* genes may or may not be present (Haft *et al*., 2005). Each CC subtype was defined on the basis of a cluster of genes that are always found together (sub-type specific) and named after an organism that contains only this sub-type specific cluster (E.g. Cse, Cas-sub-type *E. coli*). A recent re-evaluation of this classification system and a

comparison of distant evolutionary relationships between many sub-type specific *cas* genes, led to a consolidation of CC systems in to three major types (I-III) ((Makarova *et al*., 2011b) and Figure 1.1). Two or more sub-types are contained within each of the major CC system types, and each was given a single letter designation (Types I-A-F, II-A and B, and III-A and B) ((Makarova *et al*., 2011b) and Figure 1.1)). CC system sub-types may also be referred to by their previously assigned "Cas sub-type" name (Cse, Csy, Csa, etc.) (Haft *et al*., 2005). Distantly related sub-type specific *cas* gene orthologs shared between CC systems were re-named Cas7, 8, 9, or 10, while those not shared between systems retained their sub-type specific names (Haft *et al*., 2005; Makarova *et al*., 2011b).

In addition to a high degree of variability in *cas* gene content and organization between the different CC systems, the distribution of CC modules between prokaryotic genomes is also highly variable. Many bacteria and some archaea completely lack CC, whereas other species contain multiple CC modules (Godde and Bickerton, 2006; Makarova et al., 2002). This discrepancy is apparent even among some closely related species, such as the genus of marine bacteria *Shewanella* (Godde and Bickerton, 2006). Thus, it has been suggested that CC modules are themselves mobile genetic elements that are exchanged between prokaryotic genomes via horizontal gene transfer (Godde and Bickerton, 2006; Makarova *et al*., 2002).

Stages of CC immunity

CC immunity is broken down into three mechanistic stages (Figure 1.2). During the first stage, adaptation, a short fragment of invader DNA is inserted into the host CRISPR at the leader-proximal end of the array. The second stage, crRNA biogenesis, is characterized by transcription of the CRISPR array and processing to mature crRNAs that are each composed of a

single invader-derived (or guide) sequence and a portion of the direct repeat at the 3', 5' end, or both. The final stage is invader silencing (or interference), during which the crRNA guides effector Cas proteins to silence DNA or RNA containing a sequences complementary to the crRNA. Current progress regarding each mechanistic stage of CC-mediated immunity is discussed in detail below.

CRISPR adaptation

A defining feature of the CC prokaryotic immune systems is their ability to incorporate fragments of foreign DNA (proto-spacers) into the leader-proximal end a CRISPR array. Although thought to be a universal feature of CC systems, the process by which new spacer sequences are acquired remains the least understood aspect of CC-mediated immunity. A number of recent studies have uncovered stimulating insights regarding the how new spacer sequences are selected, as well the genetic requirements for CRISPR adaptation, however the precise mechanism of adaptation has yet to be elucidated.

Non-random selection of proto-spacers

In silico analyses and early findings suggested that CC systems select novel spacers by identifying short sequence motifs outside of the proto-spacer region, termed proto-spacer adjacent motifs (PAM) (Deveau *et al*., 2008; Horvath *et al*., 2008; Mojica *et al*., 2009). Bioinformatic analysis of a large set of spacers that map to known extra-chromosomal elements revealed that there are distinct di- or trinucleotide PAM sequences associated with many Type I and Type II CC systems (Mojica *et al*., 2009). The two Type II-A CC modules (see Figure 1.1) found in *Streptococcus thermophilus* have been shown to preferentially incorporate spacers from

phages or plasmids that contain distinct PAMs (Deveau *et al*., 2008; Horvath *et al*., 2008; Levin *et al*., 2013; Paez-Espino *et al*., 2013). Experimental analysis of adaptation in *E. coli* (Type I-E) (Datsenko *et al*., 2012; Swarts *et al*., 2012; Yosef *et al*., 2012) and *Sulfolobus solfataricus* (Types I-A and III-B) (Erdmann and Garrett, 2012) have also demonstrated that the majority of novel spacers are derived from regions of the invader that contain adjacently encoded PAM sequences. In addition to their role in adaptation, PAM sequences are also critical for efficient invader silencing in Type I and Type II CC systems (Deveau *et al*., 2008; Garneau *et al*., 2010; Jore *et al*., 2011; Magadan *et al*., 2012; Westra *et al*., 2012a).

Genetic requirements for CC adapation

Recent advances have revealed that the ubiquitous Cas1 and Cas2 proteins are involved in CC adaptation, and two other Cas proteins have been implicated in the process as well. The presence of Cas1 and Cas2 in all CC systems, coupled with their predicted functions as a an integrase and nuclease, respectively, suggested that these proteins may be involved in the process of spacer acquisition (Makarova *et al*., 2006). Two recent studies done in *E. coli* have shown that the Cas1 and Cas2 proteins associated with the Type I-E CC module are both necessary to facilitate incorporation of spacer sequences from M13 phage or a plasmid into the leader-proximal end of a CRISPR array (Datsenko *et al*., 2012; Yosef *et al*., 2012).

In the first study, Yosef *et al.* 2012 showed that new spacers were readily incorporated into a plasmid encoded CRISPR array that was also over-expressing *E. coli* K12 Cas1 and Cas2 in *E. coli* BL21, which lacks an endogenous CC system. The majority of the spacers incorporated were derived from the plasmid containing the CRISPR array and *cas1* and *cas2*. The minimal requirements for adaptation in this Type I-E CC system were determined to be

over-expression Cas1 and Cas2, a sixty base pair region of the CRISPR leader sequence (150 base pairs total) proximal to the first direct repeat, and a single CRISPR repeat sequence (Yosef *et al*., 2012).

In a complementary study published shortly following the work of Yosef *et al.* 2012, it was shown that Cas1 and Cas2 were required for spacer acquisition from M13 phage as well (Datsenko *et al*., 2012). In this study, *E. coli* K12 cells inducibly expressing the Type I-E *cas* genes found in this organism (Figure 1.1) were shown to acquire M13-derived spacers following infection (Datsenko *et al*., 2012). Adaptation was shown to require both Cas2 and a catalytically active Cas1 protein (Babu *et al*., 2011; Datsenko *et al*., 2012). Interestingly, when a spacer sequence targeting M13 was engineered into the CRISPR array, spacer acquisition following infection with M13 was greatly stimulated when the crRNA biogenesis and effector Cas proteins were co-expressed along with Cas1 and Cas2 (Datsenko *et al*., 2012). This phenomenon was termed "priming" (Datsenko *et al*., 2012).

Whereas there was no obvious strand bias for spacer acquisition in the absence of priming, the overwhelming majority of the new spacers incorporated in the primed system were derived from the same strand of M13 genome as the engineered spacer targeting the phage (Datsenko *et al*., 2012). This strand bias was confirmed by switching the orientation of the engineered M13-targeting spacer to the opposite strand (Datsenko *et al*., 2012). Supporting evidence for a strand bias in spacer acquisition was provided in a study in which *E. coli* K12 cells acquired spacers from a high copy number plasmid (Swarts *et al*., 2012). It has been speculated that the priming effect would allow the CC system to rapidly respond to phage "escape" mutants that have mutations within their genomes that allow them to evade CC-mediated silencing (Datsenko *et al*., 2012; Swarts *et al*., 2012). A separate study revealed that *E.*

coli Cas1 physically and genetically interacts with proteins involved in DNA repair, suggesting that non-Cas proteins may also be critical in CC adaptation (Babu *et al*., 2011).

In addition to Cas1 and Cas2, the Type II-A specific protein Csn2 (formerly known as Cas7) and Cas4 have been implicated in adaptation. Deletion of the *csn2* gene from the Type II-A CC module in *Streptococcus thermophilus* appears to block the ability of this module to acquire spacer sequences from either an invading phage or plasmid (Barrangou *et al*., 2007; Garneau *et al*., 2010). Indirect evidence also suggests that Cas4 could be involved in adaptation. Genes encoding Cas4 are found associated with several Type I CC systems and do not appear to be required for crRNA biogenesis or invader silencing in these systems (Makarova *et al*., 2011b). Additionally, in some Type I CC modules, the *cas4* gene is fused with *cas1*, suggesting the two proteins are co-functional (Makarova *et al*., 2011b). Furthermore, Type II-A and II-B CC modules have the same *cas* gene organization, except that the *csn2* gene associated with Type II-A systems is replaced by *cas4* in Type II-B CC modules, implying that Cas4 may function in adaptation as well (Makarova *et al*., 2011b). Finally, the Cas4 protein from *S. solfataricus* has been shown to possess exonuclease activity, which may be required for insertion of new spacer sequences (Zhang *et al*., 2012a).

Biochemical insights into possible CC adaptation mechanism(s)

A number of biochemical analyses have been reported for proteins that are believed to mediate spacer acquisition. How these properties relate to the mechanism of adaptation, however, remains unclear. The Cas1 protein from *Psuedomonas aeruginosa* (Type I-F CC module) was determined to cleave long double-stranded and single-stranded DNA endonucleolytically, generating fragments ~80 base pairs/nucleotides in length, suggesting

perhaps that Cas1 cleaves the genome of an invader to generate proto-spacers (Wiedenheft *et al*., 2009). Conversely, the Cas1 protein from *E. coli* (Type I-E) was shown to cleave single stranded DNA and a number of branched DNA structures corresponding to recombination intermediates (Babu *et al*., 2011). It was therefore speculated that this Cas1 protein may function as a novel holliday junction resolvase during spacer addition (Babu *et al*., 2011). On the other hand, a Cas1 ortholog from *S. solfataricus*, which contains a Type I-A and two Type III-B modules, was shown bind single- or double-stranded RNA and DNA with a high affinity ($K_d \approx$ 18-50 nM) and promote annealing of complementary single stranded DNA, however no nuclease activity was detected for this protein (Han *et al*., 2009).

Reports describing distinct *in vitro* activities of Cas2 proteins have also been reported. *S. solfataricus* contains two Cas2 orthologs. One of these Cas2 proteins (SSO1404) was shown to endonucleolytically cleave single-stranded RNA preferentially in U-rich regions (Beloglazova *et al*., 2008). However, the other Cas2 protein from this organism was shown to cleave double stranded RNA or DNA preferentially at G:C base pairs (Han and Krauss, 2009). The Cas2 protein from *Bacillus halodurans* (Type I-C CC system) displayed endonucleolytic cleavage of double stranded DNA as well (Nam *et al*., 2012). However, a Cas2 ortholog from *Desulfovibrio vulgaris*, which also contains a Type I-C CC module, displayed no apparent cleavage activity against either DNA or RNA (Samai *et al*., 2010).

Comparative structural analysis of Cas2 proteins revealed that they are each composed of a single ferrodoxin fold, a divalent metal binding site, and a variable loop region (Beloglazova *et al*., 2008; Nam *et al*., 2012; Samai *et al*., 2010). It has been speculated that the different activities observed within Cas2 protein orthologs can be attributed to the divergent loop region. How these proteins may function within spacer acquisition remains obscure (Nam *et al*., 2012).

Studies describing Cas4 have recently been published. Members of the Cas4 protein family are found associated with several Type I and III CC modules and Type II-B CC systems (Makarova *et al*., 2011b). They are characterized by a conserved cysteine-rich C-terminal cluster, and were predicted to be RecB-like exonucleases (Makarova *et al*., 2006). Cas4 from *S. solfataricus* was found to cleave single stranded DNA and RNA (only in the presence of $Mn2^{+}$) exonucleolytically with a 3' to 5' directionality (Zhang *et al*., 2012a). Interestingly, a crystal structure of the protein revealed that the conserved cysteine-rich motif forms an iron-sulfur cluster very similar to that observed in the AddB exonuclease from *Bacillus subtillis*, which is involved in DNA double-stranded break repair (Yeeles and Dillingham, 2007; Zhang *et al*., 2012a). It was proposed that Cas4 functions in resecting a single DNA strand of the proto-spacer and/or the host CRISPR array during the insertion of a novel spacer (Zhang *et al*., 2012a).

The Csn2 family are a highly divergent group of proteins found exclusively associated with Type II-A CC modules, and genetic evidence in *S. thermophilus* suggests that Csn2 is required for CC adaptation (Barrangou *et al*., 2007; Garneau *et al*., 2010; Haft *et al*., 2005; Makarova *et al*., 2011b). Despite the high degree of sequence variability within this protein family, crystal structures of four Csn2 orthologs demonstrate that these proteins have a very similar overall architecture (Ellinger *et al*., 2012; Koo *et al*., 2012; Lee *et al*., 2012). Each of the Csn2 proteins that have been crystallized adopt a homotetrameric structure that contains a wide central pore lined with basic amino acid residues (Ellinger *et al*., 2012; Koo *et al*., 2012; Lee *et al*., 2012). Although no enzymatic activity has been detected for Csn2, each of the four proteins studied bind double-stranded DNA non-specifically, suggesting that this protein may play a role in identifying DNA invaders (Ellinger *et al*., 2012; Koo *et al*., 2012; Lee *et al*., 2012).

crRNA biogenesis in Type I and Type III CC systems

The generation of mature crRNAs is a critical step in CC-mediated immunity. Mature crRNAs are composed of a single invader-targeting guide sequence and a portion of the CRISPR repeat at the 5', 3', or both ends. Early evidence suggested that the entire CRISPR repeat-spacer array is transcribed from the AT-rich leader sequence. Subsequent primary processing of pre-crRNAs occurs via endonucleolytic cleavage within each direct repeat, followed in some cases by trimming to smaller species (Hale *et al*., 2008; Lillestol *et al*., 2006; Lillestol *et al*., 2009; Tang *et al*., 2005). Subsequent studies have shown that primary processing of pre-crRNAs in Type I and Type III CC systems is carried out by Cas6 superfamily (or Cas5d) endoribonucleases followed in some cases by further 3' processing ((Brouns *et al*., 2008; Carte *et al*., 2008; Haurwitz *et al*., 2010; Lintner *et al*., 2011; Richter *et al*., 2012) and Figure 1.3). The first direct evidence for this type of primary processing was provided by Brouns *et al* in 2008 when they found that the Cas6e protein (formerly known as Cse3 or CasE) associated with the Type I-E CC system in *E. coli* is a divalent metal-independent endoribonuclease that cleaves within each CRISPR repeat (Brouns *et al*., 2008). This cleavage generates RNAs composed of a single guide sequence, an 8 nucleotide 5' repeat-derived tag sequence, and a 21 nucleotide 3' repeat tag (Brouns *et al*., 2008).

Cas6 superfamily ribonucleases in primary pre-crRNA processing

Similar Cas6-like metal-independent ribonucleases have been identified associated with each of the other Type I CC systems, as well as both Type III-A and III-B CC systems, and have been shown to carry out primary processing of pre-crRNAs in these CC systems (Carte *et al*., 2008; Garside *et al*., 2012; Hatoum-Aslan *et al*., 2011; Haurwitz *et al*., 2010; Lintner *et al*.,

2011; Przybilski *et al*., 2011; Richter *et al*., 2012). With only one exception (Cas5d, discussed below), Cas6-like nucleases cleave CRISPR repeats exactly 8 nucleotides upstream of the 5' end of the guide sequence generating the mature 5' crRNA ends. These proteins are highly divergent in terms of amino acid sequences. However, detailed biochemical and structural analyses have revealed many common features. All Cas6 superfamily enzymes were initially classified as Repeat-associated mysterious proteins (RAMPs), which are characterized by a glycine-rich C-terminal motif and at least one predicted ferrodoxin fold, and were predicted to be RNA-binding proteins (Haft *et al*., 2005; Makarova *et al*., 2002; Makarova *et al*., 2006). Although similar in structure and cleavage activity, the modes of RNA binding and the predicted active site amino acids vary considerably among members of this superfamily.

Cas6 from *Pyrococcus furiosus*

The first Cas6 superfamily member for which detailed biochemical and structural studies were performed was Cas6 from *Pyrococcus furiosus* (Pf Cas6) (Carte *et al*., 2008). *P. furiosus* harbors seven CRISPR loci, each containing nearly identical direct repeats and three distinct *cas* gene modules: Type I-A, I-B, and III-B CC systems. Pf Cas6 was shown to be responsible for primary processing of pre-crRNAs in *P. furiosus*, producing 1X intermediates (a single guide with partial repeats at either end) that are then sorted and loaded into three distinct protein complexes, each composed of sub-type specific proteins ((Carte *et al*., 2010; Hale *et al*., 2009; Hale *et al*., 2012b) and Terns lab unpublished data)). Cleavage by Pf Cas6 within the CRISPR repeat produces RNAs with 5' hydroxyl and 2'-3' cyclic phosphate chemical end groups, consistent with its divalent metal independent cleavage activity (Carte *et al*., 2010; Carte *et al*., 2008).

The crystal structure of Pf Cas6 revealed that it is composed of duplicated ferrodoxin folds separated by a central cleft, which contains a number of positively charged amino acids as well as the C-terminal glycine rich loop characteristic of Cas6 proteins (Carte *et al*., 2008). Sequence conservation within closely related Cas6 family members, and a similar architecture to the archaeal tRNA-splicing endonuclease led to the prediction that the active site is composed of a tyrosine, histidine, and lysine catalytic triad (Calvin and Li, 2008; Carte *et al*., 2008). Individual mutations of these residues to alanine resulted in a loss of cleavage activity, but did not interfere with the ability of Pf Cas6 to bind the repeat RNA, confirming the importance of these residues in catalysis (Carte *et al*., 2010). A general acid-base catalytic mechanism was proposed in which the tyrosine and histidine act as a proton donor/acceptor pair.

Pf Cas6 was found to have an intriguing mode of RNA substrate recognition. Using a combination of RNA mutagenesis and RNA footprinting, Cas6 was found to specifically recognize nucleotides 2-10 of the repeat sequence, while cleavage occurs distally between nucleotides 22 and 23 (8 nucleotides upstream of the 5' end of the guide sequence) (Carte *et al*., 2010; Carte *et al*., 2008). The CRISPR repeats found in *P. furiosus* were found to be unstructured in solution, consistent with an *in silico* analysis performed on CRISPR repeat sequences that placed this repeat sequence in an unstructured subtype (Carte *et al*., 2010; Kunin *et al*., 2007).

A co-crystal structure of Pf Cas6 bound with a non-hydrolyzable CRISPR repeat RNA confirmed that the protein binds single-stranded RNA and that nucleotides 2-10 directly contact the protein along the positively charged central cleft (Wang *et al*., 2011). The remainder of the RNA could not be resolved in the crystal structure. However, the location of RNA binding indicates that the RNA is bound on the reverse face of the protein and suggests that it wraps

around through the central cleft to the opposite face where the putative active site is located, positioning the scissile phosphate for cleavage (Wang *et al*., 2011).

A co-crystal structure of a distantly related Cas6 ortholog from *Sulfolobus solfataricus* P2 with its cognate repeat RNA provided additional insight into the mode of Cas6 substrate recognition (Shao and Li, 2013). In contrast to the co-crystal structure of PF Cas6 with its repeat RNA, the portion of the repeat RNA proximal to the cleavage site was resolved in this crystal structure. The co-structure shows that this Cas6 protein also interacts with ssRNA, however the protein appears to stabilize a short stem-loop structure on the RNA directly upstream of the scissile phosphate that is not present in the unbound form of the RNA (Shao and Li, 2013). The authors suggest that cleavage downstream of a stem-loop by Cas6 superfamily members may be a universal feature among Cas6 superfamily members despite the highly divergent nature of both the proteins and the repeat RNAs (Shao and Li, 2013).

Other members of the Cas6 superfamily

Detailed structural and biochemical analyses have been performed for a number of other Cas6 superfamily members (Gesner *et al*., 2011; Haurwitz *et al*., 2010; Haurwitz *et al*., 2012; Lintner *et al*., 2011; Przybilski *et al*., 2011; Richter *et al*., 2012; Sashital *et al*., 2011; Scholz *et al*., 2013; Sternberg *et al*., 2012). These studies reveal a number of intriguing similarities and differences with Pf Cas6. The structure of Cas6e protein from the *Thermus thermophilus* Type I-E CC system demonstrated that this protein has a very similar overall architecture as Pf Cas6 despite a complete lack sequence identity outside of the C-terminal glycine-rich motif (Ebihara *et al*., 2006). The protein is composed of tandem ferredoxin folds separated by a central cleft that contains a number of basic residues. Two later studies were also published simultaneously

describing Cas6e from *T. thermophilus* (Gesner *et al*., 2011; Sashital *et al*., 2011). Both groups found that the protein cleaves its cognate CRISPR repeat RNA directly downstream of a short stem-loop structure generating the mature 5' repeat tag in a metal-independent manner, consistent with a previous study with a Cas6e ortholog from *E. coli* (Brouns *et al*., 2008).

The mode of substrate RNA recognition by Cas6e was revealed by co-crystal structures with Cas6e bound to a fragment of the CRISPR repeat (Gesner *et al*., 2011; Sashital *et al*., 2011). The partial repeat RNAs form a duplex that contacts the protein along the positively charged groove between the ferredoxin folds, in a similar location to where Pf Cas6 binds to its single stranded repeat RNA (Gesner *et al*., 2011; Sashital *et al*., 2011; Wang *et al*., 2011).

The predicted active site of Cas6e also differs from Pf Cas6, however there is not a consensus on exactly which amino acids are likely involved. In both studies in which co-crystal structures of Cas6e bound to repeat RNA fragments were solved, the scissile phosphate of the RNA was located in very close proximity to a conserved tyrosine residue, and both groups propose that this residue may be involved in catalysis. A previous study of Cas6e in *E. coli* showed that a conserved histidine was indispensable for cleavage activity (Brouns *et al*., 2008). In the co-crystal structure published by Sashital *et al* 2011, this histidine is positioned opposite the tyrosine residue close to the scissile phosphate, and the authors propose that this residue is directly involved in catalysis through a general acid-base catalytic mechanism. In contrast, in the Gesner *et al.* 2011 structure, the same histidine is positioned such that it appears to interact with the RNA phosphate backbone one nucleotide upstream of the cleavage site. The authors propose that instead of being directly involved in catalysis, that this histidine may play a critical role in properly orienting the repeat RNA in the active site, which they speculate is made up of

the conserved tyrosine and two arginine residues that are located closer to the scissile phosphate bond in their structure (Gesner *et al*., 2011).

Another well-studied member of the Cas6 superfamily is Cas6f (formerly Csy4) associated with the Type I-F CC module in *Psuedomonas aeruginosa* (Haurwitz *et al*., 2010; Haurwitz *et al*., 2012; Przybilski *et al*., 2011; Sternberg *et al*., 2012). Similar to other members of the Cas6 superfamily, this protein was found cleave endonucleolytically within its cognate CRISPR repeat RNA at the base of a stem-loop in a metal-independent manner, generating an 8 nucleotide 5' repeat-derived tag with a 5' hydroxyl chemical end and a 3' phosphate (or 2'-3' cyclic phosphate) (Haurwitz *et al*., 2010; Przybilski *et al*., 2011).

The structure of the protein revealed that it is composed of a single ferredoxin fold and a C-terminal alpha helical domain not found on other Cas6 enzymes, separated by a cleft containing a number of basic amino acids (Haurwitz *et al*., 2010). Co-crystal structures of Cas6f along with a non-hydrolyzable truncated CRISPR repeat RNA illustrates that the basic patch in a cleft between the ferredoxin fold and the C-terminal motif is primarily responsible for recognizing the substrate RNA through sequence and structure specific contacts (Haurwitz *et al*., 2010; Sternberg *et al*., 2012). The RNA contacts the protein in both single stranded and duplexed regions near the site of cleavage.

The predicted active site of Cas6f differs from that of both Pf Cas6 and Cas6e. The location of the scissile phosphate in close proximity to highly conserved histidine and serine residues lead the authors to propose that these residues catalyze cleavage of the phosphodiester backbone using a general acid-base catalytic mechanism. Indeed mutation of either residue abolished detectable cleavage activity with no detectable effects on RNA binding. The predicted

acid-base catalytic mechanism has been further supported by subsequent kinetic analyses and additional structural studies (Haurwitz *et al*., 2012).

Cas6 superfamily members have also been identified associated with Type I-A, I-B, I-D, and Type III-A CC systems ((Hatoum-Aslan *et al*., 2011; Lintner *et al*., 2011; Richter *et al*., 2012; Scholz *et al*., 2013) and Chapter 2)). RNA-deep sequencing in *Clostridium thermocellum* and *Methancoccus maripaludis*, each harboring a single Type I-B CC system, revealed that nearly all of the crRNAs sequenced contain an 8 nucleotide repeat-derived tag sequence, characteristic of cleavage by a Cas6 enzyme (Richter *et al*., 2012). A single gene with weak sequence identity to Pf Cas6 was identified in the *cas* gene clusters of both *C. thermocellum* and *M. maripaludis*, and was given the name Cas6b. Cas6b from *M. maripaludis* was shown to cleave its cognate CRISPR repeat downstream of a predicted stem-loop structure 8 nucleotides upstream of the guide sequence, but did not cleave a repeat from another CRISPR array (Richter *et al*., 2012).

A structural model generated of *M. maripaludis* Cas6 shows a very similar overall architecture to Pf Cas6, composed of duplicated ferredoxin folds separated by a central cleft (Richter *et al*., 2012). Mutation of two conserved histidine residues results in a loss of cleavage, and the authors speculated that this histidine dyad forms the active site of Cas6b, which further highlights the diversity observed among Cas6 superfamily active site compositions (Richter *et al*., 2012).

Sulfolobus solfataricus contains six CRISPR arrays, Type I-A and III-B *cas* gene clusters, and four predicted Cas6 proteins (Lillestol *et al*., 2006). One of the Cas6 proteins (SSO2004) was found to associate directly with a multi-protein complex composed of Type I-A proteins,

and was shown to cleave the CRISPR repeat RNAs derived from pre-crRNAs that were found to associate Type I-A complex in metal-independent fashion (Lintner *et al*., 2011).

Strong genetic evidence in *Staphylococcus epidermidis* and biochemical evidence in *Streptococcus thermophilus* indicate that the Cas6 protein associated with Type III-A CC modules carries out pre-crRNA primary processing in this CC system ((Hatoum-Aslan *et al*., 2011; Marraffini and Sontheimer, 2008) and Chapter 2). A conserved histidine in this Cas6 was shown to be required for RNA cleavage, but not RNA binding, suggesting that it may play a key role in catalysis (Chapter 2).

An interesting case of independently functioning Cas6 proteins in the same organism was presented in a study done in the cyanobacterium *Synechocystis 6803* (Scholz *et al*., 2013). This microbe harbors a mega-plasmid that encodes three separate CC modules, each with its own CRISPR array and an adjacently encoded *cas* gene cluster. The sequences of the repeats vary considerably between the CRISPR arrays. Two of the modules, CRISPRs 1 and 2 (Type I-D and III-B, respectively) encode highly divergent *cas6* genes. RNA deep sequencing revealed that each of the three CRISPR arrays are transcriptionally active and the crRNAs generated from CRISPRs 1 and 2 retain an 8 nucleotide repeat-derived tag sequence characteristic of cleavage by a Cas6 enzyme (Scholz *et al*., 2013). Individual knockouts of either *cas6* gene lead to a loss in crRNA processing from its cognate CRISPR array, but not either of the other two, suggesting that each Cas6 protein functions specifically and independently in biogenesis of crRNAs (Scholz *et al*., 2013).

Primary processing in Type I-C CC systems

Among the Type I CC systems, only the Type I-C CC system does not contain an obvious Cas6 superfamily member. Primary processing of pre-crRNAs in Type I-C CC systems was therefore predicted to be carried out by either Cas5 or Cas7 (Makarova *et al*., 2011a; Makarova *et al*., 2011b). A recent study has shown that crRNA primary processing in this CC system is carried out by a member of the Cas5 superfamily, Cas5d (Garside *et al*., 2012).

Cas5 proteins were initially classified as a group of orthologous proteins containing a conserved 43 amino acid N-terminal region with subtype specific regions outside of the N-terminus (Haft *et al*., 2005). The Cas5d proteins from *Mannheimia succiniciproducens* and *Thermus thermophilus* were shown to possess CRISPR repeat-specific divalent metal-independent cleavage activity, but rather than generating an 8 nucleotide repeat-derived tag sequence, cleavage occurs 11 nucleotides upstream of the 5' end of the guide sequence (Garside *et al*., 2012).

The crystal structure of *M. succiniciproducens* Cas5d protein showed that, like many Cas6 proteins, is composed of duplicated ferredoxin folds separated by a cleft which contains a large basic patch (Garside *et al*., 2012). Using structural alignments with Cas6e and Cas6f, a similar mechanism for substrate recognition was proposed in which CRISPR repeat RNA forms a duplex that is docked onto the protein along the basic patch found the in central cleft of Cas5d (Garside *et al*., 2012).

Identification of catalytic residues proved elusive as mutations within conserved residues did not result in an appreciable effect on cleavage activity. Mutation of a conserved glutamate found in a loop within the central cleft, on the other hand, abolished RNA binding, and it was speculated that this residue is critical substrate recognition (Garside *et al*., 2012). Thus, the

enzymes responsible (or likely responsible) for primary processing of pre-crRNAs in all Type I and both Type III CC systems have been identified. However, much less is known about how maturation of crRNA intermediates in occurs these systems.

crRNA maturation in Type I CC systems

In Type I-E and Type I-F CC systems, cleavage of pre-crRNAs by Cas6e or Cas6f produces RNAs that function as the mature crRNAs in these systems (Jore *et al*., 2012; Wiedenheft *et al*., 2011b). However, in other Type I CC systems, as well as Type III CC systems, cleavage by Cas6 produces a 1X crRNA intermediates that undergo further processing whereby variable amounts of the 3' repeat sequences, and sometimes a portion of the guide, are removed. In Type I-A, I-B, and I-D CC systems variable amounts of the 3' repeat sequence are removed by unknown mechanisms.

In *Sulfolobus solfataricus* and *S. acidocalderins*, both of which contain Type I-A CC systems, northern analysis revealed that the 1X crRNA intermediates are trimmed at the 3' end leading to the accumulation of several discreet species down to a size corresponding roughly to a crRNA composed of an 8 nucleotide 5' repeat tag and a single guide sequence (Lillestol *et al*., 2006; Lillestol *et al*., 2009; Tang *et al*., 2005). This processing pattern was supported by a later study in which crRNAs from *S. solfataricus* were cloned and sequenced (Lintner *et al*., 2011).

In the Type I-B CC systems of *Clostridium thermocellum* and *Methanococcus maripaludis*, deep RNA sequencing revealed that the most abundant crRNAs sequenced retained 2-5 nucleotides of 3' repeat sequence (Richter *et al*., 2012). Similarly, RNA deep sequencing and Northern analysis of RNAs derived from the cyanobacterium *Synechocystis* 6803 revealed that crRNAs derived from the Type I-D CC module contain variable 3' ends ranging from the

Cas6 generated 1X intermediate to species lacking ~5 nucleotides of the guide sequence (Scholz *et al*., 2013).

Maturation of crRNAs in Type III CC systems

Though 3' end processing of the Cas6 generated 1X crRNA intermediates in Type I CC systems (where it occurs) produces crRNAs with a high degree of 3' variability, 3' end processing in Type III CC systems gives rise to discreet crRNAs of a defined length from the mature 5' end. crRNAs associated with the Type III-B Cmr complex from *Pyrococcus furiosus* have been shown to be either 39 or 45 nucleotides in length, independent of the length of the guide sequence (Hale *et al*., 2009). The mature crRNAs retain little to no 3' repeat sequence and may contain a 3' truncated guide sequence. Both the 39 and 45 nucleotide species have been shown to be functional in guiding the complex to cleave complementary RNAs. However, the mechanism by which the 3' processing occurs is poorly understood (Hale *et al*., 2009; Hale *et al*., 2012b). In addition, the most abundant crRNAs found associated with Cmr complex from *Sulfolobus solfataricus* are 46 nucleotides in length, also independent of the length of the guide sequence (Zhang *et al*., 2012b).

Analogously to Cmr-associated crRNAs, mature crRNAs found with the Type III-A CC system in *Staphylococcus epidermidis* are 43 and 37 nucleotides in length, contain little to no 3' repeat sequence and a truncated guide on the 37 nucleotide crRNA species (Hatoum-Aslan *et al*., 2011). Mutational analysis of *cas* genes in this organism illustrate that in the absence of the Csm2, Csm3, and Csm5 proteins, processing of the 1X intermediate to the mature forms is disrupted, suggesting that these proteins may be directly involved in this processing step (Hatoum-Aslan *et al*., 2011).

Type II CC system crRNA biogenesis

Whereas primary processing of pre-crRNAs in Type I and Type III CC systems is carried out by a Cas6 (or Cas5) endoribonuclease, in Type II CC systems this step is surprisingly catalyzed by the non-Cas protein RNase III, and requires a small non-coding RNA known as a *trans*-activating (tra)crRNA (Deltcheva *et al*., 2011). RNase III is a double-stranded RNA-specific endoribonuclease that, in bacteria, plays critical roles in biogenesis of rRNA and regulation of gene expression by cleaving certain mRNAs (reviewed in (Drider and Condon, 2004)).

Through differential RNA deep sequencing in *Streptococcus pyogenes*, Deltcheva *et al.* 2011 discovered a small non-coding RNA (termed tracrRNA) encoded directly upstream of the Type II-A CC module in this organism (Deltcheva *et al*., 2011). Primary transcripts of 171 and 89 nucleotides were detected, along with a smaller 75 nucleotide processed form. Sequence analysis of the primary tracrRNA transcripts revealed a 25 nucleotide region which contains near perfect complementary (single mis-match) to the CRISPR repeats encoded within the downstream CRISPR array. The RNA-seq analysis also revealed that mature crRNAs from this Type II CC system contain a 3' repeat-derived tag sequence 19-22 nucleotides in length and a truncated guide sequence of 20 nucleotides (average spacer length = 30 bp). Alignment of the 75 nucleotide processed tracrRNA with individual sequencing reads of the 3' repeat tag region of mature crRNAs revealed a two base-pair overhang at the end of the duplex, which is characteristic of cleavage by RNase III (Drider and Condon, 2004).

Genetic evidence for a role of RNase III in crRNA biogenesis in Type II CC systems was provided when the gene encoding RNase III was knocked out in *S. pyogenes* (Deltcheva *et al*., 2011). crRNA biogenesis was disrupted and no mature crRNAs were detected by Northern

analysis. In the absence of RNase III, processing of the tracrRNA was also disrupted, with the 75 nucleotide processed form no longer being detected. In addition, deletion of the tracrRNA also led to the loss of crRNA processing. Taken together, these results suggest that the tracrRNA, once transcribed, base-pairs with the direct repeat sequences within the CRISPR primary transcript, and this duplex is cleaved by RNase III, producing 1X crRNA intermediates composed of a single guide sequence and fragments of the repeat at either end (Figure 1.4).

Analogously to Type I and Type III CC systems, primary processing in Type II CC systems by RNase III and tracrRNA produces 1 X crRNA intermediates. However, processing of this intermediate to the mature form is quite different in Type II CC systems. In contrast to the 3' end trimming occurs in Type I and Type III CC systems, Type II mature crRNAs are generated by removal of the 5' repeat fragment and ~10 nucleotides of the guide sequence. The mechanism by which this occurs has not yet been determined. However, it has been speculated that maturation of the 1X intermediates is carried out by Cas9 (formerly Csn1 or Cas5) ((Deltcheva *et al*., 2011; Makarova *et al*., 2011a) and Chapter 2). Cas9 proteins are large, multi-domain proteins that contain both HNH and RuvC nuclease active sites (Haft *et al*., 2005; Makarova *et al*., 2006). Each active site has been shown to cleave a single strand of a double-stranded invader DNA in a crRNA-dependent manner (Gasiunas *et al*., 2012; Jinek *et al*., 2012). However, it has also been shown that Cas9 is required for the accumulation of mature crRNAs, leaving open the possibility that this protein also catalyzes maturation of 1X crRNA intermediates to mature crRNAs ((Deltcheva *et al*., 2011) and Chapter 2).

Invader Silencing in Type I CC systems

Type I CC modules are a highly diverse group of CC systems found in both bacteria and archaea that are characterized by the presence of the core *cas1-3* and *cas6* genes along with a set of sub-type specific *cas* genes ((Makarova *et al*., 2011b) and Figure 1.1)). Silencing of invaders within these diverse systems has been shown in some cases, and predicted in others, to be mediated by a common mechanism. In general, the sub-type specific Cas proteins form (or are predicted to form) multi-protein surveillance complexes composed of sub-type specific Cas proteins that use a crRNA guide to detect foreign DNA invaders, and once detected, the complex recruits a Cas3 nuclease/helicase which then cleaves the invader ((Brouns *et al*., 2008; Cady *et al*., 2012; Lintner *et al*., 2011; Makarova *et al*., 2011b; Wiedenheft *et al*., 2011b) and Figure 1.5)). Among the Type I CC systems, Types I-A, I-B, I-E, and I-F have been shown to be capable of silencing invader DNA (Brouns *et al*., 2008; Cady *et al*., 2012; Elmore *et al*., 2013; Fischer *et al*., 2012; Gudbergsdottir *et al*., 2011; Peng *et al*., 2013).

A number of Type I systems have been investigated biochemically. However, the Type I-E CC system present in *E. coli* is the most well-studied among them. Type I-E modules are composed of *cas1-3*, *cas6e*, *cse1-2*, as well as *cas5* (*cas5e*) and *cas7* (*cse4*) orthologs (Haft *et al*., 2005; Makarova *et al*., 2011b). A complex containing *E. coli* K12 Cse1, Cse2, Cas5, Cas7, Cas6e, and mature crRNA was isolated following over-expression the entire CC module in *E. coli* BL21, which lacks an endogenous CC system (Brouns *et al*., 2008). The crRNP was given the name Cascade (CRISPR-associated complex for anti-viral defense) (Brouns *et al*., 2008). Analogous Cascade-like complexes associated with Type I-A and Type I-F CC systems have been isolated (Lintner *et al*., 2011; Wiedenheft *et al*., 2011b).

Over-expression of both *E. coli* K12 Cascade and Cas3 in *E. coli* BL21 was shown to provide defense against lambda phage when an artificial CRISPR array containing spacers targeting the phage genome was also expressed (Brouns *et al*., 2008). Cas3 proteins associated with several Type I CC systems in (including *E. coli* Type I-E) have been shown to possess both DNA endonuclease and DExD-box mediated helicase activities *in vitro*, consistent with a role as an effector nuclease (Beloglazova *et al*., 2011; Mulepati and Bailey, 2011; Sinkunas *et al*., 2011; Sinkunas *et al*., 2013). A point mutation within the predicted HD nuclease active site within Cas3 blocked phage resistance, suggesting that Cas3 directly cleaves invader DNA (Brouns *et al*., 2008).

Purified *E. coli* Cascade has been shown to bind double- and single-stranded DNA containing a crRNA target sequence. There was a higher affinity when a PAM sequence was present (Jore *et al*., 2011; Westra *et al*., 2012a; Wiedenheft *et al*., 2011a). Structural, biochemical and genetic evidence from multiple studies suggest that primary DNA target recognition is mediated by the PAM, and that the first 7 nucleotides of the proto-spacer target proximal to the PAM (termed the seed sequence) must be completely complementary with the crRNA guide in order for efficient recognition and silencing to occur (Jore *et al*., 2011; Semenova *et al*., 2011; Westra *et al*., 2012b). In the DNA target bound *E. coli* Cascade complex, structural and biochemical evidence suggests that the Cse1subunit recruits Cas3, which then degrades the target DNA (Westra *et al*., 2012b; Wiedenheft *et al*., 2011a). The Type I-E CC module from *Streptococcus thermophilus* has been also shown to silence DNA invaders when expressed in *E. coli*, and the recombinant Cascade complex plus Cas3 also cleaves target DNA in a PAM dependent manner (Sapranauskas *et al*., 2011; Sinkunas *et al*., 2013). An analogous

mechanism of invader detection by a sub-type specific crRNP and silencing by a Cas3 protein has been predicted for the other, less studied, Type I CC systems (Makarova *et al*., 2011b).

Invader Silencing in Type II CC systems

Type II CC systems are composed of only four cas genes: *cas1*, *cas2*, *cas9* (formerly *csn1* or *cas5*), and either *csn2* (Type II-A) or *cas4* (Type II-B) and are found exclusively in bacterial genomes ((Haft *et al*., 2005; Makarova *et al*., 2011a) and Figure 1.1)). The first report of CC-mediated adaptive immunity was reported in a study conducted in *Streptococus thermophilus* (Barrangou *et al*., 2007). Infection of *S. thermophilus* with the virulent phage phi2972 resulted in a number of phage-resistant colonies, each of which had acquired 1-4 phage-derived spacer sequences within one of the two CRISPR arrays associated with Type II-A CC modules (later named CRISPR 1) found in this organism (Barrangou *et al*., 2007). In a later study, the other Type II-A CC module in this system (named CRISPR 3) was also shown to acquire phage-resistance by incorporating invader-derived spacers within its CRISPR array (Horvath *et al*., 2008). The two Type II-A CC modules in *S. thermophilus* contain the same *cas* gene content and organization, however the CRISPR repeat sequences and Cas protein sequences are highly divergent (Horvath *et al*., 2008). In addition, it has been shown that *S. thermophilus* CRISPR 1 and CRISPR 3 rely on distinct PAM sequences for both spacer acquisition and invader silencing, NNAGAAW, and NGGNGN respectively (Deveau *et al*., 2008; Garneau *et al*., 2010; Gasiunas *et al*., 2012; Jinek *et al*., 2012; Magadan *et al*., 2012).

Genetic evidence for the protein responsible for carrying out invader silencing in Type II CC systems was provided when a gene knock-out of *cas9* from *S. thermophilus* CRISPR 1 resulted in a loss of defense in previously a previously resistant strain, suggesting that this

protein may be an effector nuclease (Barrangou *et al*., 2007). Subsequent studies have demonstrated that both Type II CC systems in *S. thermophilus* are capable of cleaving plasmid or phage DNA that contain complementary sequences to crRNAs and a PAM sequence (Garneau *et al*., 2010; Magadan *et al*., 2012). Cleavage of target DNA produces a blunt cut that maps to 3 base pairs upstream of the 3' end of the target sequence (Garneau *et al*., 2010; Magadan *et al*., 2012). Cleavage of invader DNA was disrupted in the absence of either Cas9 protein, providing further evidence that Cas9 is the effector nuclease associated with Type II CC systems (Garneau *et al*., 2010; Magadan *et al*., 2012). In a separate study, the Cas9 protein from *S. thermophilus* CRISPR 3 and a CRISPR array were sufficient for silencing an invader which contained proto-spacer matching the CRISPR locus when both were expressed in *E. coli*, further supporting the hypothesis that Cas9 is the effector nuclease associated with Type II CC systems (Sapranauskas *et al*., 2011).

Cas9 proteins are large multi-domain proteins that contain two nuclease active sites, a HNH restriction enzyme-like site, and a RuvC resolvase-like site (Haft *et al*., 2005; Makarova *et al*., 2006; Makarova *et al*., 2011b). Three recent studies have shown that Cas9 proteins from *S. thermophilus* CRISPR3 and *Streptococcus pyogenes* utilize a crRNA guide to directly cleave double stranded DNAs which contain a target sequence and intact PAM (Gasiunas *et al*., 2012; Jinek *et al*., 2012; Karvelis *et al*., 2013). Interestingly, in addition to a crRNA guide, in order for Cas9 to cleave target DNA the tracrRNA, which is involved in crRNA biogenesis in Type II CC systems, must be present as well ((Gasiunas *et al*., 2012; Jinek *et al*., 2012; Karvelis *et al*., 2013) and Figure 1.5)). Furthermore, evidence suggests that each strand of the target DNA is cleaved independently by one of the nuclease active sites. Mutation of either the HNH or RuvC active sites resulted in a nicking of one strand or the other, suggesting that both active sites cooperate to

generate a double stranded break in the target DNA. The strand complementary to the crRNA is cleaved by the HNH active site, and cleavage of the non-complementary strand requires the RuvC active site ((Gasiunas *et al*., 2012; Jinek *et al*., 2012) and Figure 1.5)). Given its ability to generate site-specific double-stranded DNA breaks, Cas9 has recently been utilized in genome editing in humans, zebrafish, yeast, and bacteria (Cho *et al*., 2013; Cong *et al*., 2013; Dicarlo *et al*., 2013; Hwang *et al*., 2013; Jiang *et al*., 2013; Jinek *et al*., 2013; Mali *et al*., 2013; Ramalingam *et al*., 2013).

Invader silencing in Type III CC systems

Type III CC modules are present in many bacterial and archaeal genomes, commonly as accessory CC systems; that is, they are frequently present in organisms that contain at least one other CC system, and they are often encoded directly adjacent to another CC module (Haft *et al*., 2005). The Type III-A and III-B CC are both encode of a number of RAMP proteins as well as a Cas10 ortholog ((Haft *et al*., 2005; Makarova *et al*., 2011b) and Figure 1.1)). Although similar in gene content, the two modules differ in their silencing activities, with Type III-A CC systems targeting invader DNA (Marraffini and Sontheimer, 2008), while Type III-B targets RNA and perhaps DNA (Hale *et al*., 2012a; Hale *et al*., 2009; Zhang *et al*., 2012b).

DNA silencing in Type III-A CC systems

Type III-A CC systems are composed of three RAMP proteins, Csm3-5, as well as Csm2, Csm6, and a Cas10 (Csm1) protein (Haft *et al*., 2005; Makarova *et al*., 2011b). The Type III-A CC module from *Staphylococcus epidermidis* was shown to silence plasmid DNA (Bikard *et al*., 2012; Marraffini and Sontheimer, 2008). *S. epidermidis* harbors a single CRISPR array

composed of four direct repeats and three unique spacers. The first spacer in the CRISPR array perfectly matches the *nickase* (*nes*) gene found on all *Staphlyococcal* conjugative plasmids and crRNAs containing this guide sequence as well as the Cas6 generated 8 nucleotide repeat-derived tag sequence have been detected (Marraffini and Sontheimer, 2008; Marraffini and Sontheimer, 2010). It was shown that this spacer conferred resistance to *S. epidermidis* from either conjugation or transformation of a plasmid containing the *nes* gene (Marraffini and Sontheimer, 2008). Nickase activity is required for plasmid conjugation, so in order to determine whether the Type III-A system targets plasmid DNA or the *nes* mRNA, a self-splicing intron was introduced within the *nes* target sequence (Marraffini and Sontheimer, 2008). This blocked CC mediated-plasmid silencing, indicating that this Type III-A CC module targets DNA (Marraffini and Sontheimer, 2008).

Unlike Type I and Type II CC modules, Type III CC systems do not appear to require a PAM sequence for their silencing activities. However, since Type III-A systems target DNA, it would be expected that a mechanism existed to allow the silencing complex to distinguish invader DNA from its own genome, given that both contain crRNA target sequences. Rather than using a defined PAM sequence, the Type III-A CC system from *S. epidemidis* distinguishes self from non-self using the crRNA 5' repeat tag sequence. Extensive mutational analysis of both the crRNA 5' tag sequence and an 8 base pair region just upstream a complementary DNA target sequence revealed that silencing by this Type III-A CC system is blocked when the crRNA tag is able to base-pair with the upstream region of a target at any single position between -2 and -4 of the target sequence (Marraffini and Sontheimer, 2010). Given that this base pairing would occur if the effector complex targeted the host genome (since the 5' tag would be completely

complementary to the CRISPR repeat), a basis for discrimination between a non-self target DNA and the chromosome (self) was established (Marraffini and Sontheimer, 2010).

Type III-B CC systems silence invader RNA and possibly DNA

While Type III-A CC systems (and all other CC systems described) silence invader DNA, the Type III-B Cas RAMP module (Cmr) complex has been shown to target and cleave complementary invader single stranded RNA ((Hale *et al*., 2009; Zhang *et al*., 2012b) and Figure 1.5)). The Cmr module is a six (sometimes seven) gene cluster that contains four RAMP proteins (cmr1, 3, 4, and 6), *cmr5*, and a *cas10* (*cmr2*) ortholog (and sometimes a *cmr7* gene, see below) (Haft *et al*., 2005). A complex containing Cmr1, Cmr3-6, Cas10, and crRNAs of 39 and 45 nucleotides in length was isolated from *Pyrococcus furiosus* (Hale *et al*., 2009). This complex was shown to cleave single-stranded RNAs complementary to crRNAs contained within the purified complex into several discreet products (Hale *et al*., 2009). Cleavage of complementary RNAs, like Type III-A silencing, is not thought to require a PAM sequence (Hale *et al*., 2012a; Hale *et al*., 2009; Zhang *et al*., 2012b). Reconstitution of the complex with recombinant proteins expressed and isolated from *E. coli* demonstrated a similar RNA cleavage activity as the native complex, and required a crRNA guide (Hale *et al*., 2009). All of the proteins, with the exception of Cmr5, were shown to be required for efficient cleavage by the recombinant complex (Hale *et al*., 2009). A later study of the same complex demonstrated that the 5' 8 nucleotide repeat-derived tag, as well as a 5' hydroxyl group generated by Cas6 cleavage are required for cleavage activity (Carte *et al*., 2008; Hale *et al*., 2012a). Additionally, Hale *et al.* 2012 showed that the Cmr module can be programmed to silence RNAs of choice by simply appending a target sequence to the 5' 8 nucleotide repeat tag sequence, opening the possibility

for selective gene silencing in prokaryotes analogous to RNAi-mediated gene knockdown in eukaryotes (Hale *et al*., 2012a).

An orthologous Cmr complex isolated from *Sulfolobus solfataricus* was also shown to cleave complementary RNAs as well (Zhang *et al*., 2012b). This complex contains an addition Cmr protein, Cmr7, that is only present in Cmr modules found in the order Solfalobales (Zhang *et al*., 2012b). Although both *P. furiosus* Cmr and *S. solfataricus* Cmr modules cleave complementary RNAs, the two complexes differ in their respective sites of cleavage within target RNAs. Evidence suggests that the *P. furiosus* Cmr module cleaves at a fixed distance (14 bp) from the 3' end of the base-pairing between the crRNA and the target in a sequence non-specific manner (Hale *et al*., 2012a; Hale *et al*., 2009). The *S. solfataricus* Cmr module, on the other hand, cleaves preferentially between AU dinucleotides (Zhang *et al*., 2012b). Further highlighting the differential cleavage activities between the two complexes, the *S. solfataricus* Cmr module appears to cleave the crRNA guide as well as the target RNA, whereas the *P. furiosus* Cmr module does not seem to do so (Hale *et al*., 2012a; Hale *et al*., 2009; Zhang *et al*., 2012b).

Interestingly, in addition to RNA silencing, certain Type III-B CC systems may also be capable of DNA silencing. A recent study demonstrated that one of the two Cmr modules found in *Sulfolobus islandicus* is able to block plasmid transformation when the plasmid contains a crRNA target sequence that is transcribed (Deng *et al*., 2013). Like Cmr cleavage of complementary RNAs, this plasmid silencing does not require a PAM sequence (Deng *et al*., 2013). In addition to the Cmr complex proteins, this silencing appears to require the Csx1 protein (Deng *et al*., 2013). Proteins of the Csx1 family are highly diverse and are found associated with ~50% of archaeal Type III CC modules (Garrett *et al*., 2011; Haft *et al*., 2005).

The structure of Csx1 from *P. furiosus* has been solved; however, the role played by this protein within the putative Cmr-mediated, transcription-dependent DNA silencing remains unclear (Kim *et al*., 2013).

Phage response to CC silencing

CC systems identify and silence genome invaders. However, phages have developed a number of strategies to counter this silencing, and in at least one case have even hijacked an entire CC module for its own use in evading host defense mechanisms (Bondy-Denomy *et al*., 2013; Seed *et al*., 2013). Silencing by CC systems relies on sequence complementarity between the crRNA guide and the invader DNA (or RNA). It therefore stands to reason that if a phage or virus were able to alter its genome such that this base pairing were disrupted, then CC defense would be neutralized. Indeed many CRISPR escape mutant (CEM) phages have been isolated from *Streptococcus thermophilus* and *E. coli* (Barrangou *et al*., 2007; Datsenko *et al*., 2012; Deveau *et al*., 2008; Levin *et al*., 2013). These CEMs almost invariably contain mutations within the seed sequence of the proto-spacer or within the PAM, which presumably blocks CC-mediated silencing (Datsenko *et al*., 2012; Levin *et al*., 2013).

In addition to simple mutations, a recent study reported that some phages encode proteins that actively block CC silencing (Bondy-Denomy *et al*., 2013). Three CC-resistant lysogenic phages were isolated from *Pseudomonas aerugninosa* that contained intact proto-spacer and PAM sequences and therefore would be expected to be silenced by the Type I-F CC module found in the host (Bondy-Denomy *et al*., 2013). Further analysis revealed that between these three phages, there were eight genes of unknown function capable of ablating CC defense (Bondy-Denomy *et al*., 2013). The presence of these genes on an invader blocked CC

interference but did not appear to alter the expression levels of either the endogenous *cas* genes or crRNA. Thus, the mechanism by which these predicted proteins evade CC defense remains a mystery (Bondy-Denomy *et al*., 2013).

In another recent study, five CC-resistant lytic bacteriophages isolated from *Vibrio Cholera* were found to encode their own Type I-F CC modules with CRISPR arrays containing spacer sequences matching the *V. cholera* genome (Seed *et al*., 2013). When either the *V. cholera*-targeting spacer sequences in the phage CRISPR array, or the proto-spacer sequence from the host chromosome were deleted, bacteriophages were no longer able to efficiently infect *V. cholera* (Seed *et al*., 2013). In addition, the low number of phages from which the *V. cholera*-targeting spacers had been deleted that were able to re-infect the host had incorporated new host-derived spacers within their CRISPR arrays, suggesting that the phage encoded CC module functions an adaptive immune system which allows the phage to evade host defense (Seed *et al*., 2013).

Overview

The aim of this work was to understand crRNA biogenesis in *Streptococcus thermophilus* DGCC7710. *S. thermophilus* is a gram positive lactic acid bacterium that is important for the dairy industry in the production of yogurt and various cheeses. This organism contains four CC systems (CRISPRs 1-4), including a representative of each of the three major types (I-III), and has become a model for studying CC defense. Each of the CC systems is composed of a distinct CRISPR repeat-spacer array and *cas* gene cluster. At the outset of this work, nothing was known about crRNA expression or biogenesis pathways from the four CC systems present in *S. thermophilus*.

Through RNA deep sequencing and Northern analysis we show that each of the four CRISPR arrays in *S. thermophilus* is transcriptionally active and produces mature crRNAs. The crRNAs derived from each CC system contain distinct combinations of guide and repeat tag sequences. In addition, the factors involved in crRNA biogenesis were investigated. We found that the Cas6 protein associated with CRISPR 2 (Type III-A CC system) specifically binds and cleaves pre-crRNAs from its cognate CRISPR array generating the mature 5' crRNA ends. The Cas6e protein encoded within the Type I-E CRISPR 4 CC module, on the other hand, specifically recognizes and cleaves pre-crRNAs from CRISPR 4 generating the mature 5' end of these crRNAs. In addition, we show that the Cas9 protein associated with CRISPR 1 (Type II-A) is required for maturation of crRNAs from this CC system. The absence of Cas9 from CRISPR 1 had no effect on crRNA processing from other three CRISPR arrays, notably CRISPR 3, which is also a Type II-A CC system.

Taken together, these findings indicate that crRNA biogenesis from each of the four CC systems found in *S. thermophilus* is carried out by Cas proteins found within the adjacently encoded *cas* gene cluster independently of the other CC systems. These results provide a model for understanding organisms with multiple CC systems and suggest that bacterial strains may be engineered with co-existing CC modules without cross-talk between the systems.

References

Agrawal, N., P.V. Dasaradhi, A. Mohmmed, P. Malhotra, R.K. Bhatnagar, and S.K. Mukherjee. 2003. RNA interference: biology, mechanism, and applications. *Microbiology and Molecular Biology Reviews : MMBR*. 67:657-685.

Aravin, A.A., G.J. Hannon, and J. Brennecke. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*. 318:761-764.

Babu, M., N. Beloglazova, R. Flick, C. Graham, T. Skarina, B. Nocek, A. Gagarinova, O. Pogoutse, G. Brown, A. Binkowski, S. Phanse, A. Joachimiak, E.V. Koonin, A. Savchenko, A. Emili, J. Greenblatt, A.M. Edwards, and A.F. Yakunin. 2011. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Molecular Microbiology*. 79:484-502.

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

Beloglazova, N., G. Brown, M.D. Zimmerman, M. Proudfoot, K.S. Makarova, M. Kudritska, S. Kochinyan, S. Wang, M. Chruszcz, W. Minor, E.V. Koonin, A.M. Edwards, A. Savchenko, and A.F. Yakunin. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *The Journal of Biological Chemistry*. 283:20361-20371.

Beloglazova, N., P. Petit, R. Flick, G. Brown, A. Savchenko, and A.F. Yakunin. 2011. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *The EMBO Journal*. 30:4616-4627.

Bikard, D., A. Hatoum-Aslan, D. Mucida, and L.A. Marraffini. 2012. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell host & Microbe*. 12:177-186.

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

Bondy-Denomy, J., A. Pawluk, K.L. Maxwell, and A.R. Davidson. 2013. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*. 493:429-432.

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

Cady, K.C., J. Bondy-Denomy, G.E. Heussler, A.R. Davidson, and G.A. O'Toole. 2012. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *Journal of Bacteriology*. 194:5728-5738.

Calvin, K., and H. Li. 2008. RNA-splicing endonuclease structure and function. *Cellular and Molecular Life Sciences : CMLS*. 65:1176-1185.

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

Carte, J., R. Wang, H. Li, R.M. Terns, and M.P. Terns. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & Development*. 22:3489-3496.

Cho, S.W., S. Kim, J.M. Kim, and J.S. Kim. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology*.

Chopin, M.C., A. Chopin, and E. Bidnenko. 2005. Phage abortive infection in lactococci: variations on a theme. *Current Opinion in Microbiology*. 8:473-479.

Cong, L., F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, and F. Zhang. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*.

Datsenko, K.A., K. Pougach, A. Tikhonov, B.L. Wanner, K. Severinov, and E. Semenova. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*. 3:945.

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

Deng, L., R.A. Garrett, S.A. Shah, X. Peng, and Q. She. 2013. A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. *Molecular Microbiology*. 87:1088-1099.

Deveau, H., R. Barrangou, J.E. Garneau, J. Labonte, C. Fremaux, P. Boyaval, D.A. Romero, P. Horvath, and S. Moineau. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology*. 190:1390-1400.

Dicarlo, J.E., J.E. Norville, P. Mali, X. Rios, J. Aach, and G.M. Church. 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Research*.

Drider, D., and C. Condon. 2004. The continuing story of endoribonuclease III. *Journal of Molecular Microbiology and Biotechnology*. 8:195-200.

Ebihara, A., M. Yao, R. Masui, I. Tanaka, S. Yokoyama, and S. Kuramitsu. 2006. Crystal structure of hypothetical protein TTHB192 from T*hermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Science : a Publication of the Protein Society*. 15:1494-1499.

Ellinger, P., Z. Arslan, R. Wurm, B. Tschapek, C. MacKenzie, K. Pfeffer, S. Panjikar, R. Wagner, L. Schmitt, H. Gohlke, U. Pul, and S.H. Smits. 2012. The crystal structure of the CRISPR-associated protein Csn2 from *Streptococcus agalactiae*. *Journal of Structural Biology*. 178:350-362.

Elmore, J.R., Y. Yokooji, T. Sato, S. Olson, C.V. Glover, 3rd, B.R. Graveley, H. Atomi, R.M. Terns, and M.P. Terns. 2013. Programmable plasmid interference by the CRISPR-Cas system in *Thermococcus kodakarensis*. *RNA Biology*. 10.

Erdmann, S., and R.A. Garrett. 2012. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Molecular Microbiology*. 85:1044-1056.

Fischer, S., L.K. Maier, B. Stoll, J. Brendel, E. Fischer, F. Pfeiffer, M. Dyall-Smith, and A. Marchfelder. 2012. An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *The Journal of Biological Chemistry*. 287:33351-33363.

Garneau, J.E., M.E. Dupuis, M. Villion, D.A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A.H. Magadan, and S. Moineau. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 468:67-71.

Garrett, R.A., G. Vestergaard, and S.A. Shah. 2011. Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends in Microbiology*. 19:549-556.

Garside, E.L., M.J. Schellenberg, E.M. Gesner, J.B. Bonanno, J.M. Sauder, S.K. Burley, S.C. Almo, G. Mehta, and A.M. MacMillan. 2012. Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*. 18:2020-2028.

Gasiunas, G., R. Barrangou, P. Horvath, and V. Siksnys. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*.

Gesner, E.M., M.J. Schellenberg, E.L. Garside, M.M. George, and A.M. Macmillan. 2011. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nature Structural & Molecular Biology*. 18:688-692.

Godde, J.S., and A. Bickerton. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*. 62:718-729.

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

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

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

Hale, C., S. Majumdar, J. Elmore, N. Pfister, M. Compton, S. Olson, A.M. Resch, C.V. Glover, 3rd, B.R. Graveley, R.M. Terns, and M.P. Terns. 2012a. Essential features and rational

design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell*. 45:292-302.

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

Hale, C.R., S. Majumdar, J. Elmore, N. Pfister, M. Compton, S. Olson, A.M. Resch, C.V. Glover, 3rd, B.R. Graveley, R.M. Terns, and M.P. Terns. 2012b. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell*. 45:292-302.

Han, D., and G. Krauss. 2009. Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2. *FEBS Letters*. 583:771-776.

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

Hatoum-Aslan, A., I. Maniv, and L.A. Marraffini. 2011. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proceedings of the National Academy of Sciences of the United States of America*. 108:21218-21222.

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

Haurwitz, R.E., S.H. Sternberg, and J.A. Doudna. 2012. Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *The EMBO Journal*. 31:2824-2832.

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

Hwang, W.Y., Y. Fu, D. Reyon, M.L. Maeder, S.Q. Tsai, J.D. Sander, R.T. Peterson, J.R. Yeh, and J.K. Joung. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology*.

Ishino, Y., H. Shinagawa, K. Makino, M. Amemura, and A. Nakata. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia col*i, and identification of the gene product. *Journal of Bacteriology*. 169:5429-5433.

Jansen, R., J.D. Embden, W. Gaastra, and L.M. Schouls. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*. 43:1565-1575.

Jiang, W., D. Bikard, D. Cox, F. Zhang, and L.A. Marraffini. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology*.

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

Jinek, M., A. East, A. Cheng, S. Lin, E. Ma, and J. Doudna. 2013. RNA-programmed genome editing in human cells. *eLife*. 2:e00471.

Jore, M.M., S.J. Brouns, and J. van der Oost. 2012. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harbor Perspectives in Biology*. 4.

Jore, M.M., M. Lundgren, E. van Duijn, J.B. Bultema, E.R. Westra, S.P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M.R. Beijer, A. Barendregt, K. Zhou, A.P.

Snijders, M.J. Dickman, J.A. Doudna, E.J. Boekema, A.J. Heck, J. van der Oost, and S.J. Brouns. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology*. 18:529-536.

Karvelis, T., G. Gasiunas, A. Miksys, R. Barrangou, P. Horvath, and V. Siksnys. 2013. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. *RNA Biology*. 10.

Kim, Y.K., Y.G. Kim, and B.H. Oh. 2013. Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of *Pyrococcus furiosus*. *Proteins*. 81:261-270.

Koo, Y., D.K. Jung, and E. Bae. 2012. Crystal structure of *Streptococcus pyogenes* Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PloS One*. 7:e33401.

Kunin, V., R. Sorek, and P. Hugenholtz. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*. 8:R61.

Labrie, S.J., J.E. Samson, and S. Moineau. 2010. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*. 8:317-327.

Lee, K.H., S.G. Lee, K. Eun Lee, H. Jeon, H. Robinson, and B.H. Oh. 2012. Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins*. 80:2573-2582.

Levin, B.R., S. Moineau, M. Bushman, and R. Barrangou. 2013. The Population and Evolutionary Dynamics of Phage and Bacteria with CRISPR-Mediated Immunity. *PLoS Genetics*. 9:e1003312.

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

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

Lintner, N.G., M. Kerou, S.K. Brumfield, S. Graham, H. Liu, J.H. Naismith, M. Sdano, N. Peng, Q. She, V. Copie, M.J. Young, M.F. White, and C.M. Lawrence. 2011. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *The Journal of Biological Chemistry*. 286:21643-21656.

Magadan, A.H., M.E. Dupuis, M. Villion, and S. Moineau. 2012. Cleavage of Phage DNA by the *Streptococcus thermophilus* CRISPR3-Cas System. *PloS One*. 7:e40913.

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

Makarova, K.S., L. Aravind, Y.I. Wolf, and E.V. Koonin. 2011a. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology Direct*. 6:38.

Makarova, K.S., N.V. Grishin, S.A. Shabalina, Y.I. Wolf, and E.V. Koonin. 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. *Biology Direct*. 1:7.

Makarova, K.S., D.H. Haft, R. Barrangou, S.J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F.J. Mojica, Y.I. Wolf, A.F. Yakunin, J. van der Oost, and E.V. Koonin. 2011b.

Evolution and classification of the CRISPR-Cas systems. *Nature Review Microbiology*. 9:467-477.

Mali, P., L. Yang, K.M. Esvelt, J. Aach, M. Guell, J.E. Dicarlo, J.E. Norville, and G.M. Church. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science*.

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

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

Millen, A.M., P. Horvath, P. Boyaval, and D.A. Romero. 2012. Mobile CRISPR/Cas-mediated bacteriophage resistance in *Lactococcus lactis*. *PloS One*. 7:e51663.

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

Mojica, F.J., C. Diez-Villasenor, J. Garcia-Martinez, and E. Soria. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*. 60:174-182.

Mulepati, S., and S. Bailey. 2011. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *The Journal of Biological Chemistry*. 286:31896-31903.

Nam, K.H., F. Ding, C. Haitjema, Q. Huang, M.P. DeLisa, and A. Ke. 2012. Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *The Journal of Biological Chemistry*. 287:35943-35952.

Nozawa, T., N. Furukawa, C. Aikawa, T. Watanabe, B. Haobam, K. Kurokawa, F. Maruyama, and I. Nakagawa. 2011. CRISPR inhibition of prophage acquisition in *Streptococcus pyogene*s. *PloS One*. 6:e19543.

Paez-Espino, D., W. Morovic, C.L. Sun, B.C. Thomas, K. Ueda, B. Stahl, R. Barrangou, and J.F. Banfield. 2013. Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature Communications*. 4:1430.

Peng, W., H. Li, S. Hallstrom, N. Peng, Y.X. Liang, and Q. She. 2013. Genetic determinants of PAM-dependent DNA targeting and pre-crRNA processing in *Sulfolobus islandicus*. *RNA Biology*. 10.

Pingoud, A., M. Fuxreiter, V. Pingoud, and W. Wende. 2005. Type II restriction endonucleases: structure and mechanism. *Cellular and Molecular Life Sciences : CMLS*. 62:685-707.

Pourcel, C., G. Salvignol, and G. Vergnaud. 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.

Przybilski, R., C. Richter, T. Gristwood, J.S. Clulow, R.B. Vercoe, and P.C. Fineran. 2011. Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biology*. 8:517-528.

Ramalingam, S., N. Annaluru, and S. Chandrasegaran. 2013. A CRISPR way to engineer the human genome. *Genome biology*. 14:107.

Richter, H., J. Zoephel, J. Schermuly, D. Maticzka, R. Backofen, and L. Randau. 2012. Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic acids research*.

Samai, P., P. Smith, and S. Shuman. 2010. Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. *Acta crystallographica. Section F, Structural Biology and Crystallization Communications*. 66:1552-1556.

Sapranauskas, R., G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys. 2011. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Research*. 39:9275-9282.

Sashital, D.G., M. Jinek, and J.A. Doudna. 2011. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nature Structural & Molecular Biology*. 18:680-687.

Scholz, I., S.J. Lange, S. Hein, W.R. Hess, and R. Backofen. 2013. CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein. *PloS One*. 8:e56470.

Seed, K.D., D.W. Lazinski, S.B. Calderwood, and A. Camilli. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*. 494:489-491.

Semenova, E., M.M. Jore, K.A. Datsenko, A. Semenova, E.R. Westra, B. Wanner, J. van der Oost, S.J. Brouns, and K. Severinov. 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America*. 108:10098-10103.

Shao, Y., and H. Li. 2013. Recognition and Cleavage of a Nonstructured CRISPR RNA by Its Processing Endoribonuclease Cas6. *Structure*.

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

Sinkunas, T., G. Gasiunas, S.P. Waghmare, M.J. Dickman, R. Barrangou, P. Horvath, and V. Siksnys. 2013. In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *The EMBO Journal*.

Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna. 2012. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*. 18:661-672.

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

Tang, T.H., N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J.P. Bachellerie, and A. Huttenhofer. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Molecular Microbiology*. 55:469-481.

Terns, M.P., and R.M. Terns. 2011. CRISPR-based adaptive immune systems. *Current Opinion in Microbiology*. 14:321-327.

Tock, M.R., and D.T. Dryden. 2005. The biology of restriction and anti-restriction. *Current Opinion in Microbiology*. 8:466-472.

Tomari, Y., and P.D. Zamore. 2005. Perspective: machines for RNAi. *Genes & Development*. 19:517-529.

Wang, R., G. Preamplume, M.P. Terns, R.M. Terns, and H. Li. 2011. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*. 19:257-264.

Westra, E.R., B. Nilges, P.B. van Erp, J. van der Oost, R.T. Dame, and S.J. Brouns. 2012a. Cascade-mediated binding and bending of negatively supercoiled DNA. *RNA Biology*. 9:1134-1138.

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

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

Wiedenheft, B., E. van Duijn, J.B. Bultema, S.P. Waghmare, K. Zhou, A. Barendregt, W. Westphal, A.J. Heck, E.J. Boekema, M.J. Dickman, and J.A. Doudna. 2011b. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences of the United States of America*. 108:10092-10097.

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

Yeeles, J.T., and M.S. Dillingham. 2007. A dual-nuclease mechanism for DNA break processing by AddAB-type helicase-nucleases. *Journal of Molecular Biology*. 371:66-78.

Yosef, I., M.G. Goren, and U. Qimron. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research*. 40:5569-5576.

Zhang, J., T. Kasciukovic, and M.F. White. 2012a. The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PloS One*. 7:e47232.

Zhang, J., C. Rouillon, M. Kerou, J. Reeks, K. Brugger, S. Graham, J. Reimann, G. Cannone, H. Liu, S.V. Albers, J.H. Naismith, L. Spagnolo, and M.F. White. 2012b. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular Cell*. 45:303-313.

Figure 1.1. Overview of *cas* gene organization and distribution. CC systems are divided into three major types (I-III) and sub-types. CC sub-type nomenclature according to Haft *et al*. 2006 is shown to the right as well. The typical *cas* gene content and organization for each sub-type are shown. Genes encoding proteins that have been shown, or are predicted to be involved in spacer acquisition are shown in green. Pre-crRNA cleaving endonucleases of the *cas6* superfamily are shown in blue. Distantly related orthologs that are shared between multiple CC sub-types are shown in red. Nucleases responsible for invader silencing are shown in purple. Note that in Type I-F CC systems the *cas2* and *cas3* genes are fused. Figure adapted from (Makarova *et al*., 2011b).

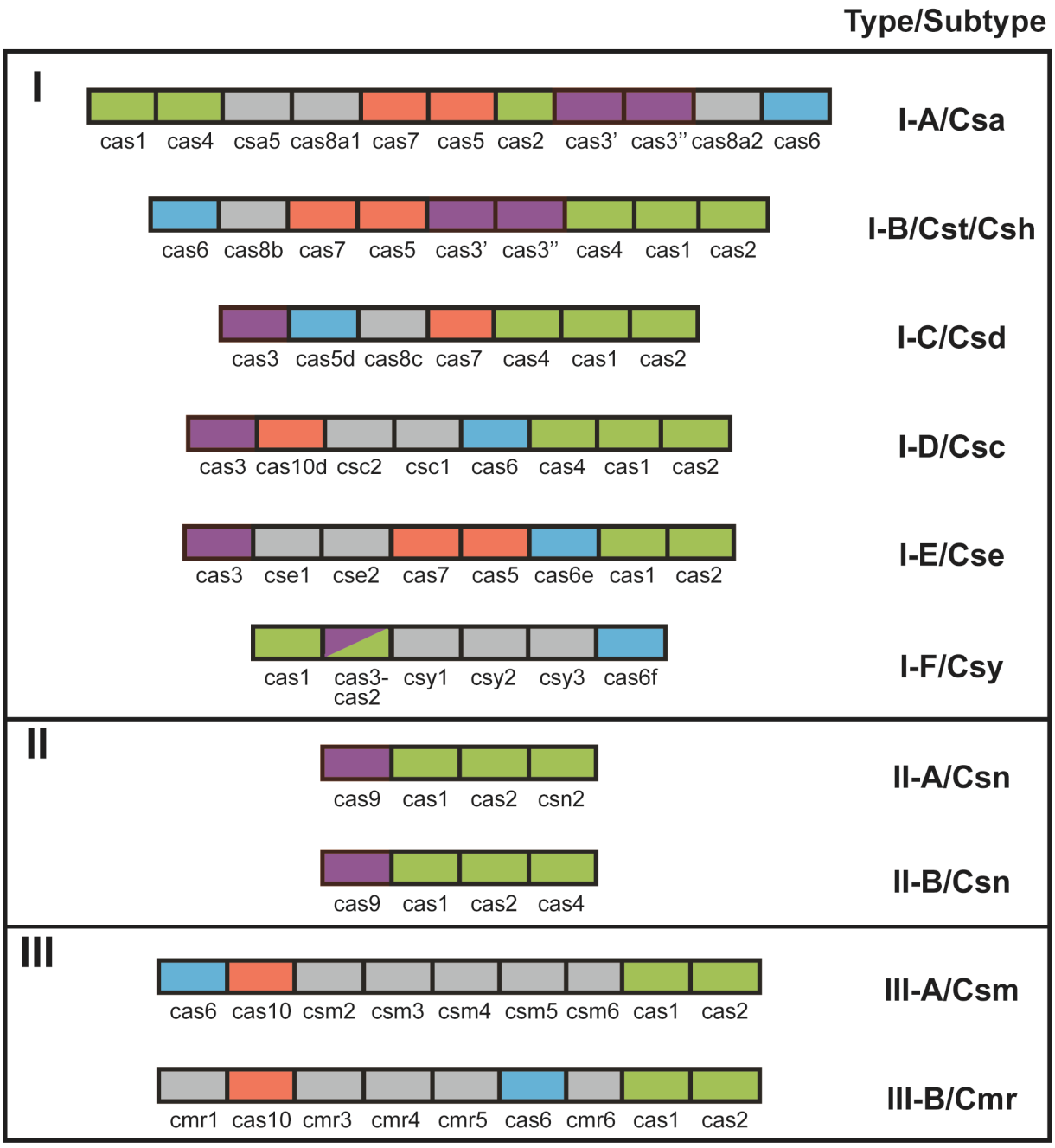

Type/Subtype
I
cas1 cas4 csa5 cas8a1 cas7 cas5 cas2 cas3' cas3'' cas8a2 cas6
I-A/Csa
cas6 cas8b cas7 cas5 cas3' cas3'' cas4 cas1 cas2
I-B/Cst/Csh
cas3 cas5d cas8c cas7 cas4 cas1 cas2
I-C/Csd
cas3 cas10d csc2 csc1 cas6 cas4 cas1 cas2
I-D/Csc
cas3 cse1 cse2 cas7 cas5 cas6e cas1 cas2
I-E/Cse
cas1 cas3- cas2 csy1 csy2 csy3 cas6f
I-F/Csy
II
cas9 cas1 cas2 csn2
II-A/Csn
cas9 cas1 cas2 cas4
II-B/Csn
III
cas6 cas10 csm2 csm3 csm4 csm5 csm6 cas1 cas2
III-A/Csm
cmr1 cas10 cmr3 cmr4 cmr5 cas6 cmr6 cas1 cas2
III-B/Cmr

Figure 1.2. Overview of the three stages of the CC immune mechanism. Fragments derived from foreign genetic elements (proto-spacers), such as phages, are inserted into the host CRISPR locus (adaptation). The CRISPR array is transcribed and processed into a pool of crRNAs each containing a single invader-targeting guide sequence and a fragment of the CRISPR repeat (crRNA biogenesis). The crRNAs are then loaded into a Cas protein-containing effector complex and guide the complex to silence invader (invader silencing). Figure adapted from (Terns and Terns, 2011).

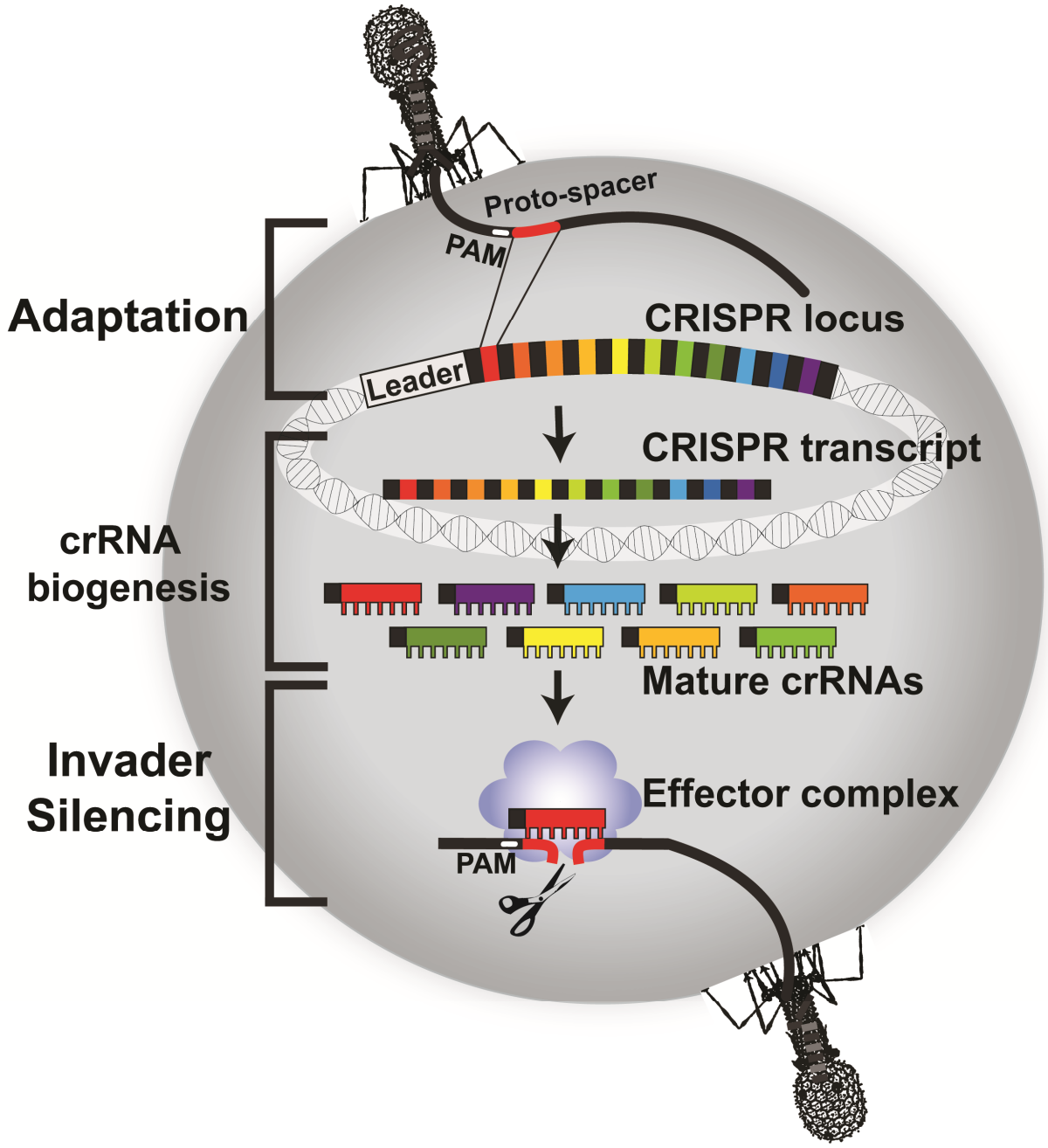

Proto-spacer
PAM
Adaptation
CRISPR locus
Leader
CRISPR transcript
crRNA
biogenesis
Mature crRNAs
Invader
Silencing
Effector complex
PAM

Figure 1.3. Overview of crRNA biogenesis in Type I and Type III CC systems. CRISPR primary transcripts (pre-crRNAs) are cleaved endonucleolytically by a member of the Cas6 (*or Cas5d in Type I-C CC systems) superfamily member producing RNAs that contain a single guide sequence flanked by fragments of the CRISPR repeat. In Type I-E and I-F CC systems, this RNA functions as the mature crRNA. In other Type I CC systems, the 1X intermediate is further processed producing crRNAs with highly variable 3' ends. In Type III CC systems, the 1X intermediate is trimmed to one or two discreet crRNA species that are a defined length from the 5' end of the crRNA.

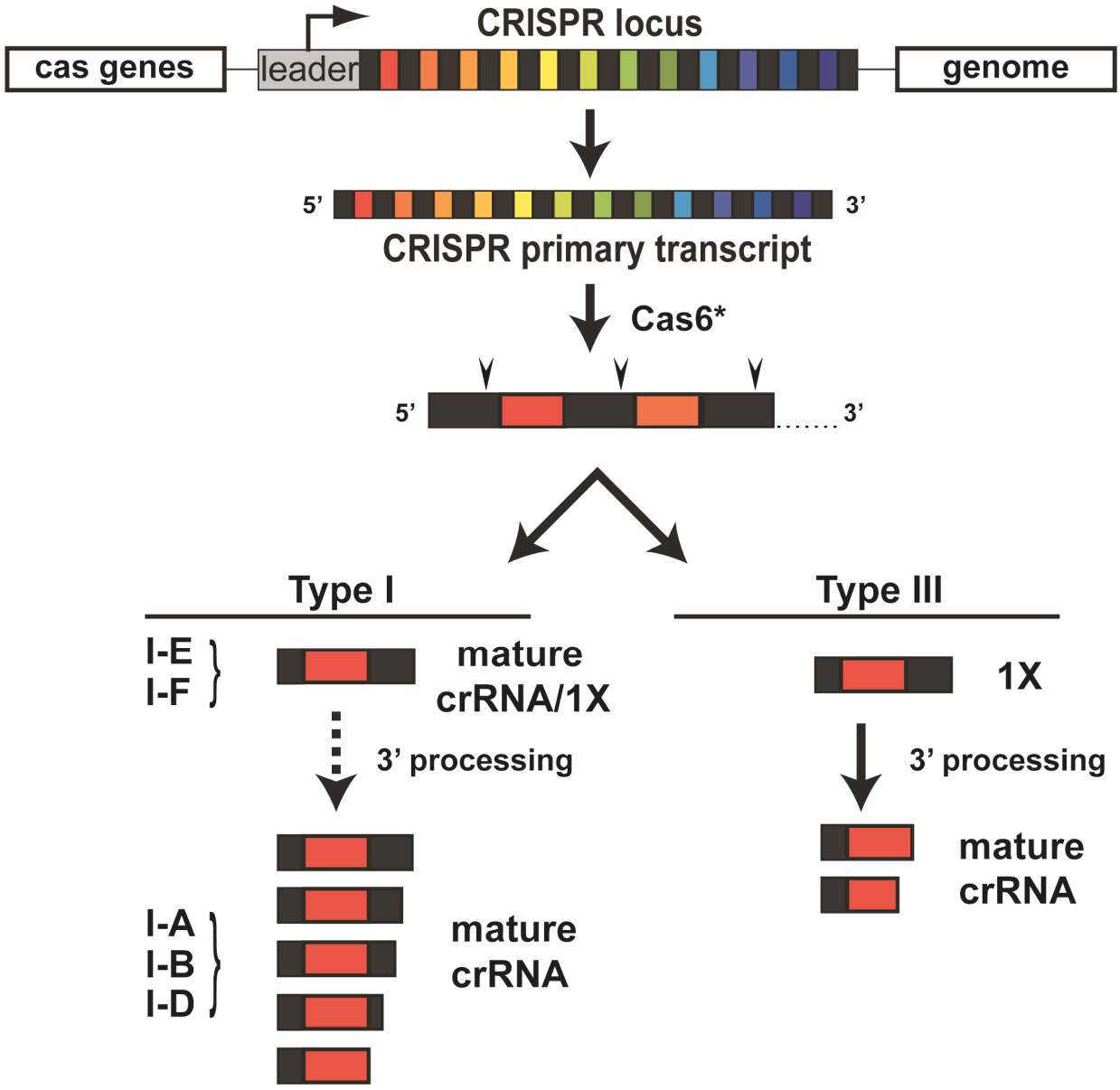

CRISPR locus
cas genes
leader
genome
5'
3'
CRISPR primary transcript
Cas6*
5'
3'
Type I
Type III
I-E
I-F
mature
crRNA/1X
3' processing
1X
3' processing
mature
crRNA
I-A
I-B
I-D
mature
crRNA

Figure 1.4. crRNA biogenesis in Type II CC systems. A *trans*-activating (tra)crRNA hybridizes with each direct repeat within a pre-crRNA. These RNA duplexes are then cleaved by RNase III, producing 1X crRNA intermediates composed of a single guide sequence flanked by fragments of the CRISPR repeat. The 1X intermediates are processed to mature crRNA by removal of the 5' repeat fragment and ~10 nucleotides of the guide sequence.

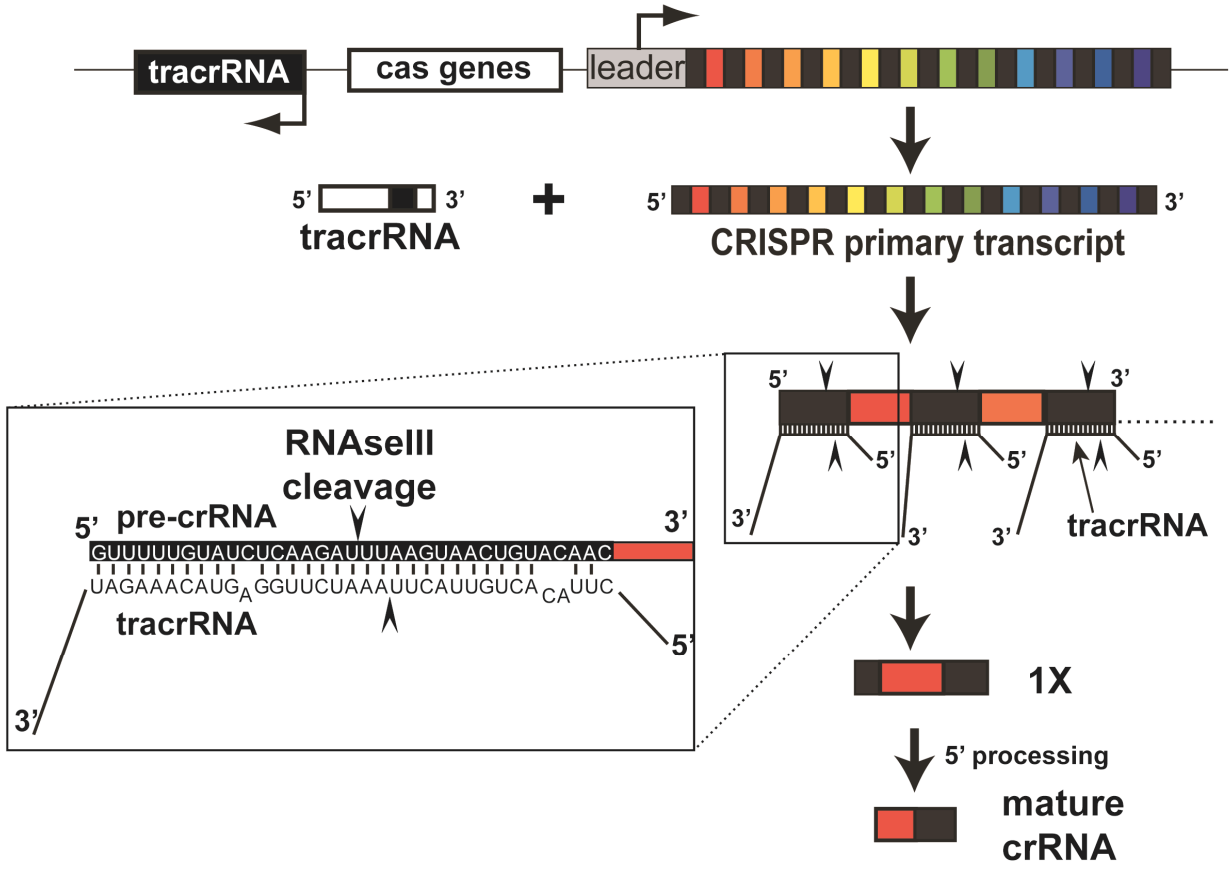
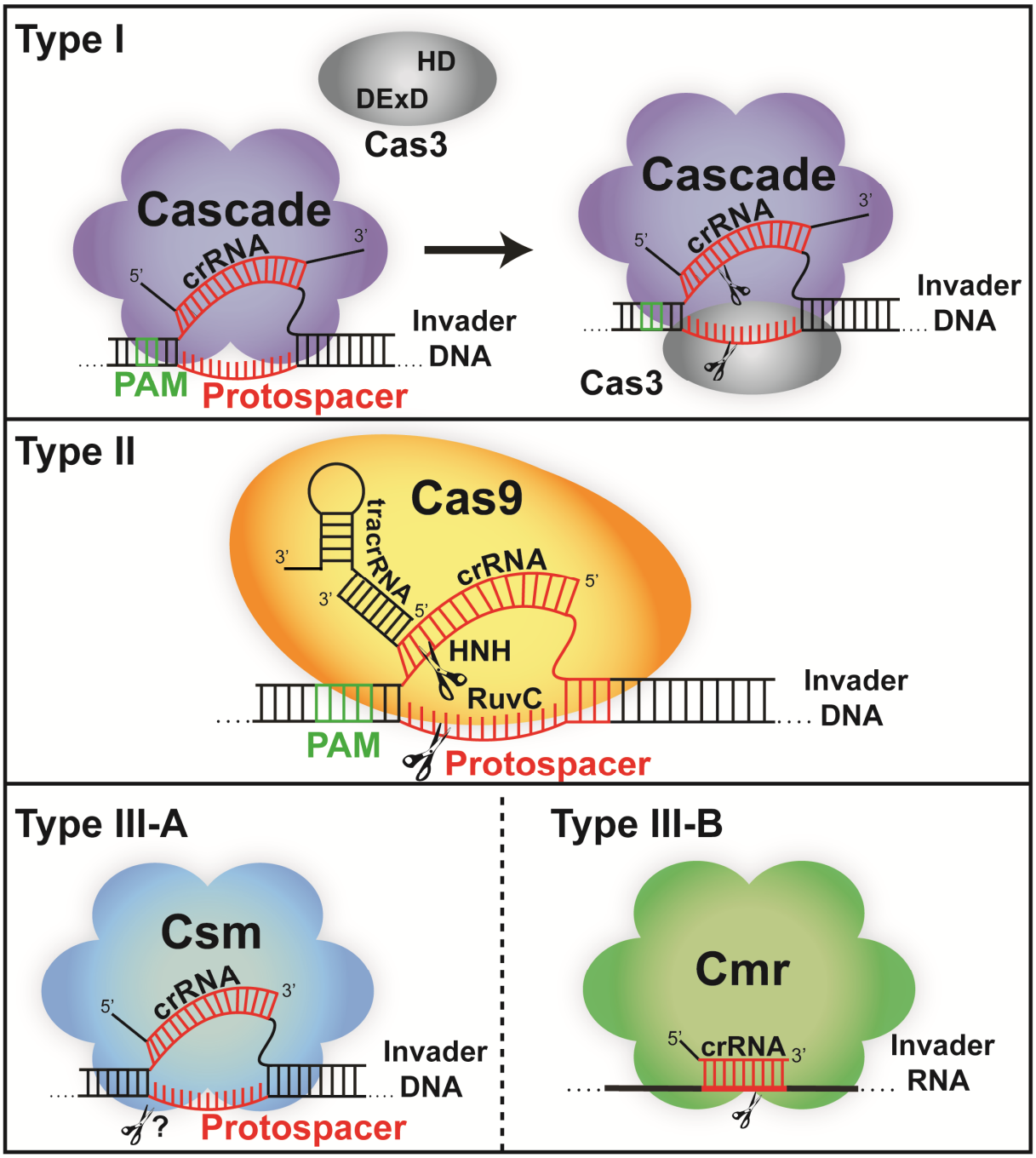

tracrRNA
cas genes
leader
5'
3'
tracrRNA
CRISPR primary transcript
RNAseIII
cleavage
pre-crRNA
GUUUUUGUAUCUCAAGAUUUAAGUAACUGUACAAC
UAGAAACAUG A GGUUCUAAAUUCAUUGUCA CA UUC
tracrRNA
1X
5' processing
mature
crRNA

Figure 1.5. Silencing complexes associated with each type of CC system. Type I CC modules are, or are hypothesized to be, characterized by a surveillance complex composed of sub-type specific proteins and a crRNA (Cascade) that detects an invader using the crRNA guide. Target-bound Cascade then recruits the Cas3 effector nuclease/helicase to cleave invader DNA (scissors). Invader silencing in Type II CC systems is carried out by the dual-RNA guided (tracrRNA and crRNA) Cas9 protein. The HNH and RuvC active sites each cleave one strand of invader DNA, producing a blunt double-stranded break. Type I and II CC systems rely on a PAM sequence within the invader for efficient silencing. Type III CC systems do not appear to require a PAM sequence and the identities of the effector nucleases are not yet known. Silencing complexes associated with Type III-A CC system have not been defined, however a putative Csm complex is predicted to cleave invader DNA. In Type III-B CC systems, a complex composed of Cmr-associated proteins, along with a crRNA guide, cleave single-stranded invader RNA.


Type I
HD
DExD
Cas3
Cascade
crRNA
5'
3'
Invader
DNA
PAM
Protospacer
Cascade
crRNA
3'
5'
Invader
DNA
Cas3
Type II
Cas9
tracrRNA
3'
3'
5'
crRNA
5'
HNH
RuvC
Invader
DNA
PAM
Protospacer
Type III-A
Csm
crRNA
5'
3'
Invader
DNA
?
Protospacer
Type III-B
Cmr
5'
crRNA
3'
Invader
RNA

CHAPTER 2

MULTIPLE CRISPR-CAS SYSTEMS FUNCTION INDEPENDENTLY IN *STREPTOCOCCUS THERMOPHILUS*[1]

---

[1]Carte, J, R. T. Christopher, J. T. Smith, S. Olson, R. Barrangou, S. Moineau, C. V. C. Glover III, B. Graveley, R. M. Terns, M. P. Terns. To be submitted to *RNA*.

Abstract

CRISPR-Cas systems are small RNA-based immune systems that protect prokaryotes from invaders such as viruses and plasmids. There are multiple CRISPR-Cas systems comprised of distinct sets of Cas proteins and CRISPR (cr)RNA species. We have investigated the biogenesis of crRNAs in *Streptococcus thermophilus* (*Sth*), which possesses 4 CRISPR-Cas system modules. crRNAs are processed from CRISPR locus transcripts, which contain multiple invader-targeting (or guide) sequences separated by a repeat sequence. Our results indicate that the crRNAs from each CRISPR locus are specifically processed by Cas proteins associated with the respective locus into divergent crRNA species. We find that the crRNAs from the Csm Type III-A system in *Sth* are initially processed by the Cas6 protein and ultimately trimmed to RNAs with an 8-nucleotide 5' CRISPR repeat sequence tag and ~35-nucleotide invader-targeting sequence. The Cse Type I-E system crRNAs are processed by Cse3 (Cas6e) and possess a 21-nucleotide 3' repeat tag as well as an 8-nucleotide 5' tag flanking the guide sequence. The crRNAs from the two Csn Type II-A systems present in *Sth* are similar overall, consisting of a 5'-truncated targeting sequence and a 3' tag, but interestingly, are distinct in the sizes of these 2 component elements. Moreover, we find that the Csn-1 (Cas9) protein associated with one Csn locus functions specifically in the production of crRNAs from that locus. Our findings indicate that multiple CRISPR-Cas systems can function independently in crRNA biogenesis in a given organism. This knowledge is important to understand and engineer co-existing CRISPR-Cas pathways.

Introduction

The CRISPR-Cas systems are adaptive RNA-directed immune systems present in many bacteria and archaea. Small, host-encoded CRISPR (cr)RNAs function along with CRISPR-associated (Cas) proteins to destroy invading nucleic acids. CRISPR loci are characterized by short direct repeats (24-37 bp in length) separated by similarly sized, invader-derived sequences (termed spacers) (Grissa et al., 2007). When transcribed and processed, these invader-derived elements guide effector Cas proteins to cleave foreign DNA or RNA (for reviews see (Barrangou and Horvath, 2012; Jore et al., 2012; Terns and Terns, 2011; Wiedenheft et al., 2012)).

There are multiple, diverse CRISPR-Cas systems defined by distinct sets of *cas* genes. Each of the 10 or so systems includes the universal *cas1* and *cas2* genes along with a module of subtype-specific *cas* genes named for an organism where the system is found (e.g. the cas subtype *E. coli* or *cse* genes) (Haft et al., 2005; Makarova et al., 2006; Makarova et al., 2011). Distant relationships have been identified among some components of different CRISPR-Cas systems allowing the definition of broad superfamilies of *cas* genes and the classification of the systems into three "Types": I - III (Makarova et al., 2011). The Cse system, for example, is the Type I-E system. The distinct CRISPR-Cas systems carry out essential steps in the defense pathway, including crRNA biogenesis, by different mechanisms.

CRISPR-Cas defense includes three mechanistic stages. In the adaptation stage, an invader-derived segment of DNA (termed a proto-spacer) is integrated into the microbe's CRISPR array. Evidence indicates that the common Cas1 and Cas2 proteins function in adaptation (Datsenko et al., 2012; Yosef et al., 2012). The biogenesis stage produces a panel of crRNAs that target the individual collected invader sequences. CRISPR locus transcripts are

processed to generate functional crRNAs that contain a common repeat sequence as well as invader-specific guide sequence (Brouns et al., 2008; Carte et al., 2008; Deltcheva et al., 2011; Hale et al., 2008; Hale et al., 2009; Hatoum-Aslan et al., 2011; Nam et al., 2012; Richter et al., 2012). In the interference stage, mature crRNAs guide effector Cas proteins to destroy invader DNA or RNA (Garneau et al., 2010; Gasiunas et al., 2012; Hale et al., 2009; Jinek et al., 2012; Magadan et al., 2012; Marraffini and Sontheimer, 2008; Sinkunas et al., 2013).

crRNA biogenesis occurs by diverse mechanisms in the distinct CRISPR-Cas systems. crRNAs are released from the CRISPR transcript by endonucleolytic cleavages within the conserved repeat elements, generating RNAs with repeat sequence flanking the guide sequence at both ends. In some cases, these 1X unit crRNAs are further processed. The initial repeat cleavage is catalyzed by "Cas6 superfamily" endoribonucleases in multiple CRISPR-Cas systems: the Cas6 protein in Csa/I-A, Cst/I-B, Csc/I-D, Csm/III-A, and Cmr/III-B systems, Cse3 (Cas6e) in the Cse/I-E system and Csy4 (Cas6f) in the Csy/I-F system (Brouns et al., 2008; Carte et al., 2008; Haurwitz et al., 2010; Lintner et al., 2011; Nam et al., 2012; Richter et al., 2012; Scholz et al., 2013). The 8-nucleotide repeat sequence tag generated at the 5' end of the crRNAs by the Cas6 superfamily endoribonucleases is retained in mature crRNA species (Brouns et al., 2008; Hale et al., 2008; Hatoum-Aslan et al., 2011; Lintner et al., 2011; Richter et al., 2012; Scholz et al., 2013) and has been shown to be critical for function (Hale et al., 2012). Cse3 (Cas6e) and Csy4 (Cas6f) recognize and cleave downstream of a stem-loop structure in associated CRISPR repeats, and that repeat stem-loop is retained at the 3' end of the Cse/I-E and Csy/I-F crRNAs (Brouns et al., 2008; Haurwitz et al., 2010; Nam et al., 2012; Richter et al., 2012; Sashital et al., 2011). The 3' ends of crRNAs generated by Cas6 (for Csa/I-A, Cst/I-B,

Csm/III-A and Cmr/III-B systems) are processed to remove some or all of the 3' repeat sequence, and sometimes a portion of the guide sequence, by unknown mechanisms (Hale et al., 2008; Hale et al., 2012; Hatoum-Aslan et al., 2011; Richter et al., 2012; Zhang et al., 2012). In another CRISPR-Cas system, Csd/I-C, CRISPR transcripts are processed by one of the Cas5 superfamily proteins, Cas5d, which generates an 11-nt 5' repeat tag (Garside et al., 2012; Nam et al., 2012). In Csn/II-A systems, base-pairing of a small non-coding RNA containing complementarity to the CRISPR repeat, termed the *trans*-activating crRNA or tracrRNA, triggers cleavage of CRISPR locus transcripts by RNase III (Deltcheva et al., 2011). The tracrRNA is encoded in the Csn/II-A CRISPR-Cas module. The 5' ends of Csn/II-A crRNAs undergo further processing that removes the 5' repeat fragment and ~10 nucleotides of the guide sequence (Deltcheva et al., 2011). The Csn/II-A crRNAs retain 19-22-nucleotide 3' repeat tag (Deltcheva et al., 2011).

Streptococcus thermophilus DGCC7710 (*Sth*) has four CRISPR-Cas modules, each containing a cluster of *cas* genes and a CRISPR array ((Horvath and Barrangou, 2010) and Figure 2.2.1)). Two of the CRISPR-Cas modules (1 and 3) are Csn/II-A systems and both of these are known to be active in both adaptation and interference in *Sth* (Barrangou et al., 2007; Deveau et al., 2010; Garneau et al., 2010; Levin et al., 2013; Magadan et al., 2012). The effector complex of the Csn system, comprised of Csn1 (Cas9) and associated RNAs, was recently successfully adapted as a promising tool for genome editing in human cells and other systems (Chang et al., 2013; Cho et al., 2013; Cong et al., 2013; Dicarlo et al., 2013; Hwang et al., 2013; Jiang et al., 2013; Jinek et al., 2013; Mali et al., 2013). *Sth* also contains a Csm/III-B and a Cse/I-E CRISPR-Cas system (modules 2 and 4, respectively). We refer to the *Sth* systems by

their 3-letter designations below: Csn, Csm, Cse. Here, we have investigated the composition and biogenesis of crRNAs from the four co-existing CRISPR-Cas systems in *Sth*.

Results

crRNAs associated with 4 CRISPR-Cas systems in *Sth*

To characterize the crRNAs generated by the CRISPR-Cas systems in *Sth*, we deep sequenced total *Sth* RNA and analyzed the RNAs derived from each of the CRISPR loci. Prior to construction of cDNA libraries, the RNAs were phosphatase and kinase treated to allow sequencing of RNAs containing diverse 5' and 3' chemical end groups (See Materials and Methods). We found that all four CRISPR arrays in the *Sth* genome are expressed and give rise to multiple small crRNAs that each contain an invader targeting sequence (Figure 2.2A). Probable transcription promoters are identifiable within the leader regions of CRISPRs 1, 2 and 3 in *Sth* (at approximately -35 and -10 relative to the 5' ends observed in sequenced RNAs from the region) (Figure 2.S1). CRISPR 4 may be transcribed as part of the upstream *cas* gene operon; the translational stop codon of the *cas2* gene is located just 20 nts upstream of the first repeat element of CRISPR locus 4 (Figure 2.S1). The greatest numbers of crRNAs were cloned from the loci with the strongest homology to the σ^{70} consensus sequence (2.2A and 2.S1).

The sequencing revealed that the crRNAs produced from each of the four CRISPR loci in *Sth* are distinct; the RNAs from each locus are comprised of a characteristic combination of invader-targeting sequence and repeat sequence tag elements. The distinctive patterns can be seen by mapping the ends of the sequenced RNAs from each locus relative to the repeat and guide element boundaries (Figure 2.2B).

The RNAs from the two Csn systems in *Sth* (CRISPRs 1 and 3) are similar in overall structure, but nonetheless distinct. The crRNAs from both of the Csn-associated loci contain substantial 3' tags comprised of the conserved repeat sequence (Figure 2.2B). However, the 3' repeat tags of the RNAs from the two loci are distinct in length as well as sequence. crRNAs from locus 1 primarily have a 17-nucleotide 3' repeat tag (guuuuuguacucucaag), and CRISPR 3 crRNAs include a 22-nucleotide 3' repeat tag that differs significantly in sequence beyond the first 5 nts (guuuuagagcuguguuguuucga). In addition, the Csn crRNAs from both loci only include ~2/3 of the guide sequence encoded in the genome. The crRNAs from loci 1 and 3 lack the first 9 and 10 nucleotides of the guide sequence, respectively (which is 30 nts in total average length for both locus 1 and 3). The most abundant RNAs detected from these CRISPR loci in Northern analysis (Figure 2.S2) correspond in size with the sequenced RNA species of 38 (21-nucleotide guide and 17-nucleotide tag) and 42 (20-nucleotide guide and 22-nucleotide tag) nucleotides respectively (Figure 2.2B). The overall features of the Csn crRNAs in *Sth* are similar to those described for *Streptococcus pyogenes* crRNAs, which consist of a 20-nucleotide guide sequence and 19-22 nucleotide 3' repeat tag (Deltcheva et al., 2011). Our results indicate that the intriguing absence of ~1/3 of the encoded guide sequence is a common feature of Csn system crRNAs. Notably, the characteristics of the crRNAs produced from the two Csn-associated CRISPRs within *Sth* are definably different.

The Csm-associated crRNAs in *Sth* are very different from the Csn crRNAs, but similar to those described for other CRISPR-Cas systems. The *Sth* Csm crRNAs include an 8-nucleotide 5' repeat tag upstream of the guide sequence (Figure 2.2B), as do crRNAs associated with Cmr, the other Type III system, and with most characterized Type I systems including Csa, Cst, Csc,

Cse and Csy (Brouns et al., 2008; Hale et al., 2008; Hale et al., 2009; Haurwitz et al., 2010; Lintner et al., 2011; Nam et al., 2012; Richter et al., 2012; Scholz et al., 2013; Zhang et al., 2012). In addition, like crRNAs associated with the Cmr effector complex (Hale et al., 2009), Csm crRNAs in *Sth* are a uniform length that is independent of the length of the guide region encoded in the CRISPR locus. The guide sequences of the three Csm crRNAs encoded in locus 2 are 36, 40 and 39 nucleotides in length, however, the guide sequences are trimmed to 35 nts in the crRNAs (Figure 2.S3). crRNAs of 43 nucleotides total length (8-nucleotide tag and 35-nucleotide guide) are the most abundant species of Csm crRNAs detected by RNA sequencing and Northern analysis (Figures 2.2B and 2.S2). (The levels of the crRNAs decline significantly with distance along the locus (Figure 2.S3), and crRNA 2.03 was not detected in Northern analysis (Figure 2.S2)) Similar processing has been reported for Csm crRNAs expressed in *Staphylococcus epidermidis* (Hatoum-Aslan et al., 2011). In that study of *S. epidermidis*, strains overexpressing crRNAs from a plasmid construct, an additional, much less abundant, 37-nucleotides species was also noted (Hatoum-Aslan et al., 2011), and a similar minor species can be observed in sequence reads and Northern analysis of *Sth* crRNA 2.01 (Figures 2.S2 and 2.S3). Each crRNA associated with the related Cmr system is found in two size forms, 39 and 45 nucleotides, that are often nearly equally abundant (Hale et al., 2009). Thus, the crRNAs associated with the two Type III CRISPR-Cas systems, Csm and Cmr, share common features including an 8-nucleotide 5' tag, RNA lengths independent of the encoded guide sequence length, and two size forms, suggesting similarities in the biogenesis pathways.

The crRNAs that we cloned from the locus associated with the Cse system in *Sth* closely resemble those characterized in *Escherichia coli* (Brouns et al., 2008; Jore et al., 2011). The

RNAs possess an 8-nucleotide repeat sequence tag upstream of the guide sequence and the remaining repeat sequence (21 nucleotides) at the 3' end (Figure 2.2B).

Together, our RNA sequencing and Northern findings indicate that the crRNAs that arise from each of the 4 CRISPR-Cas systems present in *Sth* have a characteristic composition that is different from that of the crRNAs from the other loci.

5' and 3' Termini

We also investigated the end groups of the crRNAs from the four *Sth* CRISPRs. To determine whether crRNAs possess a 5' phosphate end, total *Sth* RNA was treated with a 5' phosphate-dependent exonuclease (TEX) and the sensitivity of the crRNA to the exonuclease was assessed by Northern analysis (Figure 2.3A). Similarly, the presence of a 3' hydroxyl group was probed by treatment with poly(A) polymerase, which catalyzes the transfer of AMP to the 3' termini of RNAs that contain a 3' hydroxyl (Sippel, 1973), and Northern analysis (Figure 2.3B). Mature crRNA species from loci 1 and 3 are lost upon treatment with 5' phosphate-dependent exonuclease (Figure 2.3A), indicating that these crRNAs possess a phosphate group at the 5' end. The TEX-resistant crRNAs from loci 2 and 4 likely have 5' hydroxyl end groups. Treatment with poly(A) polymerase results in a reduction in the gel mobility of mature crRNAs from loci 1, 2 and 3 (Figure 2.3B), indicating the presence of a 3' hydroxyl end group. RNAs that are resistant to extension by poly(A) polymerase may have 3' phosphate or cyclic phosphate end groups. We found that pre-treatment with polynucleotide kinase (PNK), which removes 2′,3′ cyclic phosphate groups (Becker and Hurwitz, 1967), resulted in accessibility of the majority of the CRISPR 4 crRNAs to poly(A) polymerase (Figure 2.3B). Pre-treatment with calf intestinal

phosphatase (CIP), which removes 3' phosphates (but not 2',3' cyclic phosphates), allowed polyadenylation of a smaller fraction of the crRNAs (Figure 2.3B). The results suggest that crRNAs from locus 4 primarily possess cyclic phosphate groups at the 3' end. Similar results were described for Cse-derived crRNAs from *E. coli (Jore et al., 2011)*. The distinct end groups found on the crRNAs from each type of CRISPR-Cas system in *Sth* are indicated in Figure 2.3.

Mechanisms of crRNA biogenesis in *Sth*

Candidate biogenesis factors have been identified for each of the three types of CRISPR-Cas systems present in *Sth*. In the Csn-type system characterized in *S. pyogenes*, crRNA biogenesis was found to involve a short RNA species termed a tracrRNA that contains a region of complementarity to the CRISPR repeat sequence, and the non-Cas protein RNase III (Deltcheva et al., 2011). We observed expression of tracrRNAs associated with the two Csn systems in *Sth* in our deep RNA sequencing (Figure 2.4B and 4C, blue reads) and in Northern analysis (Figure 2.5E and 2.5F). The complementarity between the tracrRNA and the CRISPR repeat sequence can stimulate cleavage of the base-paired CRISPR transcript and tracrRNA by RNase III (Deltcheva et al., 2011). As illustrated in Figure 2.4A, tracrRNA-stimulated cleavages within the repeats of a CRISPR transcript generate 1X crRNA units with repeat sequence flanking the guide sequence at both ends (as well as cleaved tracrRNAs) (Deltcheva et al., 2011).

The 5' region of the 1X crRNAs - including the repeat sequence and a portion of the guide sequence - are not found in the mature Csn crRNAs from *Sth* loci 1 and 3 (Figures 2.2B, and 2.4B and 2.4C) or in *S. pyogenes* (Deltcheva et al., 2011), indicating that the 1X RNAs are processed at the 5' end. In addition however, our data suggest that in some cases Csn crRNAs can undergo further processing at the 3' end (and that the tracrRNA is also further processed at

the 5' end in parallel). On one hand, the most commonly observed ends of the tracrRNA and the crRNAs from *Sth* locus 3 are located at a potential RNase III cleavage site, 2 nucleotides apart within the region of tracrRNA-crRNA base-pairing (Figure 2.4C, green double arrow line). However, the observed ends of the most abundant tracrRNA and crRNA species from locus 1 are located at the edge of the region of contiguous base-pairing and are separated by 4 nts (Figure 2.4B, solid black lines), making it unlikely that these ends are generated by RNase III cleavage. We propose that RNase III cleavage of RNAs from *Sth* CRISPR 1 occurs at a site within the region of complementarity marked by less abundant crRNA and tracrRNA species with a 2-nucleotide end gap (Figure 2.4B, green double arrow line) and that the mature crRNAs are generated by 3' end processing (7 nucleotides; Figure 2.4B, solid black arrow line). The tracrRNA product of RNase III cleavage appears to undergo parallel 5' end processing (9 nucleotides; Figure 2.4B, solid black arrow line). The initial product of cleavage of the tracrRNA by RNase III is detectable by Northern analysis as well as deep sequencing (Figure 2.5E, middle band). Interestingly, there is also evidence of trimming of a fraction of the RNase III cleavage products from locus 3. CRISPR 3 crRNA species trimmed (~6-8 nucleotides) at the 3' end and tracrRNA species trimmed (~7 nucleotides) at the 5' end are also detectable by Northern and deep sequencing analysis (Figure 2.4C, dashed line and Figure 2.5F, bottom band).

In *S. pyogenes*, crRNA production also depends on Csn1 (Cas9) (Deltcheva et al., 2011). *Sth* has two Csn-type systems that each include a *csn1 (cas9)* and *csn2* gene (Figure 2.2.1). We examined the impact of loss of both Csn1 (Cas9) and Csn2 from CRISPR-Cas system 1 on crRNA biogenesis from all four CRISPR loci in *Sth* by Northern analysis of deletion strains. The profiles and levels of crRNAs from loci 2-4 were not altered in strains in which the *csn1*

(*cas9*) or *csn2* gene associated with CRISPR locus 1 was deleted (Figure 2.5, B-D). Furthermore, production of crRNAs from CRISPR 1 was not altered by deletion of the *csn2* gene (Figure 2.5A). However, deletion of the *csn1 (cas9)* gene did specifically affect biogenesis of crRNAs from locus 1. Neither mature crRNAs nor processed tracrRNAs are detected in the absence of Csn1 (Cas9) (Figure 2.5, A and E). Levels of an RNA corresponding to the 3' trimmed 1X crRNA are also significantly reduced in the absence of Csn1 (Cas9) (59 nucleotides; Figure 2.5A, double arrowhead), however, levels of the 1X crRNA are notably not reduced (66 nucleotides; Figure 2.5A, single arrowhead), suggesting that tracrRNA-stimulated production of 1X crRNAs does not require Csn1 (Cas9). (The tracrRNA product that accompanies 1X crRNA production is not observed (Figure 2.5E), which could simply reflect lower overall levels of both crRNAs and tracrRNAs and reduced traffic through the processing pathway.) Processing of the tracrRNA from CRISPR 3 is unaffected by the absence of Csn1 (Cas9) associated with CRISPR 1 (Figure 2.5F). Our results indicate that Csn1 (Cas9) functions specifically in the production or stability of 5' and/or 3' processed 1X crRNAs. Moreover, our findings reveal that the Csn1 (Cas9) protein from CRISPR-Cas system 1 functions specifically in the production of mature crRNAs from the associated CRISPR locus in *Sth*.

The Csm-type CRISPR-Cas system at locus 2 in *Sth* includes a gene encoding Cas6, an endoribonuclease that we previously found binds to and cleaves within the repeat regions of crRNA precursors associated with Csa/I-A, Cst/I-B and Cmr/III-B systems in *Pyrococcus furiosus* (Carte et al., 2010; Carte et al., 2008). Cas6 cleaves 8 nucleotides upstream of the guide region, generating the 5' repeat sequence tag that is important for crRNA function with the Cmr complex in *P. furiosus* (Hale et al., 2012). Deletion of the *cas6* gene results in loss of all species

of the Csm crRNAs in *S. epidermidis* (Hatoum-Aslan et al., 2011). To determine whether Cas6 is involved in the biogenesis of crRNAs in *Sth*, we tested the ability of the *Sth* protein to bind and cleave repeat RNAs from each of the four *Sth* CRISPR loci. In gel shift analysis, *Sth* Cas6 demonstrated significant binding to the locus 2 crRNA repeat relative to repeat sequences from the other three loci (Figure 2.6A). In addition, Cas6 specifically cleaved RNA containing the locus 2 repeat sequence (generating a product of ~32 nucleotides, consistent with cleavage approximately 8 nucleotides upstream of the guide sequence) (Figure 2.6B).

The previously determined structure of *P. furiosus* Cas6 (Carte et al., 2008) revealed a cluster of three conserved amino acids positioned similarly to a catalytic triad found in archaeal tRNA splicing endonuclease, which we found is critical for CRISPR repeat RNA cleavage, but not binding, in *P. furiosus* (Carte et al., 2010). We generated a mutation in one of the equivalent conserved amino acids in *Sth* Cas6 (H39A) and tested the activity of the protein with CRISPR 2 repeat RNAs. The mutation did not appreciably affect the affinity of the protein for CRISPR 2 repeat RNA in gel shift assays, however cleavage activity was significantly reduced (Figure 2.6C and 6D). Our findings indicate that Cas6 specifically binds and cleaves RNAs from the Csm-associated CRISPR locus, and not from the other CRISPR loci in *Sth*.

The Cse3 (Cas6e) protein that is part of Cse-type CRISPR-Cas systems like that found at locus 4 in *Sth* is essential for crRNA processing and has been found to bind and cleave crRNA repeat sequences in organisms with a lone Cse system (Brouns et al., 2008; Gesner et al., 2011). We tested the activity of *Sth* Cse3 (Cas6e) with RNAs containing repeat sequences from each of the 4 CRISPRs in *Sth*. Cse3 (Cas6e) specifically bound and cleaved RNAs containing the repeat sequence from the Cse locus, generating a product of ~21 nucleotides, but not the other three loci

(Figure 2.7A and 2.7B). Mutation in a conserved residue of *E. coli* Cse3 (Cas6e) (H20A) was previously found to disrupt crRNA biogenesis (Brouns et al., 2008)). Mutation of an equivalent residue in *Sth* Cse3 (Cas6e) (H20A) prevents cleavage (Figure 2.7D), but gel shift analysis indicates that it also disrupts interaction of Cse3 (Cas6e) with the CRISPR repeat RNA (Figure 2.7C). Our findings confirm that Cse3 (Cas6e) is the crRNA endoribonuclease of the Cse CRISPR-Cas system, but indicate that conserved residue H20 functions in RNA binding, making it more difficult to ascribe a role for this amino acid in catalysis based on existing information.

Disscussion

CRISPR-Cas systems were identified as CRISPR arrays and associated modules of Cas protein-coding genes that segregate between genomes as units, independently of other CRISPR-Cas systems. While the systems are capable of functioning independently, it is not known whether the systems share some components or processes (or may interfere with one another) when present in a common environment. Some organisms possess more than one CRISPR-Cas system (and introduction of CRISPR-Cas systems into organisms with existing CRISPR-Cas systems may be desirable for strain engineering), however the potential interactions between co-existing systems are not known. Here we have examined 4 co-existing crRNA biogenesis pathways: Cse/I-E and Csm/III-A pathways that utilize Cas6 superfamily proteins, and two side-by-side Csn/II-A pathways. Our findings indicate that not only the Cse and Csm systems, but also the two Csn systems present in *Sth*, function independently in crRNA biogenesis.

The Cas6 superfamily proteins found in the Csm/III-A and Cse/I-E systems – Cas6 and Cse3 (Cas6e) – specifically recognize and cleave only RNAs from the directly associated

CRISPR locus (Figures 2.6 and 2.7). In *Synechocystis*, deletions of Cas6 proteins from Csc/I-D and Csm or Cmr/III systems have various effects on production of crRNAs from directly associated CRISPRs consistent with independently functioning pathways, though the specificity of the effects on the associated CRISPR was not determined (Scholz et al., 2013). As summarized in Figure 2.8, cleavage by *Sth* Cas6 or Cse3 (Cas6e) generates an 8-nucleotide repeat tag sequence and 5' hydroxyl group that is retained in both the Csm and Cse crRNAs. Mature Cse crRNAs are not further processed and retain a 21-nucleotide repeat tag and a 2'-3' cyclic phosphate at the 3' end. On the other hand, the products of the Csm system Cas6 cleavage are subject to 3' end processing by an unknown mechanism that generates RNAs of a fixed length that have a 3' hydroxyl group and lack 3' repeat sequence.

Similarly, the Csn1 (Cas9) protein from locus 1 is required only for the production of Csn crRNAs from the directly associated CRISPR (Figure 2.5). In addition, the two Csn/II-A systems produce distinct crRNAs (Figure 2.2B). The crRNAs produced from the two Csn systems include ~2/3 of the encoded guide sequence and a 3' repeat tag sequence (see Figure 2.8). While very similar in architecture, the major species of crRNAs from the two loci have guides sequences of slightly different lengths (20 and 21 nucleotides) and distinct 3' tags. Processing at the 5' end of the 1X intermediate RNAs generated by tracrRNA-stimulated RNase III cleavage removes the 5' repeat sequence and 10 or 11 nucleotides of the guide sequence (from CRISPR 1 and CRISPR 3 RNAs, respectively). The initial RNase III cleavage of CRISPR 1 RNAs proposed here (Figure 2.4B) would generate crRNAs with a 24-nucleotide 3' repeat sequence, similar to the 22-nucleotide 3' tag found on the crRNAs from CRISPR 3. However, the primary CRISPR 1 species found in *Sth* has a shorter 17-nucleotide repeat tag. Our

observations of RNAs from both CRISPRs 1 and 3 indicate that Csn crRNAs may be subject to 3' end processing as well as 5' end processing (Figure 2.4). tracrRNAs appear to undergo parallel 5' end processing (Figure 2.4). It is not yet known whether the 3'-processed Csn crRNAs (and 5'-processed tracrRNAs) are functional species. The major Csn crRNA species retain 5' monophosphate and 3' hydroxyl end groups (Figure 2.3). Csn1 (Cas9) is the effector nuclease of the Csn/II-A system, cleaving DNA targets recognized by the crRNAs, and has also recently been co-opted as a tool for genome engineering and gene knockdown (Chang et al., 2013; Cho et al., 2013; Cong et al., 2013; Dicarlo et al., 2013; Gasiunas et al., 2012; Hwang et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Jinek et al., 2013; Mali et al., 2013; Qi et al., 2013). Csn1 (Cas9) DNA cleavage activity requires both a mature crRNA and tracrRNA (Gasiunas et al., 2012; Jinek et al., 2012; Karvelis et al., 2013) and understanding the precise nature of the functional forms of these RNAs will aid in optimization of Csn1 (Cas9) activity for gene engineering applications.

Our findings indicate that Csn1 (Cas9) may also play a role in the 5' or 3' processing of Csn crRNAs. In the absence of Csn1 (Cas9), RNase III production of the 1X intermediate does not appear to be reduced, but mature crRNA and tracrRNA species are lost (Figure 2.5A and 5E). Csn1 (Cas9) has two nuclease active sites (RuvC and HNH type) (Makarova et al., 2011) that are both involved in silencing, each cleaving one strand of the DNA target (Gasiunas et al., 2012; Jinek et al., 2012). One of these sites (or a yet unidentified site) may be involved in crRNA processing, however, the role of Csn1 (Cas9) in crRNA biogenesis may also be indirect (e.g. protection of the regions found in the mature crRNA and also tracrRNA). The Csn1 (Cas9) protein associated with CRISPR 1 is essential for production or accumulation of mature CRISPR

1 RNAs but not crRNAs from the other loci (Figure 2.5). The molecular basis for the interaction between Csn1 (Cas9) and crRNAs – which is likely essential for both crRNA processing and crRNA-guided DNA cleavage – is not known. Our results indicate that Csn1 (Cas9) discriminates between RNAs from the four CRISPRs in *Sth*.

Our finding that crRNA biogenesis occurs independently to generate distinct crRNA species in *Sth* suggests that these systems also likely function independently in silencing and adaptation, and other findings support this paradigm. The two Csn/II-A systems require different PAM sequences adjacent to target sequences for both adaptation of the target sequence into CRISPRs and silencing (Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008; Magadan et al., 2012). In addition, a recent study has identified a distinct PAM-like requirement for silencing by the Cse/I-E system in *Sth (Sinkunas et al., 2013)*. We predict that the distinct crRNAs produced from each locus function exclusively with their corresponding silencing complexes in this organism.

Our understanding of the co-function of specific CRISPR-Cas systems provides insight into the mechanisms of prokaryotic defense and is also important for predictable application of CRISPR-Cas technologies. Our findings indicate that Csn/II-A, Cse/I-E and Csm/III-A CRISPR-Cas systems - and even two Csn/II-A systems - can function independently within one environment to produce distinct crRNA species. However, it is likely that CRISPR-Cas systems will share components (e.g. crRNAs) or processes (e.g. crRNA biogenesis) under some conditions, and additional studies will be needed to determine the factors that restrict or allow cross-talk between systems.

Materials and Methods

RNA isolation and RNA library preparation

S. thermophilus DGCC7710 (*Sth*) cells were grown to mid-log phase in M17 media (Oxoid) supplemented with 0.5% lactose (LM17). The *csn1-1* (*cas9*) and *csn2-1* gene disruption strains were generated in a previous study (Barrangou et al., 2007). The cells were harvested by centrifugation at 10,000 x g for 10 minutes and lysed by bead-beating with a Mini-Beadbeater (Biospec Products). Trizol LS (Invitrogen) was used to extract total RNA. RNA libraries were produced as described previously (Hale et al., 2009) except that no size-selection was done prior to RNA manipulation. Briefly, 10 μg of *Sth* total RNA was treated with Thermosensitive Alkaline Phosphatase (TSAP) (Promega) followed by 3' ligation to a 5' adenylated adaptor (See Table 2.S1 for sequences) with T4 RNA ligase 2, truncated (NEB). The 3' ligated RNAs were gel purified away from free adaptor and treated with T4 polynucleotide kinase (Ambion) before 5' ligation to an additional adaptor with T4 ssRNA Ligase 1 (NEB). The 5' and 3' ligated RNAs were reverse transcribed with SuperScript II reverse transcriptase (Invitrogen), digested with RNase H (Promega), and PCR amplified with Crimson Taq DNA polymerase (NEB).

Chemical end group analysis of crRNAs

Chemical end groups of crRNAs from *Sth* were analyzed by enzymatic treatment followed by Northern analysis. To map 5' chemical end groups, reactions containing 10 μg of *Sth* total RNA were treated with Terminator 5' phosphate-dependent exonuclease (Epicentre) according to the manufacturer's protocol. The reactions were terminated by phenol/chloroform/isoamyl alcohol (PCI) extraction followed by ethanol precipitation. Half of each reaction was separated on 7M

urea TBE 15% polyacrylamide gels for Northern analysis, probing for the first spacer sequence from each CRISPR locus (Probe sequences are shown in Table 2.S1). Northern blotting was performed as described previously (Hale et al., 2009).

For 3' chemical end group analysis of crRNAs from CRISPR loci 1-3, 100 μg of *Sth* total RNA was separated on a 7M urea TBE 15% acrylamide gel and RNAs between ~30 and ~65 nucleotides were gel purified as described previously (Hale et al., 2009). Gel-purified RNAs were treated with *E. coli* poly(A) polymerase (NEB) in reactions containing 10% of the gel purified RNA, 1X poly(A) reaction buffer, and 1 mM ATP. Reactions were incubated at 37º C for 20 minutes followed by PCI extraction and ethanol precipitation. Northern analysis was performed with half of each reaction. Chemical end groups of crRNAs from CRISPR locus 4 were analyzed using combinations of poly(A) polymerase, TSAP, calf intestinal phosphatase (Promega) and T4 polynucleotide kinase. *Sth* total RNA (10 μg) was treated with either TSAP or T4 polynucleotide kinase, or untreated. Reactions were stopped by PCI extraction and ethanol precipitation. RNAs were then treated with poly(A) polymerase as described above and Northern analysis was performed following PCI extraction and ethanol precipitation.

Cloning and mutagenesis of *Sth cas6* and *cse3 (cas6e)*

Genomic DNA was isolated from *Sth* as described previously (Hill et al., 1991) with the minor modifications. The cells were not treated with proteinase K and the DNA was precipitated with isopropanol/ammonium acetate prior to phenol/chloroform extraction. DNA primers specific to *Sth cas6* and *cse3* (*cas6e*) genes (listed in Table 2.S1) were designed and ordered from Eurofins MWG Operon. The primers were used to amplify *cas6* and *cse3 (cas6e)* from *Sth* genomic DNA

using the Expand High Fidelity PCR system (Roche) according to the manufacturer's protocol. The genes were cloned into the pET24d plasmid using the In Fusion PCR cloning system (Clontech) according to the manufacturer's protocol. Colonies were screened by PCR and nucleotide sequences of positive clones were confirmed by DNA sequencing. These constructs were used to generate mutant *cas6* and *cse3 (cas6e)* constructs using specific primers (listed in Table 2.S1) and the QuikChange site-directed mutagenesis kit (Stratagene). The nucleotide sequences mutant constructs were confirmed by DNA sequencing.

Production of recombinant *Sth* Cas6 and Cse3 (Cas6e) proteins

Sth Cas6 and Cse3 (Cas6e) (both 6x N- terminally histidine tagged) were expressed and purified from *E. coli* as described previously (Hale et al., 2009) with the following modifications. The cells (from 100 mL cultures) expressing *Sth* Cas6 were resuspended in 20 mM sodium phosphate (pH 7.6, buffer A), 0.1 mM phenylmethysulfonyl fluoride (PMSF), and 5 mM imidazole. Cells expressing *Sth* Cse3 (Cas6e) were resuspended in buffer A supplemented with 50 mM NaCl and 20 mM β-mercaptoethanol. The cells were lysed by sonication and centrifuged at 14,000 rpm for 10 minutes. Proteins were isolated from the cleared lysate by batch purification using 50 μL Ni-NTA agarose beads (Qiagen). Proteins were bound by end-over-end rotation for one hour at room temperature. The beads were washed three times with buffer A and once with buffer A supplemented with 20 mM imidazole. Proteins were eluted from the beads in buffer A plus 300 mM imidazole. Elutions containing *Sth* Cas6 were dialyzed into 50 mM Tris-HCl (pH 7.0) using Slide-a-lyzer MINI dialysis cassettes (ThermoScientific). Elutions containing *Sth* Cse3 (Cas6e)

were dialyzed into the same buffer supplemented with 50 mM NaCl and 20 mM β-mercaptoethanol. Glycerol was added to 50% and the proteins were stored at -20º C until use.

RNA binding and cleavage reactions

Radiolabeled crRNA repeat substrates were generated by *in vitro* transcription as described previously ((Carte et al., 2008), See Table 2.S1 for oligo sequences). RNA binding and cleavage assays were carried out as described previously (Carte et al., 2008) with the following modifications. *Sth* Cas6 was incubated with 5,000 cpm of uniformly labeled RNA substrate in 25 mM Tris-HCl (pH 7.0), 0.75 mM DTT, 1.5 mM $MgCl_2$, 5 μg *E. coli* tRNA, and 10% glycerol. The reaction conditions for *Sth* Cse3 (Cas6e) were identical except that DTT was replaced with 5 mM β-mercaptoethanol and NaCl was added to 50 mM. Reactions were incubated at 37º C for 30 minutes prior to electrophoretic separation on both native TBE 8% polyacrylamide (RNA binding) and 7M urea TBE 15% acrylamide (RNA cleavage) gels. The gels were dried and RNAs detected by phosphor imaging (GE Life Sciences).

Acknowledgements

We would like to thank members of the Terns lab for helpful discussions, as well as Josiane Garneau and Anne Summers for technical assistance. This study was supported by funding from Dupont Nutrition and Health and NIH grant RO1 GM099876 to M.T. and R.T.

References

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

Barrangou, R., and P. Horvath. 2012. CRISPR: new horizons in phage resistance and strain identification. *Annual Review of Food Science and Technology*. 3:143-162.

Becker, A., and J. Hurwitz. 1967. The enzymatic cleavage of phosphate termini from polynucleotides. *The Journal of Biological Chemistry*. 242:936-950.

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

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

Carte, J., R. Wang, H. Li, R.M. Terns, and M.P. Terns. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & Development*. 22:3489-3496.

Chang, N., C. Sun, L. Gao, D. Zhu, X. Xu, X. Zhu, J.W. Xiong, and J.J. Xi. 2013. Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. *Cell Research*. 23:465-472.

Cho, S.W., S. Kim, J.M. Kim, and J.S. Kim. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology*.

Cong, L., F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, and F. Zhang. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*.

Datsenko, K.A., K. Pougach, A. Tikhonov, B.L. Wanner, K. Severinov, and E. Semenova. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*. 3:945.

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

Deveau, H., R. Barrangou, J.E. Garneau, J. Labonte, C. Fremaux, P. Boyaval, D.A. Romero, P. Horvath, and S. Moineau. 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of Bacteriology*. 190:1390-1400.

Deveau, H., J.E. Garneau, and S. Moineau. 2010. CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of Microbiology*. 64:475-493.

Dicarlo, J.E., J.E. Norville, P. Mali, X. Rios, J. Aach, and G.M. Church. 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Research*.

Garneau, J.E., M.E. Dupuis, M. Villion, D.A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A.H. Magadan, and S. Moineau. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 468:67-71.

Garside, E.L., M.J. Schellenberg, E.M. Gesner, J.B. Bonanno, J.M. Sauder, S.K. Burley, S.C. Almo, G. Mehta, and A.M. MacMillan. 2012. Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*. 18:2020-2028.

Gasiunas, G., R. Barrangou, P. Horvath, and V. Siksnys. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*.

Gesner, E.M., M.J. Schellenberg, E.L. Garside, M.M. George, and A.M. Macmillan. 2011. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nature Structural & Molecular Biology*. 18:688-692.

Gordon, J.J., M.W. Towsey, J.M. Hogan, S.A. Mathews, and P. Timms. 2006. Improved prediction of bacterial transcription start sites. *Bioinformatics*. 22:142-148.

Grissa, I., G. Vergnaud, and C. Pourcel. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*. 35:W52-57.

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

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

Hale, C., S. Majumdar, J. Elmore, N. Pfister, M. Compton, S. Olson, A.M. Resch, C.V. Glover, 3rd, B.R. Graveley, R.M. Terns, and M.P. Terns. 2012. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell*. 45:292-302.

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

Hatoum-Aslan, A., I. Maniv, and L.A. Marraffini. 2011. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proceedings of the National Academy of Sciences of the United States of America*. 108:21218-21222.

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

Hill, C., I.J. Massey, and T.R. Klaenhammer. 1991. Rapid method to characterize lactococcal bacteriophage genomes. *Applied and Environmental Microbiology*. 57:283-288.

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

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

Hwang, W.Y., Y. Fu, D. Reyon, M.L. Maeder, S.Q. Tsai, J.D. Sander, R.T. Peterson, J.R. Yeh, and J.K. Joung. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology*.

Jiang, W., D. Bikard, D. Cox, F. Zhang, and L.A. Marraffini. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology*.

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

Jinek, M., A. East, A. Cheng, S. Lin, E. Ma, and J. Doudna. 2013. RNA-programmed genome editing in human cells. *eLife*. 2:e00471.

Jore, M.M., S.J. Brouns, and J. van der Oost. 2012. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harbor Perspectives in Biology*. 4.

Jore, M.M., M. Lundgren, E. van Duijn, J.B. Bultema, E.R. Westra, S.P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M.R. Beijer, A. Barendregt, K. Zhou, A.P. Snijders, M.J. Dickman, J.A. Doudna, E.J. Boekema, A.J. Heck, J. van der Oost, and S.J. Brouns. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular biology*. 18:529-536.

Karvelis, T., G. Gasiunas, A. Miksys, R. Barrangou, P. Horvath, and V. Siksnys. 2013. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. *RNA Biology*. 10.

Levin, B.R., S. Moineau, M. Bushman, and R. Barrangou. 2013. The Population and Evolutionary Dynamics of Phage and Bacteria with CRISPR-Mediated Immunity. *PLoS Genetics*. 9:e1003312.

Lintner, N.G., M. Kerou, S.K. Brumfield, S. Graham, H. Liu, J.H. Naismith, M. Sdano, N. Peng, Q. She, V. Copie, M.J. Young, M.F. White, and C.M. Lawrence. 2011. Structural and functional characterization of an archaeal clustered regularly interspaced short

palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *The Journal of Biological Chemistry*. 286:21643-21656.

Magadan, A.H., M.E. Dupuis, M. Villion, and S. Moineau. 2012. Cleavage of Phage DNA by the *Streptococcus thermophilus* CRISPR3-Cas System. *PloS One*. 7:e40913.

Makarova, K.S., N.V. Grishin, S.A. Shabalina, Y.I. Wolf, and E.V. Koonin. 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. *Biology Direct*. 1:7.

Makarova, K.S., D.H. Haft, R. Barrangou, S.J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F.J. Mojica, Y.I. Wolf, A.F. Yakunin, J. van der Oost, and E.V. Koonin. 2011. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology*. 9:467-477.

Mali, P., L. Yang, K.M. Esvelt, J. Aach, M. Guell, J.E. Dicarlo, J.E. Norville, and G.M. Church. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science*.

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

Nam, K.H., C. Haitjema, X. Liu, F. Ding, H. Wang, M.P. Delisa, and A. Ke. 2012. Cas5d Protein Processes Pre-crRNA and Assembles into a Cascade-like Interference Complex in Subtype I-C/Dvulg CRISPR-Cas System. *Structure*. 20:1574-1584.

Qi, L.S., M.H. Larson, L.A. Gilbert, J.A. Doudna, J.S. Weissman, A.P. Arkin, and W.A. Lim. 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*. 152:1173-1183.

Richter, H., J. Zoephel, J. Schermuly, D. Maticzka, R. Backofen, and L. Randau. 2012. Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucleic Acids Research*.

Sashital, D.G., M. Jinek, and J.A. Doudna. 2011. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nature Structural & Molecular Biology*. 18:680-687.

Scholz, I., S.J. Lange, S. Hein, W.R. Hess, and R. Backofen. 2013. CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein. *PloS One*. 8:e56470.

Sinkunas, T., G. Gasiunas, S.P. Waghmare, M.J. Dickman, R. Barrangou, P. Horvath, and V. Siksnys. 2013. *In vitro* reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *The EMBO Journal*.

Sippel, A.E. 1973. Purification and characterization of adenosine triphosphate: ribonucleic acid adenyltransferase from *Escherichia coli*. *European Journal of Biochemistry / FEBS*. 37:31-40.

Terns, M.P., and R.M. Terns. 2011. CRISPR-based adaptive immune systems. *Current Opinion in Microbiology*. 14:321-327.

Wiedenheft, B., S.H. Sternberg, and J.A. Doudna. 2012. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*. 482:331-338.

Yosef, I., M.G. Goren, and U. Qimron. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*. 40:5569-5576.

Zhang, J., C. Rouillon, M. Kerou, J. Reeks, K. Brugger, S. Graham, J. Reimann, G. Cannone, H. Liu, S.V. Albers, J.H. Naismith, L. Spagnolo, and M.F. White. 2012. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular Cell*. 45:303-313.

Figure 2.1. Overview of the CRISPR-Cas loci found in *S. thermophilus* DGCC7710 (*Sth*). CRISPR arrays are composed of short direct repeats (black boxes) interspaced with unique invader-derived spacer sequences (colored boxes). The number of spacers is indicated for each locus. Transcription of CRISPR arrays is initiated within upstream leader sequences ("L"). Each CRISPR array contains an adjacent subset of *cas* genes (shown as colored boxes). Genes involved, or predicted to be involved in crRNA biogenesis are indicated with an asterisk. Note that *cas2-4* is a *cas2-dnaQ* fusion (Horvath and Barrangou, 2010). Diagram is not drawn to scale.

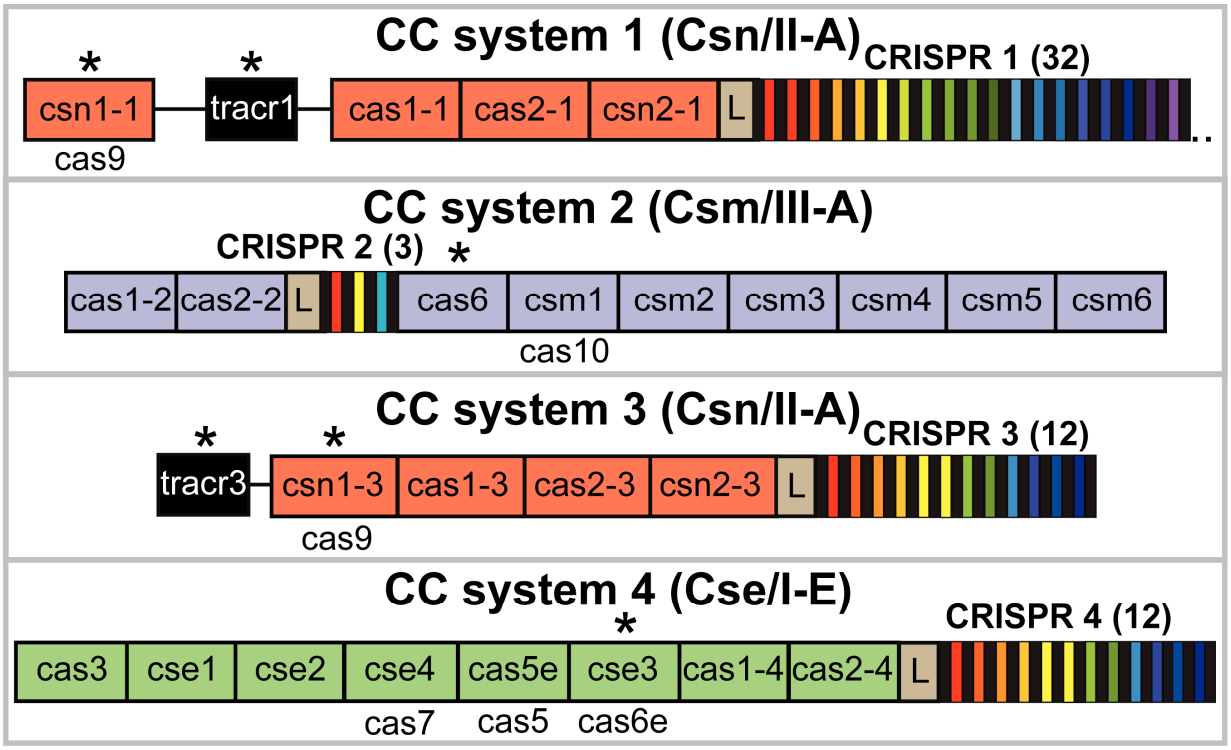

CC system 1 (Csn/II-A)
CRISPR 1 (32)
csn1-1
tracr1
cas1-1
cas2-1
csn2-1
L
cas9
CC system 2 (Csm/III-A)
CRISPR 2 (3)
cas1-2
cas2-2
L
cas6
csm1
csm2
csm3
csm4
csm5
csm6
cas10
CC system 3 (Csn/II-A)
CRISPR 3 (12)
tracr3
csn1-3
cas1-3
cas2-3
csn2-3
L
cas9
CC system 4 (Cse/I-E)
CRISPR 4 (12)
cas3
cse1
cse2
cse4
cas5e
cse3
cas1-4
cas2-4
L
cas7
cas5
cas6e

Figure 2.2. RNA deep sequencing profiles of crRNAs from *Sth*. (A) Deep sequencing reads from *Sth* total RNA map to each of the four CRISPR loci. The number of unique reads for a given nucleotide position are indicated on the Y axis in thousands. CRISPR repeats are shown as black boxes below each graph. For CRISPR3, spacers 4 and 11, as well spacers 5 and 12 are identical, and thus do not map uniquely to the genome. In both cases the total numbers of reads have been divided equally between each position (shown as dotted boxes). (B) Mapping crRNA 5' and 3' termini. The graphs show the percentages of sequenced RNAs from *Sth* total RNA that map to the indicated position relative to the repeat-guide junction. Below each graph is an illustration of the mature crRNAs from each CRISPR locus.

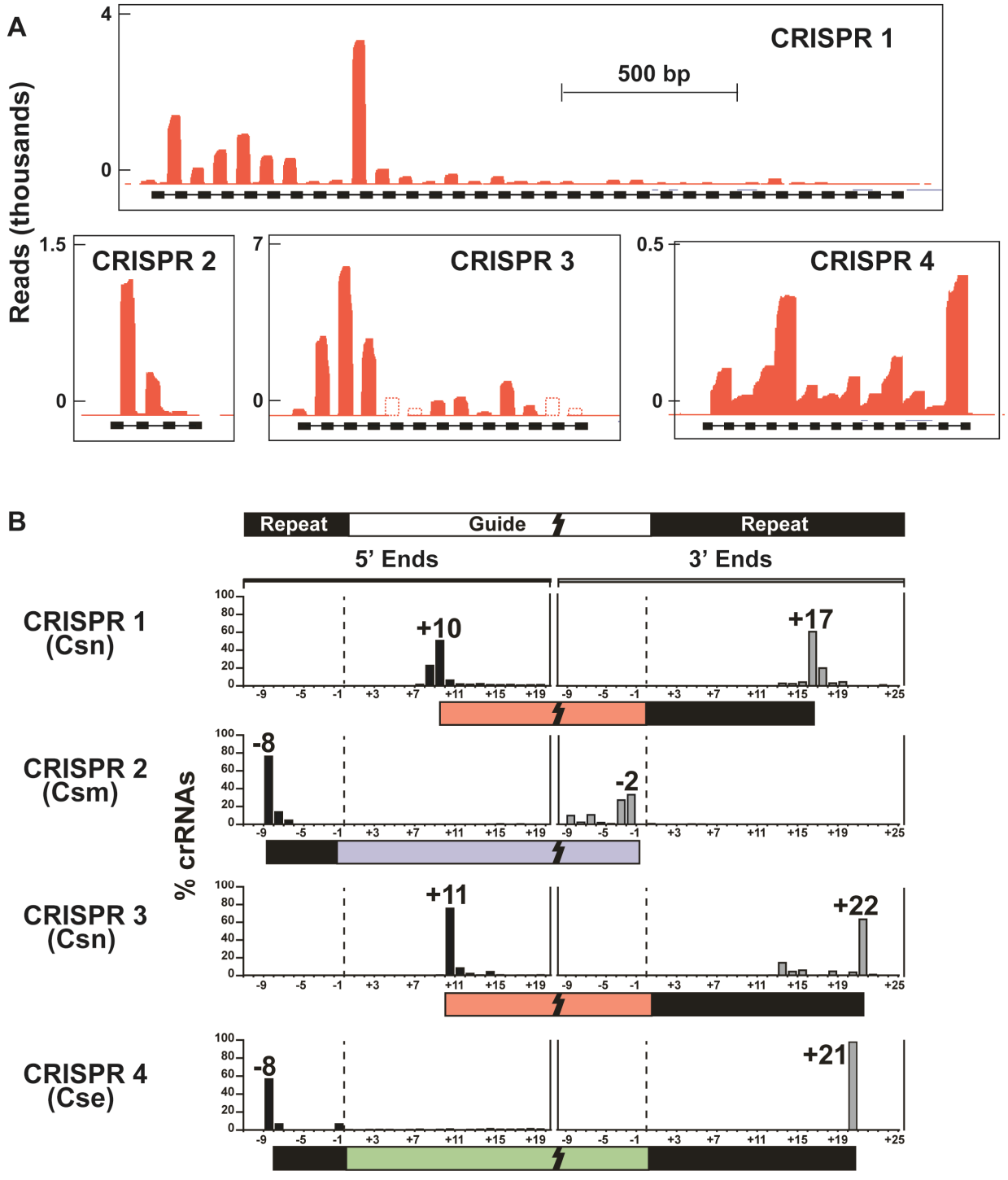

A
Reads (thousands)
CRISPR 1
500 bp
CRISPR 2
CRISPR 3
CRISPR 4
B
Repeat
Guide
Repeat
5' Ends
3' Ends
% crRNAs
CRISPR 1 (Csn)
CRISPR 2 (Csm)
CRISPR 3 (Csn)
CRISPR 4 (Cse)
+10
+17
-8
-2
+11
+22
-8
+21

Figure 2.3. Mapping *Sth* crRNA 5' and 3' chemical end groups. (A) 5' end group analysis was performed by incubating total RNA from *Sth* in the absence or presence of Terminator 5'-Phosphate-Dependent Exonuclease (TEX) followed by Northern blotting using probes indicated below each panel where 1.01 is leader proximal crRNA guide sequence, etc. (B) For 3' chemical end group analysis of crRNAs from CRISPRs 1-3, gel extracted sRNAs from *Sth* were incubated in the absence or presence of *E. coli* poly(A) polymerase (PAP) followed by Northern blotting. The chemical end groups present on crRNAs from CRISPR4 were determined by Northern analysis of *Sth* total RNA following combinations of treatments with Thermosensitive Alkaline Phosphatase (TSAP), T4 polynucleotide kinase (PNK), and *E. coli* poly(A) polymerase (PAP). Mature crRNAs are indicated by an asterisk.

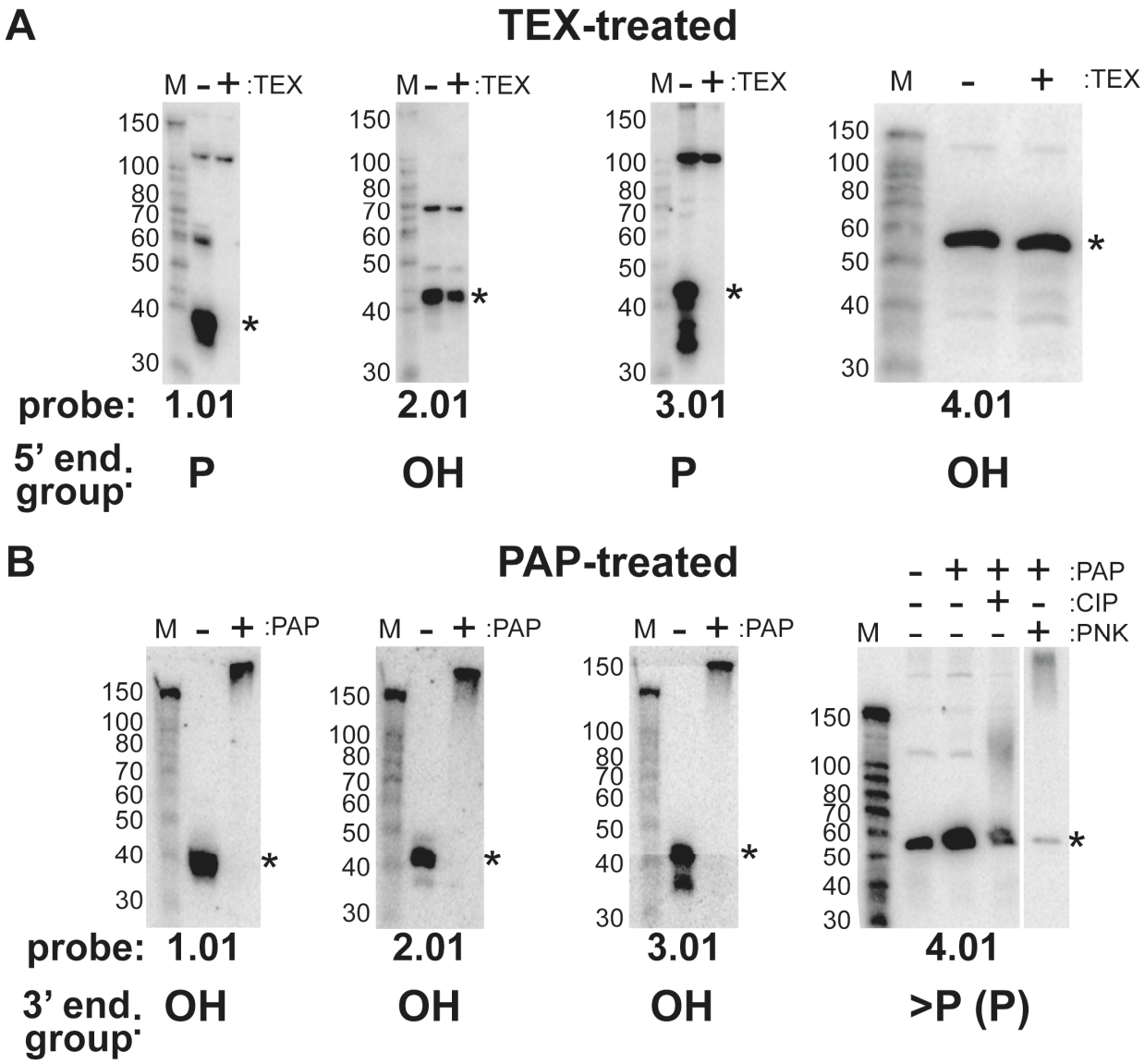

A
TEX-treated
M - + :TEX
150
100
80
70
60
50
40
30
*
probe: 1.01
2.01
3.01
4.01
5' end. group: P
OH
P
OH
B
PAP-treated
M - + :PAP
probe: 1.01
2.01
3.01
4.01
3' end. group: OH
OH
OH
>P (P)
- + + + :PAP
- - + - :CIP
- - - + :PNK

Figure 2.4. crRNA biogenesis from *Sth* CRISPRs 1 and 3. (A) Model depicting crRNA biogenesis pathway. A duplex composed of a tracrRNA and the CRISPR primary transcript is cleaved by RNase III to generate a 1X intermediate composed of a single invader-targeting sequence and fragments of the CRISPR repeat at either end. The 1X intermediate is then cleaved (by an unknown mechanism) to produce mature crRNAs containing ~20 nucleotides of guide sequence and ~20 nucleotides of CRISPR repeat sequence at the 3' end. (B) and (C) Mapping RNase III cleavage sites within tracrRNA:CRISPR repeat duplex. Deep sequencing profiles of crRNAs 1.01 and 3.01 were aligned with the profiles of their respective tracrRNAs and were used to predict the position of cleavage by RNase III (indicated by arrows). The positions of the proposed RNase III cleavage sites shown in Figure 4B are hypothesized based on the ends of low abundance RNAs (theoretical intermediates) that can be observed in the deep sequencing.)

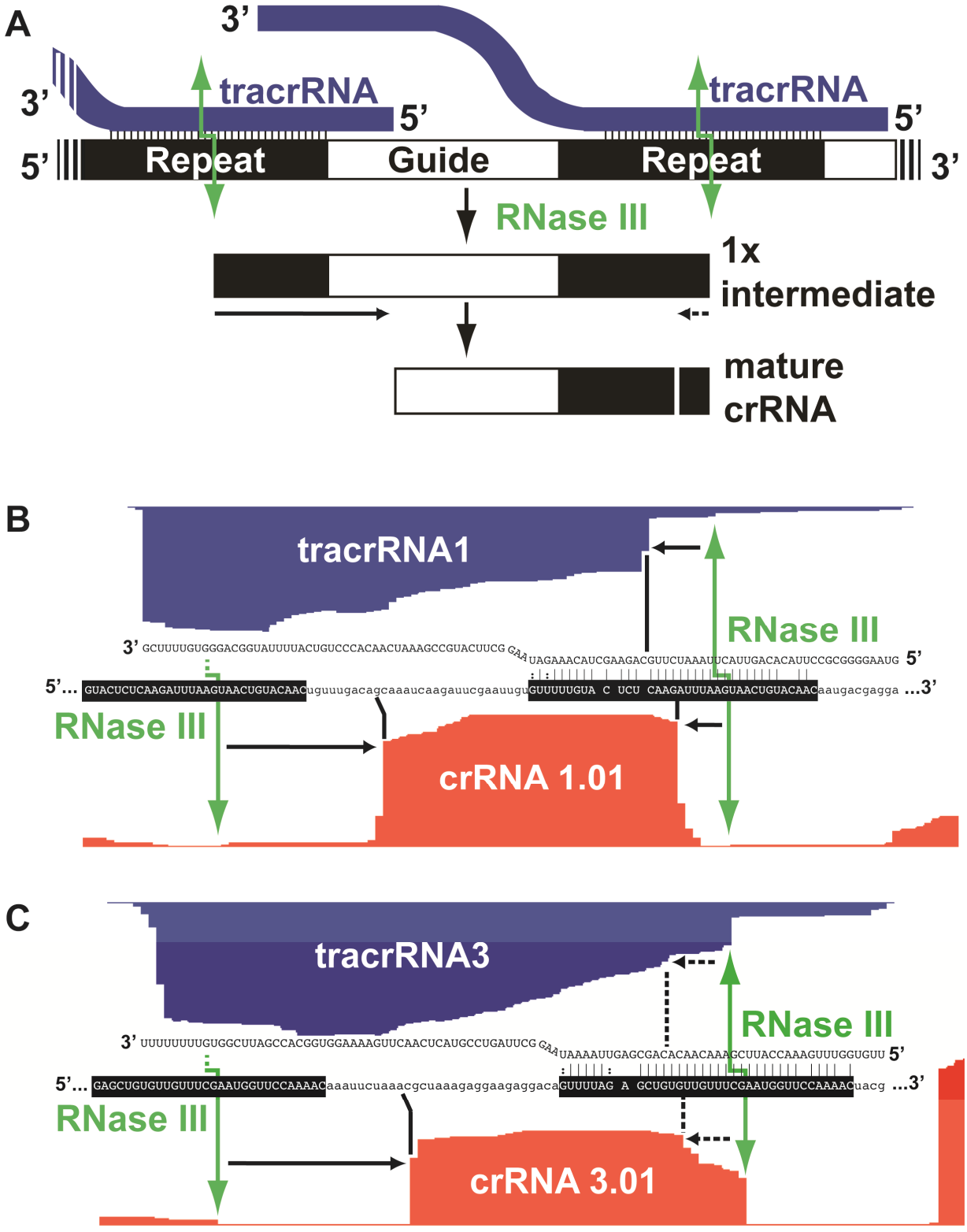

A
3'
tracrRNA
5'
3'
5'
Repeat
Guide
Repeat
3'
tracrRNA
5'
RNase III
1x
intermediate
mature
crRNA
B
tracrRNA1
RNase III
3' GCUUUUGUGGGACGGUAUUUUACUGUCCCACAACUAAAGCCGUACUUCG GAA UAGAAACAUCGAAGACGUUCUAAAAUUCAUUGACACAUUCCGCGGGGAAUG 5'
5'... GUACUCUCAAGAUUUAAGUAACUGUACAAC uguuugacagcaaaucaagauucgaauugu GUUUUUGUA C UCU CAAGAUUUAAGUAACUGUACAAC aaugacgagga ...3'
RNase III
crRNA 1.01
C
tracrRNA3
RNase III
3' UUUUUUUUGUGGCUUAGCCACGGUGGAAAAGUUCAACUCAUGCCUGAUUCG GAA UAAAAUUGAGCGACACAACAAAGCUUACCAAAGUUUGGUGUU 5'
5'... GAGCUGUGUUGUUUCGAAUGGUUCCAAAAC aaauucuaaacgcuaaagaggaagaggaca GUUUUAG A GCUGUGUUGUUUCGAAUGGUUCCAAAAC uacg ...3'
RNase III
crRNA 3.01

Figure 2.5. Csn1-1 (Cas9) is required for the accumulation of mature crRNAs as well as processed forms of tracrRNAs from CRISPR 1. Northern analysis was performed using total RNA extracted from wild type *Sth* and strains in which *csn1-1* (*cas9*) or *csn2-1* were inactivated (probes used indicated below each panel). For crRNA 1.01, the positions of the 1X intermediate and the 3' trimmed 1X intermediate are indicated by an arrow and a double arrow, respectively. Diagrams illustrating predicted tracrRNA primary transcripts and processing intermediates, as well as the position of the probes used are shown to the right of each blot.

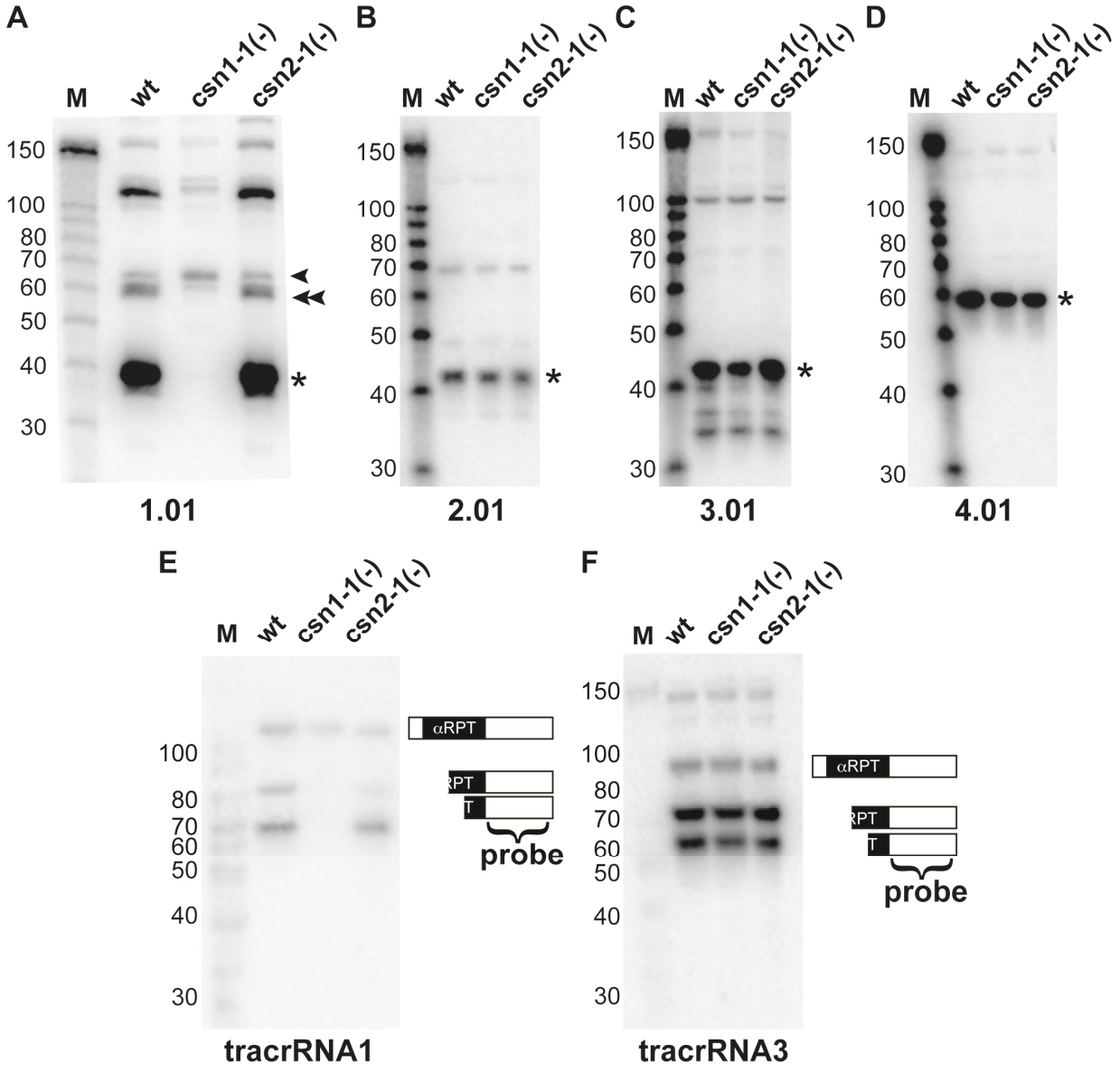

A
B
C
D
M
wt
csn1-1(-)
csn2-1(-)
150
100
80
70
60
50
40
30
*
1.01
2.01
3.01
4.01
E
F
αRPT
probe
tracrRNA1
tracrRNA3

Figure 2.6. Cas6 selectively binds and cleaves *Sth* CRISPR repeat 2 RNA. (A) and (B) ^{32}P-labeled RNAs composed of each of the four CRISPR repeats were incubated in the absence or presence of purified *Sth* Cas6 (0.2 μM). RNA binding was assessed by loading half of each reaction on a native 8% polyacrylamide gel (A) and RNA cleavage was evaluated by separation of the RNAs on a 15% denaturing (7 M urea containing) polyacrylamide gel (B). (C) and (D) His39 of Cas6 plays a critical role in catalysis but not RNA binding. ^{32}P-labeled CRISPR repeat 2 RNA was incubated with increasing concentrations (indicated in μM) of purified wild type or H39A mutant Cas6 protein. RNA binding and cleavage were assessed by native gel mobility shift (C) and denaturing PAGE (D) as in (A) and (B). Specific RNA-protein interactions (A and C) or cleavage products (B and D) are indicated by asterisks.

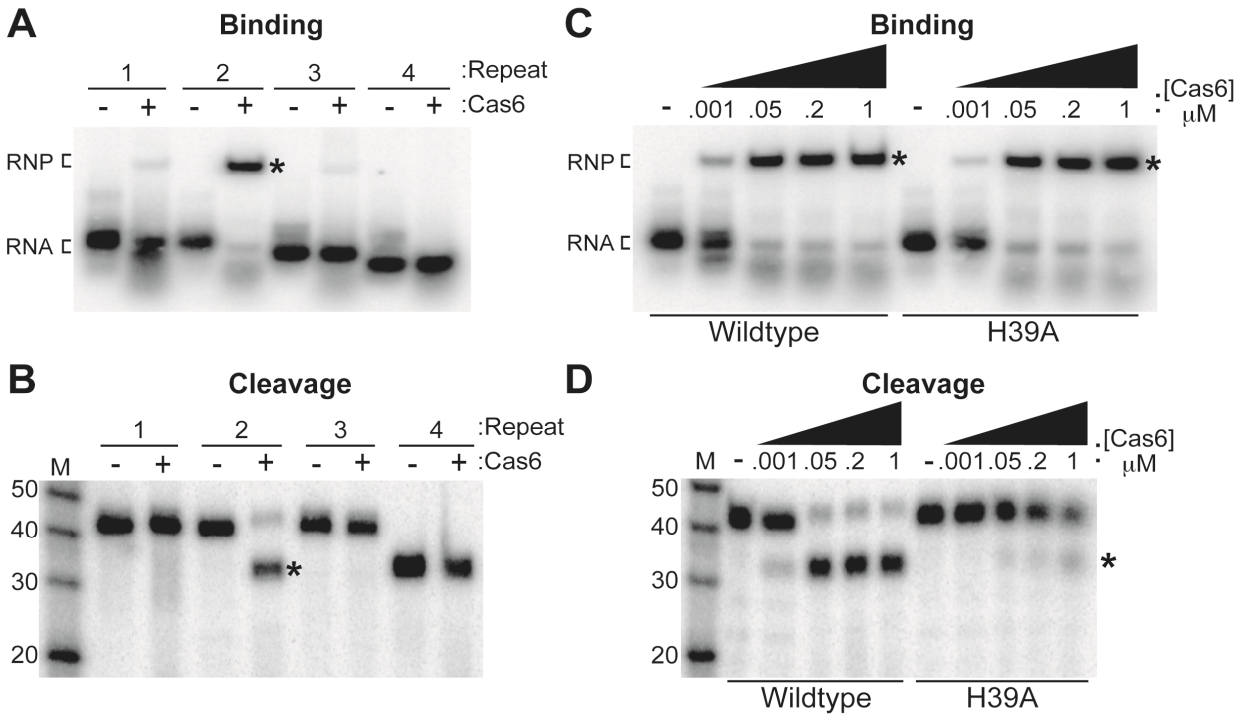

A
Binding
1
2
3
4
:Repeat
- + - + - + - +
:Cas6
RNP
RNA
*
B
Cleavage
1
2
3
4
:Repeat
M
- + - + - + - +
:Cas6
50
40
30
20
*
C
Binding
- .001 .05 .2 1 - .001 .05 .2 1
[Cas6] µM
RNP
RNA
*
*
Wildtype
H39A
D
Cleavage
M - .001 .05 .2 1 - .001 .05 .2 1
[Cas6] µM
50
40
30
20
*
Wildtype
H39A

Figure 2.7. Cse3 (Cas6e) selectively binds and cleaves *Sth* CRISPR repeat 4 RNA. (A) and (B) ^{32}P-labeled RNAs composed of each of the four CRISPR repeats were incubated in the absence or presence of purified *Sth* Cse3 (Cas6e) from CRISPR 2 (1 μM). RNA binding was assessed by loading half of each reaction on a native 8% polyacrylamide gel (A) and RNA cleavage was evaluated by separation of the RNAs on a 15% denaturing (7 M urea containing) polyacrylamide gel (B). (C) and (D) His20 of Cse3 (Cas6e) is indispensable for function. ^{32}P-labeled CRISPR repeat 4 RNA was incubated with increasing concentrations (indicated in μM) of purified wild type or H20A mutant Cse3 (Cas6e) protein. RNA binding and cleavage were assessed by native gel mobility shift (C) and denaturing PAGE (D) as in panels (A) and (B). Specific RNA-protein interactions (A and C) or cleavage products (B and D) are indicated by asterisks.

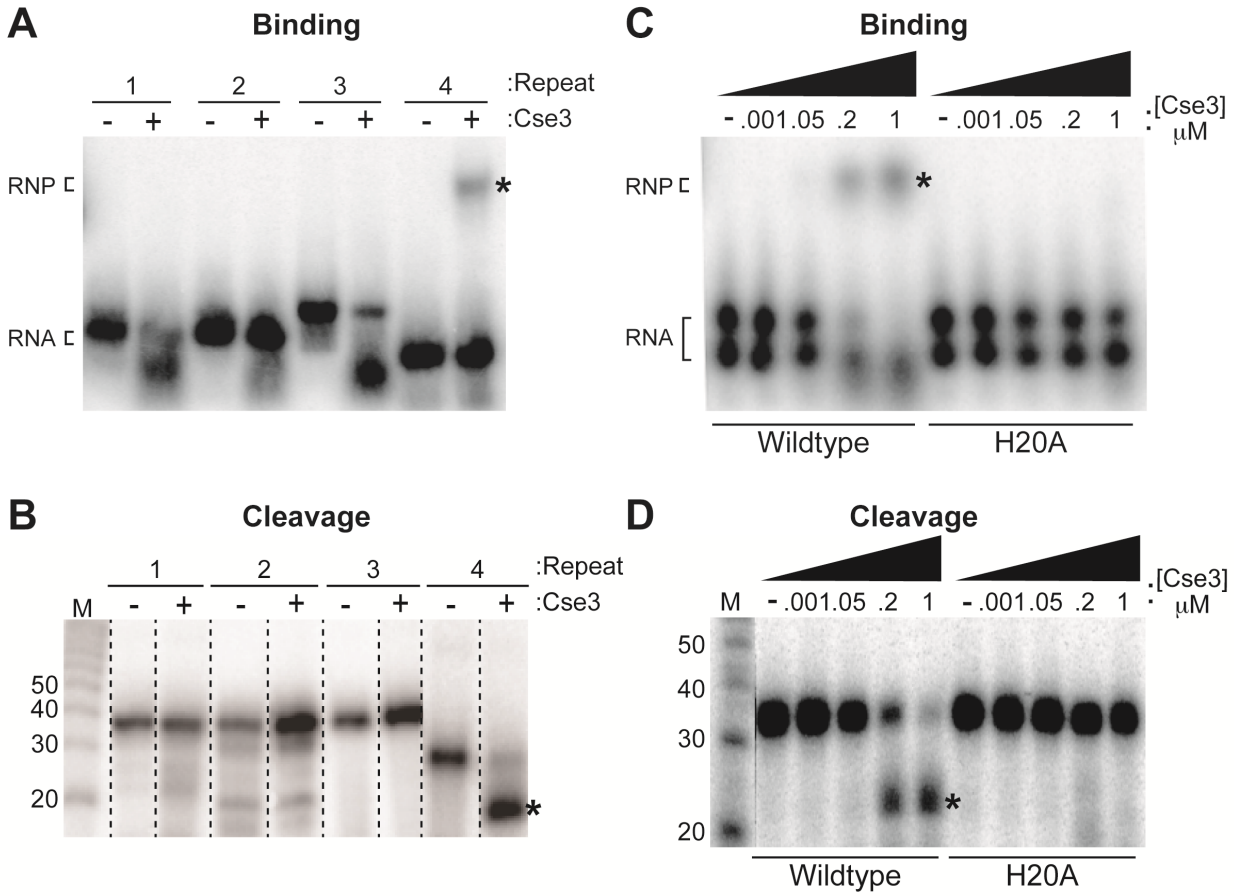

A
Binding
1
2
3
4
:Repeat
- + - + - + - +
:Cse3
RNP
RNA
*
B
Cleavage
1
2
3
4
:Repeat
M
- + - + - + - +
:Cse3
50
40
30
20
*
C
Binding
- .001 .05 .2 1 - .001 .05 .2 1
:[Cse3] μM
RNP
RNA
*
Wildtype
H20A
D
Cleavage
M
- .001 .05 .2 1 - .001 .05 .2 1
:[Cse3] μM
50
40
30
20
*
Wildtype
H20A

Figure 2.8. Proposed models for crRNA biogenesis pathways in *Sth*. Factors involved, or predicted to be involved in primary processing and maturation are indicated. The 5' and 3' chemical end groups present on RNA processing intermediates and mature crRNAs are also indicated.

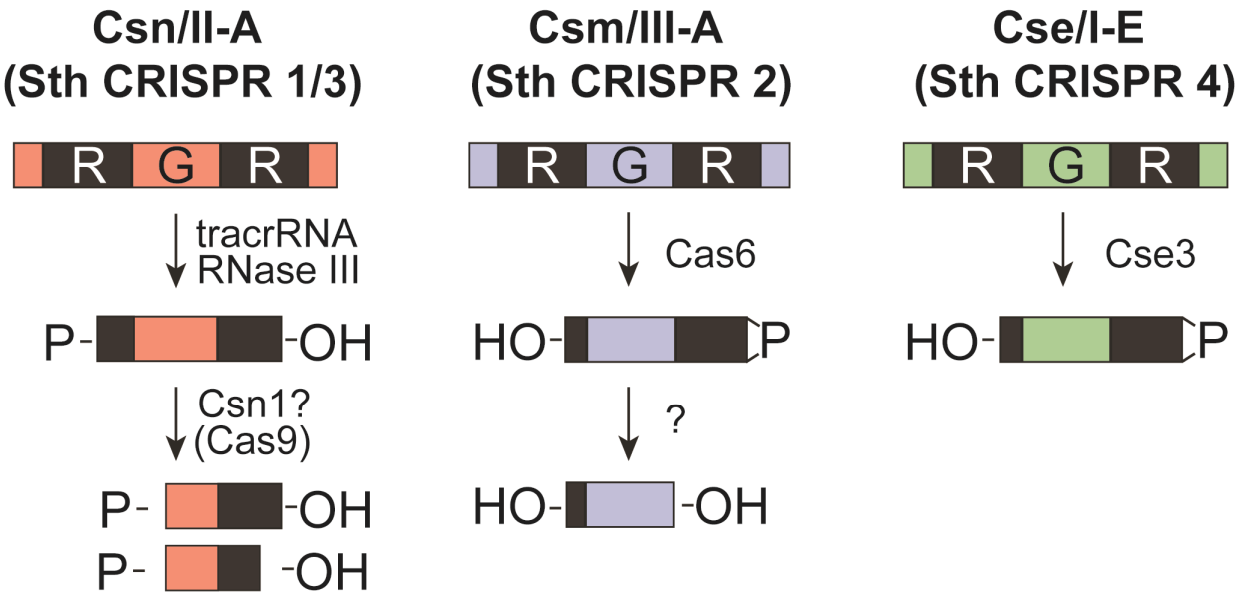

Csn/II-A
(Sth CRISPR 1/3)
R G R
tracrRNA
RNase III
P- -OH
Csn1?
(Cas9)
P- -OH
P- -OH
Csm/III-A
(Sth CRISPR 2)
R G R
Cas6
HO- P
?
HO- -OH
Cse/I-E
(Sth CRISPR 4)
R G R
Cse3
HO- P

Table 2.S1. Oligonucleotides used in this study

	Sequence (5-3')
RNA cloning	
5' Adapter	invdTGTTCrArGrArGrUrUrCrUrArGrUrCrCrGrArCrGrArUrC
3' Adapter	rAppTCGTATGCCGTCTTCTGCTTGTddC
RT primer	CAAGCAGAAGACGGCATACGA
PCR primer	AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA
Northern probes	
guide 1.01	ACAATTCGAATCTTGATTTGCTGTCAAACA
guide 1.03	GATAACAGTGGTCAGCAGATTGTCAAATCG
Repeat 1	GTTGTACAGTTACTTAAATCTTGAGAGTACAAAAAC
guide 2.01	TTTCTAGGAATGGGTAATTATAGCGAGCTAGAAAGC
guide 2.03	TTAATTGAATTATGAAATACAAGGTAATAACATATTTTGA
Repeat 2	GTTTCCGTCCCCTCTCGAGGTAATTAGGTTTATATC
guide 3.01	TGTCCTCTTCCTCTTTAGCGTTTAGAATTT
guide 3.03	ACATTATTAATAAAATGTTTACAAGAGGAA
Repeat 4	AGGATCACCCCCGCGTGTGCGGGAAAAAC
guide 4.01	GCTTATGCGTGGGAGGCCATTGATATAGGTAT
guide 4.03	TAGTACTGTTCCCATGTCTAAGGAGGGGTTGC
Repeat 4	AGGATCACCCCCGCGTGTGCGGGAAAAAC
tracrRNA1	CCCTGCCATAAAATGACAGGGTGTTG
tracrRNA3	CCGAATCGGTGCCACCTTTTCAAGTTG
PCR primers	
cas6 forward	TCACCATCACGGATCCATGAAAAAATTAGTATTTACTTTTAAAAGGATC
cas6 reverse	TGCTCGAGTGCGGCCGCTCAATCTTTTCTTTCTTCAAGCTTTATCC
cas6 H39A forward	GATAGTGACTATGTTGATTATCTGGCTCAGCAGCAAACAAATCCCTATG
cas6 H39A reverse	CATAGGGATTTGTTTGCTGCTGAGCCAGATAATCAACATAGTCACTATC
cse3 forward	TCACCATCACGGATCCATGTACATTTCTAGGGTAGAAATTGATCG
cse3 reverse	TGCTCGAGTGCGGCCGCTCATTCCTCAAGTGGTACCACTGTC
cse3 H20A forward	AGAGATTTAACGCATGTAGGAGCTTATGCTGCTTGGGTGGAAG
cse3 H20A reverse	CTTCCACCCAAGCAGCATAAGCTCCTACATGCGTTAAATCTCT
IVT Templates- for annealing	
Repeat 1 5'	TAATACGACTCACTATAGGGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Repeat 1 3'	GTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCCCTATAGTGAGTCGTATTA
Repeat 2 5'	TAATACGACTCACTATAGGGGATATAAACCTAATTACCTCGAGAGGGGACGGAAAC
Repeat 2 3'	GTTTCCGTCCCCTCTCGAGGTAATTAGGTTTATATCCCCTATAGTGAGTCGTATTA
Repeat 3 5'	TAATACGACTCACTATAGGGGTTTTAGAGCTGTGTTGTTTCGAATGGTTCCAAAAC
Repeat 3 3'	GTTTTGGAACCATTCGAAACAACACAGCTCTAAAACCCCTATAGTGAGTCGTATTA
Repeat 4 5'	TAATACGACTCACTATAGGGGTTTTTCCCGCACACGCGGGGGTGATCCT
Repeat 4 3'	AGGATCACCCCCGCGTGTGCGGGAAAAACCCCTATAGTGAGTCGTATTA

Figure 2.S1. Putative promoter sequences found in the *Sth* CRISPR leader sequences and upstream of tracrRNAs. Plus signs indicate the approximate transcript abundance levels from deep sequencing. Matches to σ^{70} consensus promoter elements are observed at approximately -35 and -10 bases (shown in blue) upstream of the observed start sites (+1, first base underlined) (Gordon et al., 2006). The first CRISPR repeat is shown in red. There were no obvious promoter sequence elements found in the region upstream of CRISPR4. The stop codon for *cas2-4* is highlighted.

Sth CRISPR Transcription Start Sites

Sth CRISPR1 (+++)

Start site: TTGACA<--14/20 bases-->TATAAT +1

TTTTAGAAAGTAAGGATTGACAAGGACAGTTATTGATTTTATAATCACTATGTGGGTATAAAAACGTCAA
AATTTCATTTGAGGTTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACTGTTTGACAGCAAATCAAGAT
TCGAATTGTGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACAATGACGAGGAGCTATTGGCACAAC
TTACAGTTTTT

Sth CRISPR2 (++)

Start site: TTGACA<--14/20 bases-->TATAAT +1

CATAGTGCTGAAAAATTAGATGACATTCTCTTCTTCTAAGCCTTTATAGACCTTTAATCATATGGTACAC
TATAGATAGTGTTTCCAGTAGGTCCTACATCTTGTGCCTCTAGCAACTGCCTAGAGCACAAGATATGGGG
ATATAAACCTAATTACCTCGAGAGGGGACGGAAACGCTTTCTAGCTCGCTATAATTACCCATTCCTAGAA
AGATATAAACCTAATTACCTCGAGAGGGGACGGAAACTCAAAATATGTTATTACCTTGTAT

Sth CRISPR3 (+++)

Start site: TTGACA<--14/20 bases-->TATAAT +1

ACATCTTGTCTAAAACGTTGATATATAAGGATTTTTAAGGTATAATAAATATAAAATTGGAATTATTTTGAAGCTGA
AGTCATGCTGAGATTAATAGTGCGATTACGAAATCTGGTAGAAAAGATATCCTACGAGGTTTTAGAGCTGTGTTGTT
TCGAATGGTTCCAAAACAAATTCTAAACGCTAAAGAGGAAGAGGACAGTTTTAGAGCTGTGTTGTTTCGAATGGTTC
CAAAACTACTGCTGTATTA

Sth CRISPR4 (+)

Start site: None obvious

ATCGAGGATTCAGTTCCTCATAGAGCTTTGAAAGATGCTAGACTAATCTATCATTTATCAACTAAAGTAAATAAATT
CTTAGCTAGAATGAAAGAAAAGTCTTAATTCCATTGGGATCTTTTAGTGTTTTTCCCGCACACGCGGGGGTGATCCT
ATACCTATATCAATGGCCTCCCACGCATAAGCGTTTTTCCCGCACACGCGGGGGTGATCCCGACACCACTAGGGCGA
GTCTGAGCGCCCCCAGGTT

Sth CRISPR1 tracrRNA (sense strand):

Start site (-35/-10): TTGACA<--14/20 bases-->TATAAT +1

TTTTAATATTAATTGTTAGAAAGTGTTGCAATTATAGTTATCATATACTATAATAATAGTGTAAGGGGCGCCTTACA
CAGTTACTTAAATCTTGCAGAAG
CTACAAAGATAAGGCTTCATGCCGAAATCAACACCCTGTCATTTTATGGCAGGGTGTTTTCGTTATTTAAAGAGGAG
AAGAAATGACTT

Sth CRISPR3 tracrRNA (sense strand):

Start site (-35/-10): TTGACA<--14/20 bases-->TATAAT +1

TTTTTTATACTTTACCCTATAGTAAATTGACATATTAGTTATAAAACTATATAATAATAATTGTGGTTTGAAACCAT
TCGAAACAACACAGCGAGTTAAA
ATAAGGCTTAGTCCGTACTCAACTTGAAAAGGTGGCACCGATTCGGTGTTTTTTTTATACACAAAGAGTCCCTTGGA
ATCATCCAAAGGA

Figure 2.S2. Detection of crRNAs in *Sth*. Northern blots were performed using *Sth* total RNA. The probes used are indicated below each blot and the mature crRNA species, if detected, are indicated with an asterisk.

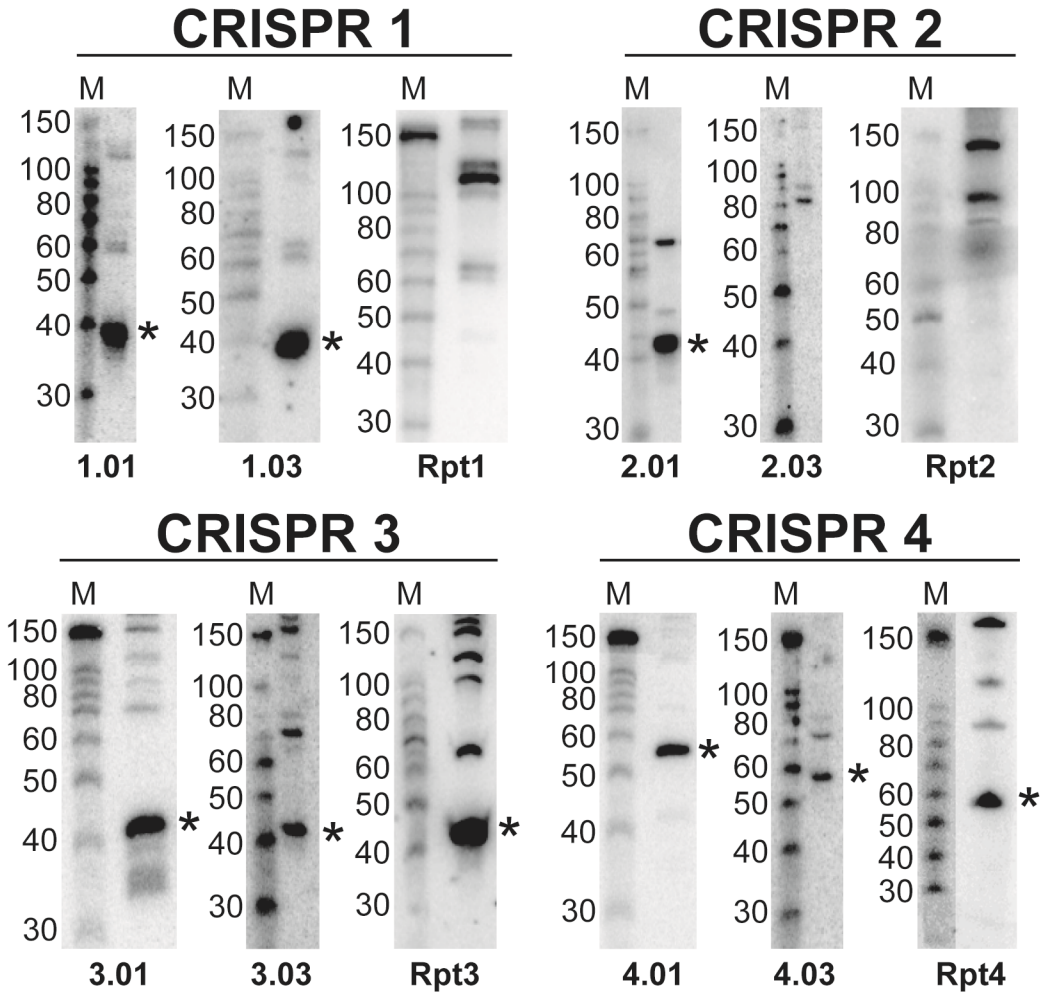
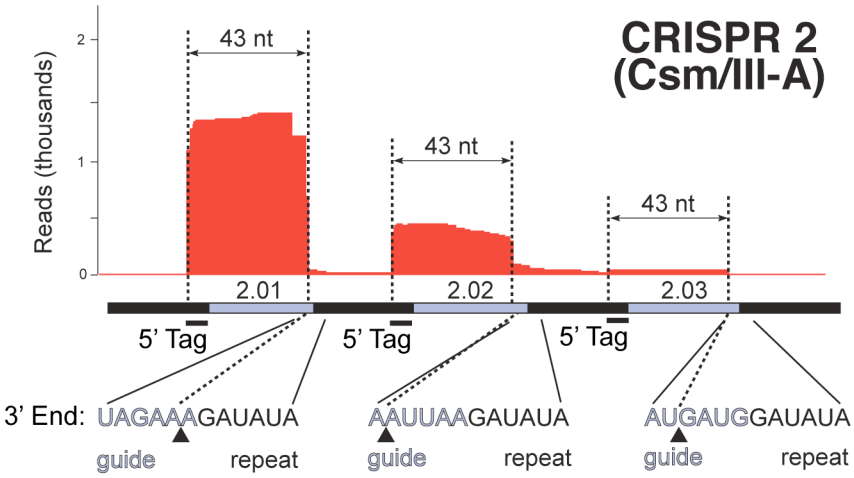

CRISPR 1
CRISPR 2
CRISPR 3
CRISPR 4
M
150
100
80
60
50
40
30
*
1.01
1.03
Rpt1
2.01
2.03
Rpt2
3.01
3.03
Rpt3
4.01
4.03
Rpt4

Figure 2.S3. The 3' end of crRNAs from *Sth* CRISPR 2 are a fixed distance from the mature 5' end and independent of genome-encoded guide length. The graph shows deep sequencing profiles of crRNAs from CRISPR2 and below the position of the mature 3' is indicated with an arrow. CrRNAs of 43 nucleotides are generated despite having encoded spacer lengths of 36, 40, and 39 base pairs, respectively.


CRISPR 2
(Csm/III-A)
Reads (thousands)
2
1
0
43 nt
43 nt
43 nt
2.01
2.02
2.03
5' Tag
5' Tag
5' Tag
3' End: UAGAAAGAUAUA
AAUUAAGAUAUA
AUGAUGGAUAUA
guide
repeat
guide
repeat
guide
repeat

CHAPTER 3

Discussion

The CC immune systems protect bacteria and archaea from genome invaders. Silencing of foreign nucleic acids by CC systems relies on small crRNAs to guide effector Cas proteins to the invader through base pairing with the invader RNA or DNA. As such, production and processing of crRNAs is a critical step in CC-mediated immunity. Here, we have investigated the crRNA biogenesis pathways of the four CC modules found in *Streptococcus thermophilus*. Prior to this work, there was nothing known about the nature of crRNAs in *S. thermophilus* or whether the different CC modules functionally interact. We found that each CRISPR array produces crRNAs composed of distinct combinations of invader-targeting guide sequence and CRISPR repeat. We showed that the Cas6 protein associated with CRISPR 2 (Type III-A) specifically binds and cleaves RNAs containing the CRISPR 2 repeat sequence, whereas the Cas6e protein associated with the Type I-E CC module 4 recognizes and cleaves repeat 4 RNAs. Furthermore, accumulation of crRNAs from CRISPR 1 (Type II-A), but not the other CC systems, requires the Cas9 protein associated with this module. Taken together, our data demonstrate that each of the four CC systems found in *S. thermophilus* have independent crRNA biogenesis pathways.

Basis for independently functioning CC modules

CC modules are composed a CRISPR array and an adjacently encoded set of *cas* genes that appear to be transmissible between bacterial and archaeal genomes as intact units. This has led to the emergence of many types and sub-type of CC systems in which *cas* genes appear to have co-evolved with the repeat sequences found in the associated CRISPR array and the upstream leader sequence (Kunin *et al.* 2007; Makarova *et al.* 2011). Given that *S. thermophilus* CRISPR 2 (Type III-A) and CRISPR 4 (Type I-E) are highly divergent in terms of *cas* gene content and repeat sequence (Figure 1.1), it is not surprising that these modules contain independent pathways for crRNA biogenesis (Chapter 2). In agreement with our findings, a recent study in done in the cyanobacterium *Synechocystis 6803* demonstrated that CC modules of three different types function independently in crRNA biogenesis (Scholz *et al.* 2013). However, here we show for the first time that two CC modules of the same sub-type, *S. thermophilus* CRISPRs 1 and 3 (both Type II-A systems) independently produce mature crRNAs of distinct compositions (Chapter 2). The basis for this autonomous function appears to be the high level of evolutionary divergence between these two modules (Figure 3.1).

Although the two Type II-A CC modules found in *S. thermophilus* contain the same *cas* gene content and organization, the Cas protein amino acid, the direct repeat, and the leader sequences are all highly divergent ((Horvath *et al.* 2008; Horvath and Barrangou 2010)and (Figure 3.1)). These two CC systems have also been shown to utilize distinct PAM sequences for spacer acquisition and invader silencing (Horvath *et al.* 2008; Garneau *et al.* 2010; Sapranauskas *et al.* 2011; Gasiunas *et al.* 2012; Sun *et al.* 2013). It therefore seems plausible that the Cas proteins found within these two modules have sufficiently different specificities to prevent cross-talk between the two systems in crRNA biogenesis.

The observation that two CC modules of the same sub-type can co-exist in the same environment yet function independently begs the question, how unalike do two CC modules have to be to prevent functional interactions between two or more CC systems present in the same organism? For example, the components of *S. thermophilus* CRISPR 3 contain much higher homology to the Type II-A CC system found in *Streptococcus pyogenes*, including a shared NGG PAM sequence (Figure 3.1). If these two systems co-existed would they interfere with one another, cooperate, or remain independently functional? If there were cross-talk between the two Type II-A CC systems, would it be restricted one or two stages within the defense pathway (E.g. only in crRNA biogenesis or invader silencing)? Addressing these questions systematically would not only provide insight into the mechanisms by which specificity within a particular CC module is mediated, but also could be very important for engineering strains with multiple CC modules, say, in an attempt to increase host fitness in an industrially important bacterial species.

Genome manipulation using CC machinery

Components of Type II CC systems have recently been utilized to edit the genomes of human, yeast, zebrafish, and bacterial cells using the dual RNA-guided Cas9 effector nuclease and a single guide (g)RNA (Chang *et al.* 2013; Cho *et al.* 2013; Cong *et al.* 2013; Dicarlo *et al.* 2013; Hwang *et al.* 2013; Jiang *et al.* 2013; Jinek *et al.* 2013; Mali *et al.* 2013). *In vitro* studies have shown that Cas9 cleavage of double-stranded DNA requires both a crRNA guide and a fragment of the *trans*-activating (tra)crRNA (Jinek *et al.* 2012; Karvelis *et al.* 2013). Further analysis revealed that a chimeric RNA that fuses the crRNA with a portion of the tracrRNA (later termed gRNA) supported efficient DNA cleavage by Cas9 (Jinek *et al.* 2012). In each of the genome editing studies, the Cas9 protein from either *Streptococcus pyogenes* or *S.*

pneumoniae was co-expressed along with gRNAs containing guide sequences engineered to target regions of the host genome (Chang *et al.* 2013; Cho *et al.* 2013; Cong *et al.* 2013; Dicarlo *et al.* 2013; Hwang *et al.* 2013; Jiang *et al.* 2013; Jinek *et al.* 2013; Mali *et al.* 2013). This selective targeting was shown to produce site-specific DNA breaks that greatly enhance the efficiency of insertions or deletions compared to the previously utilized zinc finger nucleases (ZFNs) or transcription activator-like effector nuclease (TALENs) (Chang *et al.* 2013; Hwang *et al.* 2013; Mali *et al.* 2013).

Genome editing by Cas9 requires an appropriate PAM sequence within the target in order to generate double-stranded DNA breaks (Chang *et al.* 2013; Cho *et al.* 2013; Cong *et al.* 2013; Dicarlo *et al.* 2013; Hwang *et al.* 2013; Jiang *et al.* 2013; Jinek *et al.* 2013; Mali *et al.* 2013). Site-specific editing was limited to regions which contained the downstream NGG PAM sequence utilized by *S. pyogenes* and *S. pneunomiae* Type II CC systems (Deltcheva *et al.* 2011). Adapting Cas9 proteins from Type II CC systems that utilize different PAM sequences, such as that associated with *S. thermophilus* CRISPR1 (NNAGAA), could substantially increase the number of sequences that could be targeted for editing (Deveau *et al.* 2008).

Also critical for effective genome manipulation by Cas9 is an appropriate crRNA guide sequence. Here we have shown that the crRNAs produced by the two Type II-A systems in *S. thermophilus* are composed of distinct combinations of guide and repeat tag sequence (Chapter 2). The crRNAs derived from *S. thermophilus* CRISPR 3 retain 20 nucleotides of guide sequence and a 22 nucleotide 3' repeat tag, which closely resembles those produced in *S. pyogenes* ((Deltcheva *et al.* 2011) and Chapter 2)). However, crRNAs produced from CRISPR 1 contain 21 nucleotides of and retain a 17 nucleotide 3' repeat tag (Chapter 2). This difference

could be very important for the design of gRNAs for genome editing applications using Cas9 associated with *S. thermophilus* CRISPR 1.

Role of Cas9 in crRNA biogenesis

The biogenesis pathway and composition of crRNAs from Type II CC systems differs considerably from Type I and III CC systems. Primary processing in Type II CC systems was intriguingly found to be carried out a small non-coding RNA, termed *trans*-activating tracrRNA, and the non-Cas protein RNase III (Deltcheva *et al.* 2011). Primary processing by RNase III generates 1X crRNA intermediates that undergo further processing whereby the 5' repeat fragment and ~10 nucleotides of the guide sequence are removed ((Deltcheva *et al.* 2011) and Chapter 2)). We found that crRNAs associated with one Type II-A CC module in *S. thermophlilus* (CRISPR 1) appear to undergo further 3' trimming following cleavage by RNase III, whereas those from the other Type II-A CC module (CRISPR 3) do not appear to. We show that deletion of the *cas9* gene from *S. thermophilus* CRISPR 1 disrupts crRNA biogenesis from this module, resulting in a loss of detectable mature crRNAs and what we predict to be 3' trimmed 1X intermediates. However, a low level of 1X intermediate was still produced (Chapter 2). This indicates that Cas9 is not absolutely required for primary crRNA processing, but is required for 5' processing and 3' trimming of 1X crRNA intermediate RNAs.

Since Cas9 is required for the accumulation of mature crRNAs in Type II-A CC systems, it has been speculated that this protein may directly cleave 1X crRNA intermediates to the mature form ((Deltcheva *et al.* 2011) and Chapter 2)). This could involve one of the previously characterized nuclease active sites (HNH or RuvC) that have been shown to carry out cleavage DNA invaders (Gasiunas *et al.* 2012; Jinek *et al.* 2012). On the other hand, Cas9 may contain a

cryptic nuclease motif located elsewhere on the protein that cleaves 1X crRNA intermediates to the mature form. Alternatively, like primary processing, maturation in Type II CC systems may also be carried out by a non-Cas ribonuclease such as the 5' to 3' exo-/endoribonucleases RNase J1 and J2, which is found many gram positive bacteria, including *S. thermophilus* and *S. pyogenes* (Even *et al.* 2005; Bugrysheva and Scott 2010).

CrRNA maturation in Type III CC systems

The process by which maturation of crRNAs in Type III CC systems remains a obscure. In Type III CC systems, primary processing of pre-crRNAs is carried out by Cas6, generating 1X crRNA intermediates that undergo further processing to crRNAs, which entails removal of the 3' repeat fragment, sometimes along with a portion of the guide. The mature 3' ends of Type III crRNAs appear to be measured from the mature 5' ends, independent of the length of the guide sequence by an unknown mechanism ((Hale *et al.* 2009; Hatoum-Aslan *et al.* 2011; Hale *et al.* 2012) and Chapter 2)). Here we show that Cas6 from *S. thermophilus* specifically recognizes and cleaves the repeat RNA sequence found with the Type III-A CRISPR 2 module (Chapter 2). The 1X crRNA intermediates produced by Cas6 cleavage are trimmed to exactly 43 nucleotides (with a minor species 37 nucleotides in length) despite encoding spacers of 36, 40, and 39 nucleotides in length (Chapter 2). This is consistent with previous studies describing crRNAs derived from the Type III-A CC system in *Staphylococcus epidermidis* and the Type III-B CC module in *Pyrococcus furiosus* (Marraffini and Sontheimer 2008; Hale *et al.* 2009; Hatoum-Aslan *et al.* 2011; Hale *et al.* 2012).

The chemical end groups present on 3' ends of crRNAs from Type III CC systems provide insights into the type(s) of nucleases that may be involved in the process of 3' end

trimming in these CC systems. Here, we show that crRNAs from Type III-A CC module in *S. thermophilus* retain hydroxyl groups at both the 5' and 3' termini, as has been shown previously with the Type III-A crRNAs from *S. epidermidis* (Hatoum-Aslan *et al.* 2011). Type III-B associated crRNAs, however, contain phosphates at the 3' ends (Hale *et al.* 2012). The presence of either a phosphate or hydroxyl on the 3' end of mature crRNAs from Type III-A and III-B CC systems, respectively, suggests that despite sharing an apparent 5' anchored measuring mechanism, that 3' trimming is carried out by distinct ribonucleases in these two systems.

Both Type III-A and III-B CC systems encode a number of the so-called repeat associated mysterious proteins (RAMPs) (Haft *et al.* 2005; Makarova *et al.* 2006). It was shown that deletion of either *csm2*, *csm3*, or *csm5* RAMP genes from the *S. epidermidis* Type III-A CC modules disrupted crRNA processing and resulted in the accumulation of 1X crRNA intermediate RNA (Hatoum-Aslan *et al.* 2011). This result, along with the observation that several RAMP proteins have been shown to possess ribonuclease activity and, lead to the hypothesis that maturation of 1X crRNA intermediates in Type III CC systems could be carried out by a novel RAMP ribonuclease(s) (Hatoum-Aslan *et al.* 2011). Each of the identified RAMP ribonucleases described to date are divalent metal independent and generate 3' phosphates or 2'-3' cyclic phosphates chemical end groups (Brouns *et al.* 2008; Carte *et al.* 2008; Haurwitz *et al.* 2010). Given that mature crRNAs associated with the Type III-B CC module retain 3'phosphate groups, cleavage by a novel RAMP nuclease seems possible. But the presence of hydroxyl groups on the 3' termini of Type III-A crRNAs suggest that these RNAs are cleaved by a divalent metal dependent ribonuclease, rather than a novel RAMP ribonuclease. It is conceivable that Type III-A CC systems encode a novel metal dependent ribonuclease or perhaps

a non-Cas ribonuclease is recruited to catalyze processing of 1X crRNA intermediates to the mature form.

Concluding remarks

Here we have shown that the four CC modules found in *S. thermophilus* have independently functioning crRNA biogenesis pathways and produce distinct mature crRNAs. These findings indicate that multiple CC systems are capable of co-existing the same environment without interfering with one another. This information will be important for engineering strains with multiples CC systems in order to increase host fitness. In addition, understanding the precise composition of crRNAs is critical to the design principles employed to address unanswered questions in the CC immune pathway and for CC applications such as genome engineering.

References

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**(5891): 960-964.

Bugrysheva JV, Scott JR. 2010. The ribonucleases J1 and J2 are essential for growth and have independent roles in mRNA decay in *Streptococcus pyogenes*. *Molecular Microbiology* **75**(3): 731-743.

Carte J, Wang R, Li H, Terns RM, Terns MP. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & Development* **22**(24): 3489-3496.

Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, Xiong JW, Xi JJ. 2013. Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. *Cell Research* **23**(4): 465-472.

Cho SW, Kim S, Kim JM, Kim JS. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology*.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA *et al.* 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*.

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**(7340): 602-607.

Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology* **190**(4): 1390-1400.

Dicarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Research*.

Even S, Pellegrini O, Zig L, Labas V, Vinh J, Brechemmier-Baey D, Putzer H. 2005. Ribonucleases J1 and J2: two novel endoribonucleases in *B.subtilis* with functional homology to *E.coli* RNase E. *Nucleic Acids Research* **33**(7): 2141-2152.

Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**(7320): 67-71.

Gasiunas G, Barrangou R, Horvath P, Siksnys V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*.

Haft DH, Selengut J, Mongodin EF, Nelson KE. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Computational Biology* **1**(6): e60.

Hale C, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CV, 3rd, Graveley BR, Terns RM *et al.* 2012. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell* **45**(3): 292-302.

Hale C, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**(5): 945-956.

Hatoum-Aslan A, Maniv I, Marraffini LA. 2011. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at

the precursor processing site. *Proceedings of the National Academy of Sciences of the United States of America* **108**(52): 21218-21222.

Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. 2010. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**(5997): 1355-1358.

Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**(5962): 167-170.

Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. 2008. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *Journal of Bacteriology* **190**(4): 1401-1412.

Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology*.

Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology*.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**(6096): 816-821.

Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. 2013. RNA-programmed genome editing in human cells. *eLife* **2**: e00471.

Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. 2013. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus. RNA Biology* **10**(5).

Kunin V, Sorek R, Hugenholtz P. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology* **8**(4): R61.

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* **1**: 7.

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF *et al.* 2011. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology* **9**(6): 467-477.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, Dicarlo JE, Norville JE, Church GM. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science*.

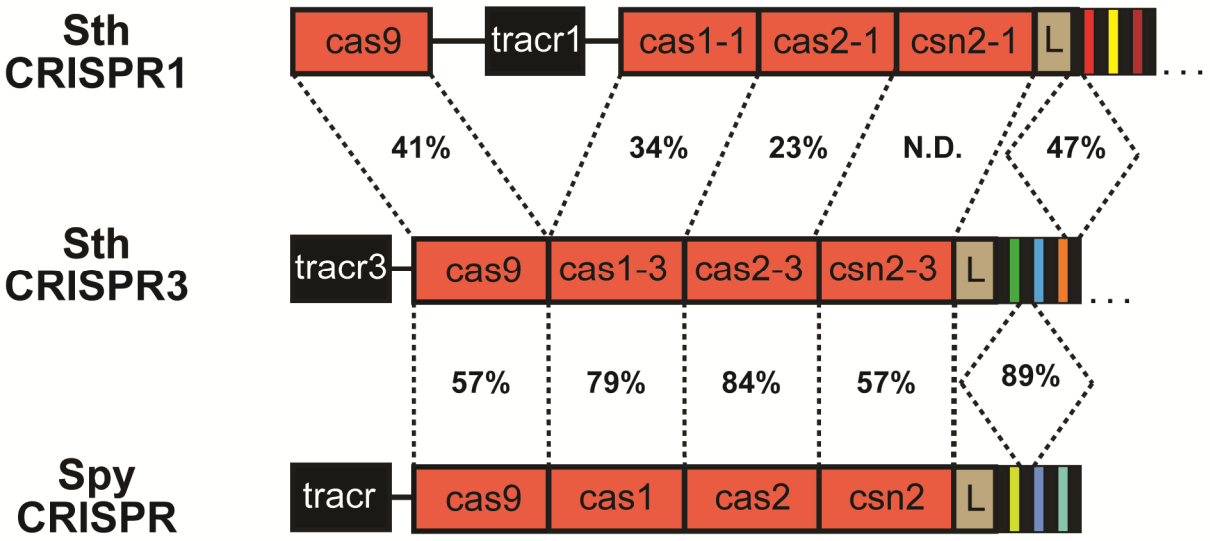
Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**(5909): 1843-1845.

Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. 2011. The Streptococcus thermophilus CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Research* **39**(21): 9275-9282.

Scholz I, Lange SJ, Hein S, Hess WR, Backofen R. 2013. CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein. *PloS One* **8**(2): e56470.

Sun CL, Barrangou R, Thomas BC, Horvath P, Fremaux C, Banfield JF. 2013. Phage mutations in response to CRISPR diversification in a bacterial population. *Environmental Microbiology* **15**(2): 463-470.

Figure 3.1. Comparison of Type II-A CC modules found in *Streptococcus thermophilus* and *S. pyogenes*. Cas protein amino acid sequence identities are indicated and nucleotide identities are indicated for CRISPR direct repeats (black boxes). There is no detectable amino acid sequence identity between the Csn2 proteins found in *S. thermophilus* (indicated by "N.D.").


Sth CRISPR1
cas9
tracr1
cas1-1
cas2-1
csn2-1
L
41%
34%
23%
N.D.
47%
Sth CRISPR3
tracr3
cas9
cas1-3
cas2-3
csn2-3
L
57%
79%
84%
57%
89%
Spy CRISPR
tracr
cas9
cas1
cas2
csn2
L