**Title**
The CRISPR endoribonuclease Csy4 utilizes unusual sequence- and structure-specific mechanisms to recognize and process crRNAs



**Author**
Haurwitz, Rachel Elizabeth

**Publication Date**
2012

Peer reviewed|Thesis/dissertation

# The CRISPR endoribonuclease Csy4 utilizes unusual sequence- and structure-specific mechanisms to recognize and process crRNAs

by
Rachel Elizabeth Haurwitz

A thesis submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Molecular and Cell Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Jennifer A. Doudna, Chair
Professor Tom Alber
Professor Jillian F. Banfield
Professor Kathleen Collins

Spring 2012

## Abstract

The CRISPR endoribonuclease Csy4 utilizes unusual sequence- and structure-specific mechanisms to recognize and process crRNAs

by

Rachel Elizabeth Haurwitz

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Jennifer A. Doudna, Chair

Many prokaryotes contain clustered regularly interspaced short palindromic repeats (CRISPRs) that together with CRISPR-associated (*cas*) genes confer resistance to invasive genetic elements. Central to this immune system is the production of CRISPR-derived RNAs (crRNAs) via enzymatic cleavage of CRISPR locus transcripts. These crRNAs serve as guides for foreign nucleic acid targeting and degradation.

Here we identify Csy4 as the endoribonuclease responsible for CRISPR transcript processing in *Pseudomonas aeruginosa* UCBPP-PA14. Biochemical assays and six co-crystal structures of Csy4 bound to substrate and product crRNAs reveal the complex mechanisms Csy4 utilizes to recognize, position, and cleave its cognate RNA substrate in order to generate mature crRNAs. Csy4 makes sequence-specific contacts to the major groove of its cognate RNA stem-loop and makes extensive electrostatic interactions with the phosphate backbone that are highly sensitive to the helical geometry of the substrate, resulting in an extremely high affinity binding interaction ($K_d \approx$ 50 pM). Csy4 has equally tight affinity for both its substrate and product RNAs and therefore functions *in vivo* as a single turnover catalyst. Phylogenetically conserved serine and histidine residues constitute a catalytic dyad in which the serine pins the ribosyl 2′-hydroxyl nucleophile in place, allowing the histidine to deprotonate the active site 2′-hydroxyl, leading to nucleophilic attack on the scissile phosphate. The Csy4 active site lacks a general acid to protonate the leaving group and positively charged residues to stabilize the transition state, explaining why the observed catalytic rate constant is ~$10^4$-fold slower than that of RNase A. The RNA cleavage step carried out by Csy4 is essential for assembly of the Csy protein-crRNA complex that facilitates target recognition. Considering that Csy4 recognizes a single cellular substrate and subsequently sequesters the cleavage product, evolutionary pressure has likely selected for substrate specificity and high-affinity crRNA interactions at the expense of rapid cleavage kinetics.

A major goal of synthetic biology is to construct reliable and predictable genetic circuits. However, synthetic genetic systems often perform unpredictably due to structural interactions between DNA, RNA, and protein components. Here we present a novel synthetic RNA processing platform utilizing Csy4 and its cognate target RNA to physically separate otherwise linked genetic elements such as promoters, ribosome binding sites, *cis* regulatory elements, and riboregulators. Implementation of this platform provides a general approach for creating context-free standard genetic elements that can be readily applied to the bottom-up construction of increasingly complex biological systems in a plug-and-play manner.

## Acknowledgements

There are many people without whom this thesis would have never been possible. I am tremendously grateful to all members of the Doudna Lab with whom I have crossed paths. They have all played important roles in shaping me as a scientist, and their insight, advice, and suggestions over the years have been quite valuable to my work. A special thanks to Blake Wiedenheft for introducing me to the fascinating field of CRISPRs and helping me to shape “the questions” I have asked during my graduate career. Thanks to Martin Jinek for being an incredible crystallography mentor/guru. I have benefitted tremendously from his patient teaching and I’m sorry for pestering him with more questions than probably all of the other lab members combined. I have a power point document compiling dozens of the kernels of wisdom I have picked up from him. Tricycle riding at the APS with Blake and Martin counts high on the list of favorite graduate school activities. I am indebted to Kaihong Zhou for patiently teaching me the basics of biochemistry and for her magical hands at the bench. Thanks to Sam Sternberg for his enthusiasm for all things Csy4 and his expertise in quantitative biochemistry. His softball pitching skills aren’t half bad, either. I am grateful to Dipa Sashital for constant advice on a variety of protein/RNA related topics. Andy Mehle could always be counted on for delightful sessions speculating about the various mechanisms of CRISPR-mediated immunity. I left those conversations feeling even more enthusiastic for the work ahead. The ever-growing Team CRISPR has been a great source of ideas and reality checks, and I look forward to seeing all the exciting work they will produce. I was lucky enough to rotate with and then join the Doudna Lab with Cameron Noland, an extremely funny and talented scientist with whom I am so glad to have shared my graduate experience. I am thankful to Lei (Stanley) Qi in Adam Arkin’s lab for inviting me to collaborate with him and giving me the opportunity to dabble in synthetic biology. And of course I am deeply grateful to Jennifer Doudna, who allowed me great leeway in my graduate studies and who supported me with guidance, ideas, and feedback throughout my time in the lab. Her constant enthusiasm for research was a positive force throughout my graduate career and I know that I have been very lucky to do my graduate work in her laboratory.

My committee members have been excellent sounding boards over the years and their advice has directly shaped my experimental work. Tom Alber’s ideas shaped my crystallographic experiments and were instrumental to being able to crystallize Csy4 in a variety of conditions and space groups. Jill Banfield’s tremendous knowledge of the CRISPR literature is a valuable resource, and I quite enjoyed the informal sessions between her lab and the Doudna lab in which we discussed ongoing efforts to understand CRISPR-mediated immunity. Kathy Collins, with whom I enjoyed rotating as a first year student, has provided tremendous feedback on all things RNA-related and helped us to better design experiments to tease out the explicit molecular mechanisms of Csy4 function.

Finally, I owe everything to my family for their constant support of my work and me. I am grateful that my parents have taken such an interest in my research, and I always really enjoy explaining to them what I study. From birth they have encouraged me to follow my

passions and to do and study what I love, and I feel exceptionally lucky that they have unconditionally supported me along the way. Their love and guidance have truly made me feel that anything is possible. And to Alex Smoligovets, my boyfriend and classmate who has been through it all with me, I owe my sanity. His love and support have been instrumental to my success and I know that I am a better person and a better scientist because of him. Together we have taken up rowing and running, and I know that our athletic pursuits have been an important balance to the science in our lives. I am lucky to share my life with him, and thank him for all he has done for me.

## Table of Contents

## List of Figures

## List of Tables

# Chapter 1

## Introduction: CRISPR-mediated immunity in prokaryotes

### 1.1 Prokaryotes rely on a variety of mechanisms to fight selfish genetic elements

In environmental samples, bacteria and archaea are outnumbered by viruses nearly ten-fold (Brussow and Hendrix, 2002). Constant viral infection leads to never ending cycles of co-evolution as the prokaryotes evade viruses, and the viruses adapt to new hosts (Labrie et al., 2010). Viruses, therefore, have a profound impact on the evolution of bacterial and archaeal species. Prokaryotes have developed a number of strategies to block or combat viral infection including blocking phage adsorption, preventing viral DNA entry into the host, cleaving viral DNA with restriction-modification enzymes, abortive infection systems, and CRISPR-mediated acquired immunity (Labrie et al., 2010).

### 1.2 CRISPR-mediated immunity in prokaryotes

Approximately half of all sequenced bacteria and nearly all sequenced archaea harbor one or more Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) loci (Grissa et al., 2007; Rousseau et al., 2009), highly repetitive genomic regions composed of a series of direct repeats separated by unique spacer sequences (Al-Attar et al., 2011). CRISPR loci co-occur with and are often physically adjacent to CRISPR-associated (*cas*) genes, which are found only in organisms that have CRISPR loci (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). Many spacers are identical in sequence to fragments of viral genomes and plasmids (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). The CRISPR locus in conjunction with the *cas* genes constitutes an acquired nucleic-acid based immune system that protects bacteria and archaea from infection by foreign genetic elements (Wiedenheft et al., 2012).

#### 1.2.1 CRISPR loci

CRISPR loci contain at least two direct repeats and as many as ~250. Many prokaryotic genomes contain a single CRISPR locus, but some have as many as 18 (Sorek et al., 2008). Repeat sequences vary between 28 and 40 nucleotides in length, and spacer sequences vary between 26 and 72 nucleotides (Al-Attar et al., 2011). Within a single CRISPR locus, all of the repeat sequences are identical or near-identical, with the exception of the final repeat which is frequently degenerate (Grissa et al., 2007). A bioinformatic analysis of repetitive sequences defined greater than 30 families of repeat sequences. Of the most abundant families, half are quasi-palindromic in nature and therefore predicted to form stable secondary structures (Kunin et al., 2007). A CRISPR locus is preceded by a leader sequence, an A/T-rich element typically hundreds of nucleotides long. Within a single organism, leader sequences are well conserved, but across distinct species, leader sequences are highly variable (Jansen et al., 2002). In *Escherichia coli* K12, the leader sequence has been demonstrated to function as a promoter element *in vitro* and *in vivo* (Pul et al., 2010) and in *Staphylococcus epidermidis* RP62a, the leader sequence is required for transcription of the CRISPR locus (Marraffini and Sontheimer, 2008).

#### 1.2.2 *cas* genes

An original bioinformatic analysis of *cas* genes identified ~45 distinct *cas* gene families (Haft et al., 2005). More recent bioinformatic work relying on *cas* gene primary sequences and biochemical and structural data on Cas proteins has combined several

of these *cas* gene families and classified CRISPR/Cas immune systems into three major types and ten sub-types (Fig. 1.1; (Makarova et al., 2011)). Cas1 is the protein hallmark of CRISPR-mediated immunity, and is the only protein found in all CRISPR-containing organisms (Makarova et al., 2006).

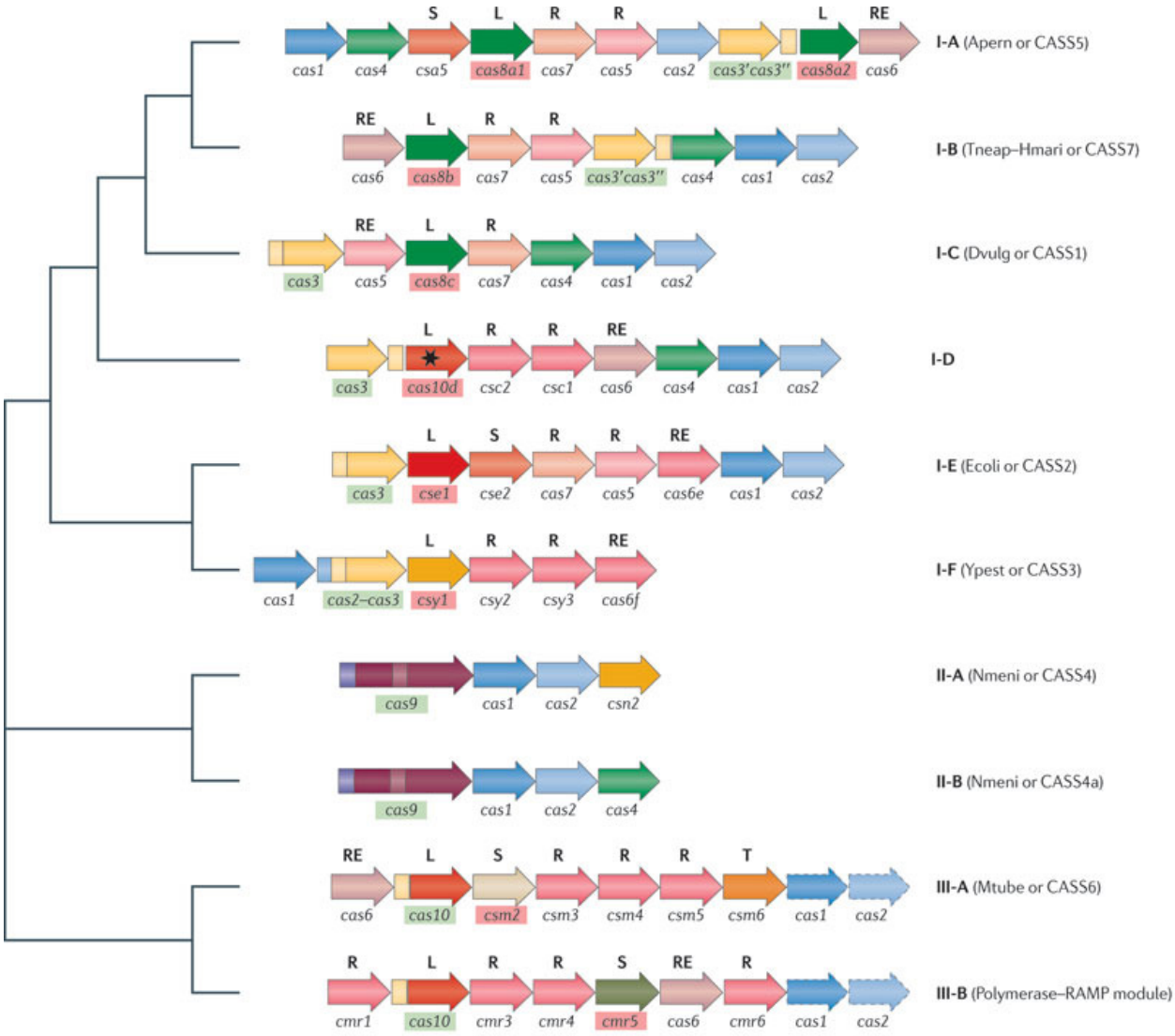



**Figure 1.1 CRISPR/Cas immune systems fall into three large types and ten sub-types based on phylogenetic analysis and protein composition.** Typical operon structures for each sub-type are shown. Orthologous genes are color coded. Large CRISPR-associated complex for antiviral defense (Cascade) subunits (L), small Cascade subunits (S), repeat-associated mysterious protein (RAMP) Cascade subunits (R), RAMP family ribonucleases involved in CRISPR RNA processing (RE), and transcriptional regulators (T) are denoted. Some of these protein classifications have been experimentally validated, while many are based on computational analyses. Adapted from (Makarova et al., 2011).

### 1.3 Three stages of immunity

CRISPR-mediated immunity occurs in three stages. First, a fragment of virus or plasmid DNA (termed the protospacer) is incorporated into the host's CRISPR locus as a new spacer. Second, the CRISPR locus is transcribed as a single RNA (pre-crRNA), which is then processed by endonucleolytic cleavage into mature crRNAs comprising a spacer sequence flanked by portions of the repetitive element. Finally, the mature crRNA is incorporated into a large multi-protein complex which targets the crRNA to invading nucleic acids via base complementarity between the spacer and the invader, triggering degradation of the invading nucleic acid (Wiedenheft et al., 2012).

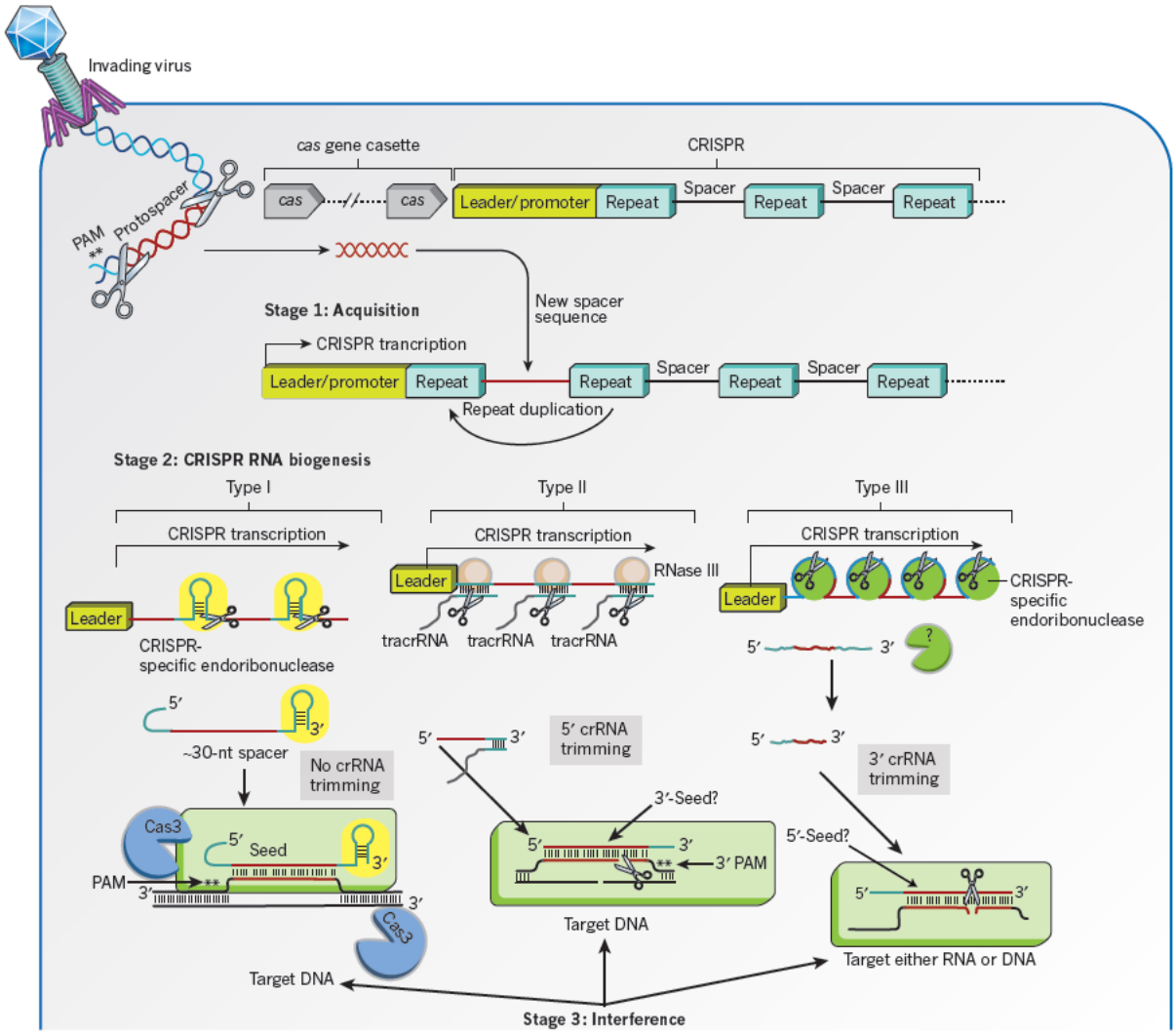



**Figure 1.2 Mechanistic overview of CRISPR-mediated immunity.** Stage 1: An infected prokaryote incorporates a piece of invading nucleic acid (the protospacer) into a genomic CRISPR locus as a new spacer at the leader-edge of the CRISPR. Stage 2: The CRISPR locus is transcribed as a single pre-crRNA and is processed into mature crRNAs via endonucleolytic cleavage. This processing step is carried out by different processing machinery in the three CRISPR/Cas types (see below). Stage 3: The crRNA is targeted to invading nucleic acid by a large multi-protein complex. Base pairing between the crRNA spacer and the target nucleic acid triggers its degradation. Adapted from (Wiedenheft et al., 2012).

### 1.3.1 Spacer acquisition

Little is known about the mechanism of spacer acquisition. Despite the widespread prevalence of CRISPR/Cas immune systems, only *Streptoccocus thermophilus* and *E. coli* have been observed to acquire spacers in the laboratory (Barrangou et al., 2007; Deveau et al., 2008; Garneau et al., 2010; Yosef et al., 2012). However, metagenomic studies involving environmental samples demonstrate a constant battle between viruses and prokaryotes in which prokaryotes acquire new spacers and the viruses rapidly mutate to evade the CRISPR system (Andersson and Banfield, 2008; Snyder et al., 2010; Tyson and Banfield, 2008). The first demonstration of acquisition by *E. coli* was in March 2012, which has opened the door to the power of *E. coli* genetics and robust biochemical strategies to study the mechanism of spacer acquisition.

In laboratory conditions, *S. thermophilus* robustly integrates one or more spacers during either plasmid or phage DNA challenge (Barrangou et al., 2007; Deveau et al., 2008; Garneau et al., 2010). Genetic analysis has demonstrated that Csn2 (formerly known as Cas7) is required for new spacer acquisition (Barrangou et al., 2007). However, Csn2 is unique to the type II-A system. It is therefore likely that one or more of the proteins found in the other subtypes are functional orthologues of Csn2 (Wiedenheft et al., 2012).

Because Cas1 is the sole protein found in all CRISPR-containing organisms (Makarova et al., 2006) and because it is not required for crRNA-guided silencing in *E. coli* K12 (Brouns et al., 2008), it has been hypothesized that Cas1 is an integrase that plays a role in CRISPR acquisition (Makarova et al., 2006). Biochemical and structural analysis of Cas1 proteins from three organisms (*Pseudomonas aeruginosa*, *E. coli* K12, and *Sulfolobus solfataricus*) has demonstrated deoxyribonuclease and nucleic acid-binding activity. Crystal structures of these Cas1 proteins reveal a homodimer with a novel fold (Babu et al., 2011; Han et al., 2009; Wiedenheft et al., 2009).

In *E. coli*, Cas1 and Cas2 are the Cas proteins necessary and sufficient to mediate spacer acquisition (Yosef et al., 2012). The *E. coli* strain BL21 (DE3) contains two genomic CRISPR loci but does not code for any *cas* genes (Grissa et al., 2007). *E. coli* K12 genome contains eight *cas* genes (Cas1-3 and CasA-E) and two CRISPR loci (Brouns et al., 2008), but the *cas* genes are transcriptionally silenced by the global transcription factor H-NS (Pougach et al., 2010; Pul et al., 2010; Westra et al., 2010). Overexpression of plasmid-encoded Cas1 and Cas2, but not either protein alone, in both of these strains leads to robust acquisition of new spacer sequences into both of the genomic CRISPR loci (Yosef et al., 2012). Understanding the mechanistic role of both Cas1 and Cas2 will be crucial to determining the molecular mechanism of spacer acquisition.

### 1.3.2 crRNA biogenesis

Precursor transcripts encompassing the full-length CRISPR locus (pre-crRNA) are transcribed and cleaved within each repeat sequence to generate mature crRNAs that consist of a spacer sequence flanked by portions of the repeat sequence (Marraffini and Sontheimer, 2010). CRISPR-Cas immune systems from the three broad types utilize distinct sets of enzymes to process pre-crRNAs (Makarova et al., 2011). In the type II CRISPR system, RNase III cleaves an RNA duplex formed by the CRISPR

repeat and a trans-activating CRISPR RNA (tracrRNA) (Fig. 1.2; (Deltcheva et al., 2011)). In the type I and type III systems, a CRISPR-specific endoribonuclease binds and cleaves the repeat elements in a sequence-specific fashion (Fig. 1.2; (Brouns et al., 2008; Carte et al., 2010; Carte et al., 2008; Gesner et al., 2011; Haurwitz et al., 2010; Haurwitz et al., 2012; Lintner et al., 2011; Sashital et al., 2011; Sternberg et al., 2012). Two of these enzymes, Cse3 and Csy4 from type I-E and type I-F, are single turnover catalysts and remain bound to the crRNA product after cleavage; both the crRNA and the processing endonuclease are components of the downstream targeting complex (Brouns et al., 2008; Haurwitz et al., 2012; Jore et al., 2011; Sashital et al., 2011; Sternberg et al., 2012; Wiedenheft et al., 2011b). However, the Cas6 CRISPR processing endonuclease from type III-B is not a component of the targeting complex (Hale et al., 2009) and the Cas6 from type I-A remains weakly bound to the archaeal Cascade (Lintner et al., 2011).

Three CRISPR-specific processing enzymes from types I and III have been biochemically and structurally characterized (Fig. 1.3). Though these three enzymes share no detectable primary sequence similarity, they all adopt ferrodoxin-like folds (Fig. 1.3; (Wiedenheft et al., 2012)). Cas6 from *Pyroccocus furiosus* consists of a double ferrodoxin-like fold and binds its cognate single stranded CRISPR RNA repeat via sequence-specific recognition of the 5′ end of the repetitive element. Co-crystal structures of Cas6 bound to crRNA have interpretable density for only the 5′ end of the CRISPR repeat, while cleavage happens in the 3′ end of the repetitive element. The RNA likely wraps around the protein to the opposite face where Cas6 cleaves the RNA in an AA dinucleotide motif (Carte et al., 2010; Carte et al., 2008; Wang et al., 2011). The mature crRNAs loaded into the targeting complex in *P. furiosus* are smaller than the initial Cas6 cleavage products, and likely result from either endonuclease or exonuclease trimming of the 3′ end (Hale et al., 2009). Like Cas6, Cse3 from *Thermus thermophilus* also adopts a double ferrodoxin-like fold. However, it binds RNA on the opposite face of the double ferrodoxin-like fold than Cas6, and it utilizes a β-hairpin to insert into the major groove of the repeat stem-loop. Additionally, Cse3 makes interactions downstream of the stem-loop that contribute to unwinding of the bottom base pair of the stem-loop which is necessary for cleavage to occur (Gesner et al., 2011; Sashital et al., 2011). Csy4 from *Pseudomonas aeruginosa* adopts an N-terminal ferrodoxin-like fold, and its C-terminal half has the same secondary structure connectivity of a ferrodoxin-like fold, but the overall architecture is altered. An arginine-rich helix inserts into the major groove of the RNA stem-loop and uses two amino acid side chains to read out the identity of the bottom two base pairs of the hairpin. Csy4 makes a sequence-specific interaction with the first single stranded nucleotide upstream of the stem-loop, but does not interact with any of the nucleotides downstream of the stem-loop (Haurwitz et al., 2010).

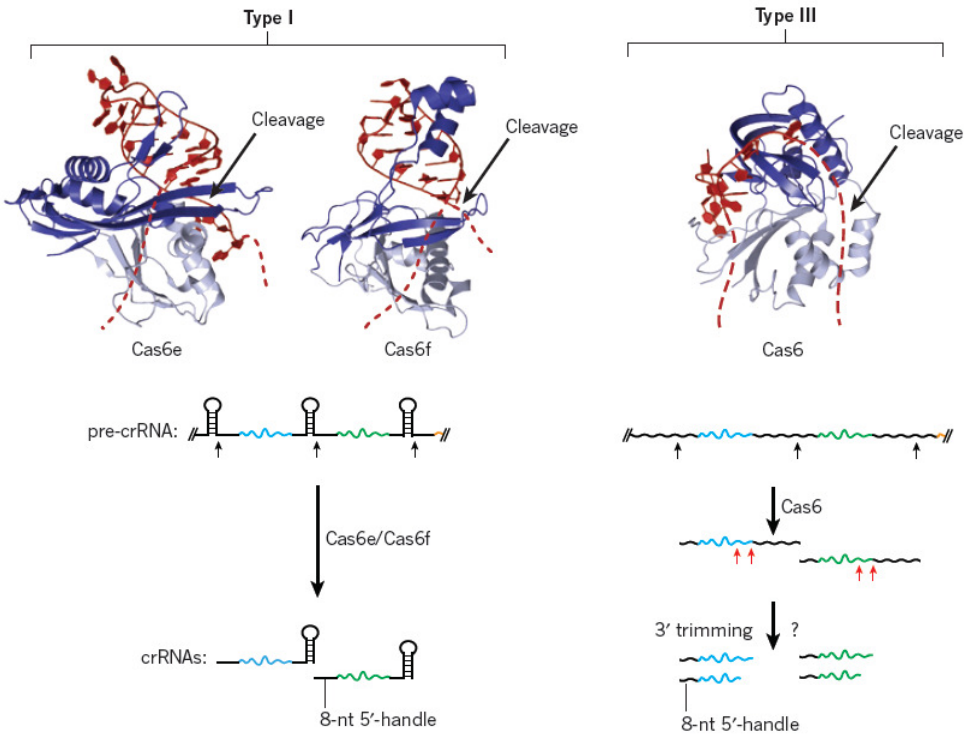



**Figure 1.3 CRISPR-specific endoribonucleases recognize and cleave their CRISPR repetitive element to generate mature crRNAs.** In the type I CRISPR system (left panel), Cse3 (aka Cas6e; PDB ID 2Y8W) and Csy4 (aka Cas6f; PDB ID 2XLK) recognize their cognate CRISPR repeats via sequence- and structure-specific interactions with a stem-loop sequence. A β-hairpin (Cse3) or an arginine-rich helix (Csy4) bind to the RNA major groove and mediate extensive electrostatic interactions with the RNA substrate. In the type III CRISPR system (right panel), Cas6 (PDB ID 3PKM) makes sequence-specific interactions with the 5′ end of the CRISPR RNA repeat on the opposite face of the protein containing the active site residues. Adapted from (Wiedenheft et al., 2012).

### 1.3.3 Targeting

In the final stage of CRISPR-mediated immunity, a large multi-protein complex targets crRNAs to foreign nucleic acids. Base complementarity between the crRNA spacer sequence and the foreign nucleic acid leads to degradation of the complementary sequence. *In vivo* activity assays have demonstrated that crRNAs targeting either coding or non-coding regions and template or non-template strands of viral and plasmid genomes lead to silencing, demonstrating that these systems directly target the viral or plasmid genomic DNA rather than messenger RNA (mRNA) (Barrangou et al., 2007; Brouns et al., 2008; Deveau et al., 2008; Garneau et al., 2010; Gudbergsdottir et al., 2011; Manica et al., 2011; Marraffini and Sontheimer, 2008). The multi-protein complexes responsible for targeting have been identified in several CRISPR/Cas types. In type I-E, the CRISPR-associated complex for antiviral defense (Cascade) consists of a non-stoichiometric distribution of Cas proteins ($CasA_1$:$CasB_2$:$CasC_6$:$CasD_1$:$CasE_1$) and a single crRNA (Brouns et al., 2008; Jore et al., 2011). A recent cryo-electron microscopy reconstruction of Cascade at ~8 Å resolution demonstrated that the crRNA runs the length of the complex, and the six

CasC subunits form a filament along the spacer sequence. CasE (also known as Cse3) remains bound to the hairpin sequence of the CRISPR repeat (Wiedenheft et al., 2011a). Together with Cas3, which contains an HD-nuclease domain and a DExD/H box helicase domain (Haft et al., 2005; Makarova et al., 2006), Cascade targets complementary invading DNAs and can reduce phage sensitivity by seven orders of magnitude (Brouns et al., 2008). *In vitro* biochemical data have demonstrated that Cas3 has ATPase activity that is stimulated by single-stranded DNA which facilitates unwinding of DNA/DNA or DNA/RNA duplexes and DNase activity, which play a key role in targeting Cascade-bound crRNAs to invading nucleic acids and subsequently destroying them (Howard et al., 2011; Sinkunas et al., 2011; Westra et al., 2012).

In type I-F, as demonstrated in *P. aeruginosa*, four Cas proteins and a crRNA assemble into the Csy complex ($Csy1_1$:$Csy2_1$:$Csy3_6$:$Csy4_1$). A low-resolution small angle x-ray scattering (SAXS) reconstruction of the Csy complex revealed a gross morphology similar to that of the type I-E Cascade (Wiedenheft et al., 2011b).

In the type I-A system, as typified in *S. solfataricus*, the archaeal Cascade (aCascade) contains Csa2, Cas5a, Cas6, Csa5, and crRNA. The minimal requirements for stable complex formation and ssDNA targeting binding are Csa2 and Cas5a. Negative stain transmission electron microscopy has identified an overall helical structure of varied length that is consistent with an RNA/protein filament similar to Cascade and the Csy complex (Lintner et al., 2011).

A large multi-protein complex from the type II-A system has not yet been identified, but the large protein Cas9 is required for virus and plasmid silencing (Barrangou et al., 2007; Garneau et al., 2010). The viral or plasmid sequence is cleaved in both strands within the region complementary to the crRNA (Garneau et al., 2010).

Not all regions of a crRNA spacer are equally important for target recognition. In both the type I-E and type I-F systems, there exists a 7- to 8-nucleotide high affinity binding seed sequence at the 5' end of the crRNA spacer. Even single mutations here abolish targeting capability, whereas as many as five mutations are tolerated elsewhere in the spacer sequence (Semenova et al., 2011; Wiedenheft et al., 2011b). Successful targeting and silencing also requires the presence of a protospacer adjacent motif (PAM), a short (2- to 5-nucleotide) motif adjacent to the protospacer sequence in the viral or plasmid genome (Mojica et al., 2009). Single nucleotide mutations in the PAM disrupt the ability of the CRISPR system to target and destroy invading DNAs (Deveau et al., 2008; Semenova et al., 2011). The PAM is likely so crucial because it plays a role in distinguishing between self and non-self target sequences. The genomic CRISPR loci themselves harbor the appropriate targeting sequence, but cleavage of the host DNA would be lethal. Therefore the CRISPR system must differentiate between self CRISPR loci and non-self protospacers. The presence of the PAM adjacent to the protospacer facilitates this differentiation (Semenova et al., 2011). A recent crystal structure of the CasA component of the type I-E Cascade was docked into the cryo EM map and uncovered a flexible loop that appears to bind where the PAM would be in a *bona fide* target DNA. Mutations to this loop demonstrated three amino acids that likely interact directly with the PAM and that are required both for Cascade assembly and DNA target binding (Sashital et al., 2012). The PAM also appears to play a direct role in spacer acquisition, as spacers acquired via the expression of only Cas1 and Cas2 (and not

Cas3 or Cascade) resulted in spacer acquisition from protospacers adjacent to the tri-nucleotide motif AWG (Yosef et al., 2012).

Not all CRISPR systems target DNA. The Cmr complexes (type III-B) from *Pyrococcus furiosus* (Hale et al., 2012; Hale et al., 2009) and *S. solfataricus* (Zhang et al., 2012) target RNA. The *P. furiosus* Cmr complex consists of six different proteins (Cmr1-6) and a crRNA. It cleaves complementary RNAs at a defined length measured from the 3' end of the crRNA and the cleaved RNA has a cyclic phosphate at its 3' terminus (Hale et al., 2012; Hale et al., 2009). The Cmr complex from *S. solfataricus* contains seven different proteins (Cmr1-7) and a crRNA. Instead of cleaving target RNAs a defined length from the crRNA terminus, it cleaves complementary RNAs within a UA dinucleotide, leaving cleavage products with a 5'-phosphate and a 3'-hydroxyl. Cleavage of target RNAs is manganese-dependent and stimulated by ATP. Electron microscopy of the *Ss*Cmr complex reveals a morphology similar to a crab claw attached to a protruding region, which is not similar to Cascade, aCascade, or the Csy complex (Zhang et al., 2012). Neither the *Pf* nor *Ss* Cmr systems require a PAM for efficient targeting (Hale et al., 2012; Zhang et al., 2012). The *in vivo* function of RNA-targeting CRISPR/Cas systems has not yet been determined.

# Chapter 2

## Identification and characterization of Csy4, the enzyme responsible for crRNA biogenesis in *Pseudomonas aeruginosa* UCBPP-PA14

*A portion of the work presented in this chapter has been previously published as part of the following paper: Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K., Doudna, J.A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science 329*, 1355-1358.

*Rachel Haurwitz performed all biochemical experiments, grew the crystals, and aided in crystallographic data collection. Dr. Martin Jinek solved the reported crystal structures and aided in crystallographic data collection and the biochemical assay reported in Fig. 2.3C. Dr. Blake Wiedenheft performed the Northern blot. Kaihong Zhou expressed and purified the selenomethionine-derivitized Csy4 and the pre-crRNAs.

## 2.1 Introduction

As discussed in Chapter 1, fragments of foreign DNA are integrated into prokaryotic clustered regularly interspaced short palindromic repeat (CRISPR) loci that are transcribed as long RNAs containing a repetitive sequence element derived from the host (Barrangou et al., 2007; Brouns et al., 2008; Carte et al., 2008; Haft et al., 2005; Hale et al., 2009; Makarova et al., 2006). These CRISPR transcripts (pre-crRNAs) are post-transcriptionally processed into short crRNAs that serve as homing oligonucleotides to prevent the propagation of invading viruses or plasmids harboring cognate sequences (Brouns et al., 2008; Marraffini and Sontheimer, 2008, 2010). CRISPR loci coexist with CRISPR-associated (Cas) proteins (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006; van der Oost et al., 2009). The opportunistic pathogen *Pseudomonas aeruginosa* UCBPP-PA14 (Pa14) harbors a CRISPR/Cas system that contains two CRISPR elements flanked by six *cas* genes (Fig. 2.1). Both CRISPRs comprise a characteristic arrangement of 28-nucleotide near-identical repeats interspersed with ~32-nucleotide spacers, some of which match sequences found in bacteriophages or plasmids (Grissa et al., 2007). Processing of primary CRISPR transcripts yields crRNAs that contain one spacer sequence flanked by sequences derived from the repeat element (Brouns et al., 2008; Carte et al., 2008; Lillestol et al., 2006; Lillestol et al., 2009; Tang et al., 2002; Tang et al., 2005).

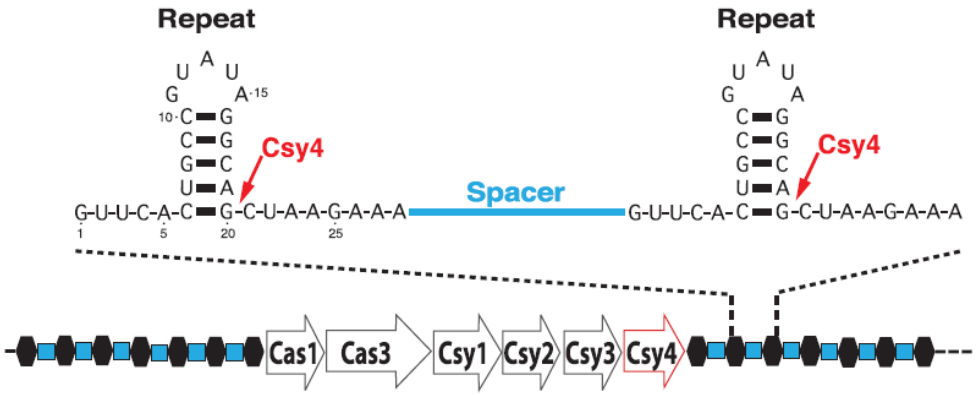



**Figure 2.1 Schematic of CRISPR/Cas locus in Pa14.** The six *cas* genes are flanked by two CRISPR loci, each consisting of a series of 28-nucleotide repeats (black lettering) separated by 32-nucleotide distinct spacer sequences (blue). Red arrows denote the Csy4 cleavage site.

## 2.2 Methods

### 2.2.1 Gene annotation, cloning, protein expression and purification

Comparative sequence analysis of Csy4 genes across species identified a conserved region 20 codons upstream of the annotated start codon in the Pa14Csy4 gene (Lee et al., 2006). The conserved full-length Csy4 (PA14_33300) sequence (residues 1-187) was PCR amplified from *Pseudomonas aeruginosa* UCBPP-PA14 genomic DNA using the following oligonucleotide primers

Forward: 5'-CACCATGGACCACTACCTCGACATTCG-3'

Reverse: 5'-GAACCAGGGAACGAAACCTCC-3'

The PCR product was subcloned using the Gateway system into the pENTR/TEV/D-TOPO entry vector (Invitrogen), followed by site-specific recombination into expression vector pHGWA or pHMGWA (Busso et al., 2005). Point mutations were introduced into Csy4 using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) and verified by DNA sequencing. The Pa14Csy4 expression plasmid was transformed into *E. coli* Rosetta 2 (DE3) cells (Novagen) or co-transformed with a pMK vector expressing a synthetic Pa14 CRISPR RNA consisting of eight repeats (from the CRISPR locus adjacent to Cas1) and seven spacers, all of which are the same sequence, synthesized by Geneart (Regensburg, Germany). Protein expression was induced with 0.5 mM isopropyl β-D-1- thiogalactopyranoside (IPTG) at an optical cell density (OD600) of ~0.5, followed by shaking at 18 ºC for 16 hours. Cells were lysed by sonication in 15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 10 mM imidazole, 5% glycerol, 0.01% Triton X-100, 100 $U.ml^{-1}$ DNase I, 1 mM Tris[2-carboxyethyl] phosphine hydrochloride (TCEP), 0.5 mM phenylmethylsulfonyl fluoride, pH 7.4, supplemented with protease inhibitors (Roche). The clarified lysate was incubated with Ni-NTA affinity resin in batch (Qiagen) and the bound protein was eluted with a high imidazole buffer (15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 300 mM imidazole, 1 mM TCEP, 5% glycerol, pH 7.4) and dialyzed overnight in a dialysis buffer (elution buffer containing 20 mM imidazole) in the presence of tobacco etch virus (TEV) protease. The released His6 or His6-MBP tag was removed by a second nickel-affinity step. The protein was further purified by size exclusion chromatography using tandem Superdex 75 (16/60) columns (GE Life Sciences) in 100 mM HEPES pH 7.5, 500 mM KCl, 5% glycerol, 1 mM TCEP. The protein was then dialyzed against 100 mM HEPES pH 7.5, 150 mM KCl, 5% glycerol, 1 mM TCEP and concentrated to 10 $mg.ml^{-1}$. Selenomethionine (SeMet)-substituted protein was expressed in BL21(DE3) cells as previously described (Wiedenheft et al., 2009).

The Ec89Csy4 coding sequence was cloned into pHMGWA and point mutants were generated by site-directed mutagenesis. Wild-type and mutant Ec89Csy4 were expressed and purified as above.

**Hint:** For greatest purity, the ortho-nickel step must be performed as a gradient elution using 5ml nickel resins from GE, not Qiagen. A gradient from 20 mM imidazole to 500 mM imidazole is sufficient. This gradient will elute three peaks: the first is non-specifically associated nucleic acids, second is Csy4, and third is His6-MBP or His6.

### 2.2.2 Northern blotting analysis

Csy4 was heterologously expressed in *E. coli* in the presence or absence of a plasmid expressing a Pa14 CRISPR transcript. Nucleic acids co-purifying with Csy4 were isolated by phenol-chloroform extraction and separated on a 16% denaturing polyacrylamide gel. Nucleic acids were transferred to a nylon membrane (Hybond-N, GE Life Sciences) using a semi-dry transfer cell (BioRad) and probed with a [5'-32P]-labeled DNA oligonucleotide complementary to the Pa14 CRISPR repeat sequence (probe sequence 5'-GCTGCCTATACGGCAG-3'). Synthetic RNA oligonucleotides corresponding to the Pa14 (5'-GUUCACUGCCGUAUAGGCAG-3') and *E. coli* K12 (5'-

UCCCCGCGCCAGCGGGGAU-3') crRNA repeats served as positive and negative controls, respectively.

### 2.2.3 Nuclease activity assays

Pa14 pre-crRNA was *in vitro* transcribed as described (Wiedenheft et al., 2009). Synthetic 16- and 28-nucleotide RNA oligonucleotides corresponding to portions of the Pa14 CRISPR repeat sequence were obtained from Integrated DNA Technologies. 75 pmol of wild-type or mutant Csy4 were incubated with 5 pmol of Pa14 pre-crRNA or 50 pmol of the short repeat-derived RNA oligonucleotides in 10 µl reactions containing 20 mM HEPES pH 7.5, 100 mM KCl buffer at 25 ºC for five minutes. Reactions were quenched by the addition of 50 µl acid phenol-chloroform (Ambion). 10 µl additional reaction buffer were added and samples were centrifuged (16,000 x g, 30 minutes). 16 µl aliquots of the aqueous phase were mixed 1:1 with formamide loading dye and separated on 15% urea denaturing polyacrylamide gel. RNA was visualized by staining with SYBR Gold (Invitrogen). To analyze the chemistry of Csy4-catalyzed RNA cleavage, a 15-nucleotide synthetic RNA (5'-CUGCCGUAUAGGCAG-3', obtained from Integrated DNA Technologies) corresponding to the Pa14 CRISPR repeat hairpin was 3'-end labeled using [5'-32P] pCp and T4 RNA ligase (New England Biolabs). This generated a minimal Csy4 substrate with the radiolabel placed at the scissile phosphate. 20 pmol of wild-type Csy4 were incubated with 0.5 pmol radiolabeled RNA substrate in 10 µl reactions containing 20 mM HEPES pH 7.5, 100 mM KCl buffer at 30 ºC for 15 minutes and the reactions were quenched by the addition of 10 ml formamide loading dye and heating at 95 ºC for 3 minutes. The products were separated on an 18% denaturing polyacrylamide gel and visualised by phosphoimaging (Storm, GE Life Sciences).

### 2.2.4 RNA binding assays

10 µM oligonucleotide was incubated with 2.5 µM, 5.0 µM, 10 µM, 20 µM and 40 µM Csy4 in 20 mM HEPES pH 7.5, 100 mM KCl in a total volume of 10 µl. The binding reactions were analyzed on a 6% native polyacrylamide gel run with 1X TBE buffer, followed by staining with Coomassie blue.

**Hints:** Coomassie blue staining permits visualization of both bound and unbound Csy4. SYBR Gold was inadequate for staining because Csy4-binding blocks the ability of SYBR Gold to stain Csy4-bound RNA.

### 2.2.5 Crystallization

For crystallization, a synthetic RNA corresponding to nucleotides C6-C21 of the Pa14 CRISPR repeat was used (obtained from Integrated DNA Technologies). The RNA carried a 2'-deoxy modification at position G20. To reconstitute the Csy4-RNA complex, purified Csy4 was mixed with RNA in a 1:2 molar ratio and incubated at 30 ˚C for 30 min. The complex was purified by size exclusion chromatography on a Superdex 75 10/300 column in 100 mM HEPES pH 7.5 and 150 mM KCl and concentrated to 5-8 $mg.ml^{-1}$. All crystallization experiments were performed at 18 ˚C using the hanging drop vapour diffusion method by mixing equal volumes (1 µl + 1 µl) of the complex and reservoir solutions. Plate-shaped crystals of the wild-type Csy4-RNA complex were grown in 200 mM sodium citrate pH 5.0, 100 mM magnesium chloride, 20% (w/v) PEG

4000 (Fig. 2.2A). These crystals belonged to the space group *C*2, contained one copy of the complex in the asymmetric unit and diffracted to 2.3 Å resolution at synchrotron X-ray sources. To facilitate phasing using heavy atom soaks, we also reconstituted the Csy4-RNA complex using the Csy4(S22C) mutant protein. Although this complex did not ultimately yield any useful heavy atom derivatives, two additional crystal forms were fortuitously obtained in 150-160 mM sodium acetate pH 4.6, 17-18% (w/v) PEG 4000. Initially, hexagonal crystals appeared within 24 hours (Figure 2.2B). These crystals diffracted to 2.6 Å resolution, belonged to space group *P*61 and contained one copy of the complex in the asymmetric unit. Three weeks later, the same crystallization condition yielded needle-shaped crystals that belonged to space group $P2_12_12_1$, contained two copies of the complex and diffracted up to 1.8 Å resolution (Figure 2.2C). For data collection, all crystal forms were cryoprotected by soaking in their respective mother liquors supplemented with 30% glycerol prior to flash cooling in liquid nitrogen.

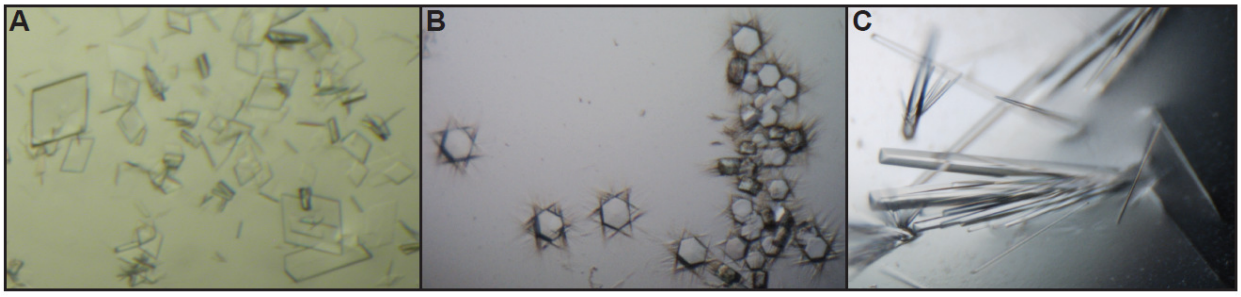



**Figure 2.2 Csy4-RNA substrate crystals.** (**A**) Plate-shaped crystals containing the wild-type Csy4-RNA complex. (**B**) Hexagonal crystals containing the Csy4(S22C) mutant in complex with RNA. (**C**) Rod-shaped crystals containing the Csy4(S22C) mutant in complex with RNA.

### 2.2.6 Structure determination

All diffraction data were collected at 100 K on beamlines 8.2.2 and 8.3.1 of the Advanced Light Source (Lawrence Berkeley National Laboratory). Data were processed using XDS (Kabsch, 2010). Experimental phases were determined from a three-wavelength multiwavelength anomalous dispersion (MAD) experiment (peak, inflection and remote data sets) using the monoclinic Csy4-RNA crystals containing selenomethionine-substituted wild-type Csy4. Two selenium sites were located using the Hybrid Substructure Search (HySS) module of the Phenix package (Grosse-Kunstleve and Adams, 2003). Substructure refinement, phasing and density modification were performed using AutoSHARP (Vonrhein et al., 2007). The resulting electron density map exhibited clear layers of density attributable to protein and RNA alternating along the *c*-axis, with the RNA layer made up of two coaxially-stacked RNA helices engaged in a "kissing loop" interaction. An initial atomic model for the Csy4 protein was obtained by automatic building using the Phenix AutoBuild module (Terwilliger et al., 2008). The complex model was completed by iterative cycles of manual building in COOT (Emsley and Cowtan, 2004) and refinement using Phenix.refine (Adams et al., 2010) against a native 2.33 Å resolution dataset, yielding a final model with a crystallographic $R_{work}$ factor of 21.8% and an $R_{free}$ factor of 24.7% (Table 2.1). The model includes RNA nucleotides C6-G20, the phosphate group

of nucleotide C21 and protein residues 1-104, 109-120 and 139-187. Owing to the layered arrangement of protein and RNA in the crystal lattice and the lack of lateral crystal contacts within the RNA layer, the RNA exhibits significant disorder, as evidenced by markedly elevated temperature factors (>100 $Å^2$) and the absence of interpretable density for the nucleotide base of U14. The disorder is also evident in protein residues 109-120, corresponding to the arginine-rich helix inserted in the major groove of the RNA, for which only the polypeptide backbone could be built (except for residues Arg115 and Arg118).

The structures of the Csy4(S22C)-RNA complex in the hexagonal and orthorhombic crystal forms were determined by molecular replacement in Phaser (McCoy et al., 2007) using the Csy4 protein (lacking the arginine-rich helix) and RNA models from the monoclinic crystal form as separate search ensembles. In both crystal forms, electron density for the arginine-rich helix and the linker region comprising Csy4 residues 105-108 was immediately noticeable in $2F_o$-$F_c$ maps obtained from the molecular replacement solutions. The structure of the Csy4(S22C)-RNA complex in the hexagonal form was refined to an $R_{work}$ factor of 24.10% and $R_{free}$ of 27.4% at 2.6 Å resolution (Table 2.1). The final model includes Csy4 residues 1-120, 124-130 (built as a polyalanine stretch) and 139-187 and RNA nucleotides C6-G20 plus the phosphate group of nucleotide C21. The orthorhombic crystal form of the Csy4(S22C)-RNA complex was solved at 1.8 Å resolution and refined to an $R_{work}$ factor of 18.4% and $R_{free}$ of 21.6%, with excellent stereochemistry (Table 2.1). Of the two complexes in the asymmetric unit, complex 1 (chains A and C) contains Csy4 residues 1-187 and RNA nucleotides C6-G20 plus the phosphate group of nucleotide C21, while the less ordered complex 2 (chains B and D) comprises Csy4 residues 1-187 (with the exception of residues 13-15 and 135-138, which show no ordered electron density) and RNA nucleotides C6-G20 and the phosphate group of nucleotide C21. The two copies of Csy4 superpose with a root-mean-square deviation (RMSD) of 1.15 Å over 179 Cα atoms, the greatest differences coming from the slightly different orientations of the arginine-rich helix relative to the remainder of the protein. The two RNA molecules in the asymmetric unit superpose with an RMSD of 1.49 Å, the largest deviation being due to the extruded nucleotide U14, which assumes different conformations in the two RNAs. Discussion and illustrations throughout this chapter are based on complex 1 in the orthorhombic crystal form. All structural illustrations were generated using Pymol (DeLano, 2002).

Coordinates and structure factors for the Csy4-crRNA complexes have been deposited in the Protein Data Bank under the accession codes 2XLI, 2XLJ, and 2XLK.

| | Native WT | Native S22C | Native S22C | SeMet WT | | |
|---|---|---|---|---|---|---|
| **Data collection** | | | | | | |
| Space group | *C 2* | $P2_12_12_1$ | $P6_1$ | *C 2* | | |
| Cell dimensions | | | | | | |
| *a, b, c* (Å) | 62.37, 46.77, 86.82 | 40.1, 78.9, 145.9 | 39.25, 39.25, 297.37 | 62.33, 47.23, 87.26 | | |
| α, β, γ (˚) | 90.0, 108.2, 90.0 | 90.0, 90.0, 90.0 | 90.0, 90.0, 120.0 | 90.0, 108.3, 90.0 | | |
| | | | | *Peak* | *Inflection* | *Remote* |
| Wavelength (Å) | 1.11159 | 0.99992 | 1.11159 | 0.97949 | 0.97971 | 0.97204 |
| Resolution (Å)* | 19.68-2.33 (2.50-2.33) | 69.4-1.80 (1.90-1.80) | 22.38-2.70 (2.80-2.70) | 82.96-2.90 (2.100-2.90) | 82.96-2.90 (2.100-2.90) | 82.96-2.90 (2.100-2.90) |
| $R_{sym}$ (%)* | 5.8 (44.6) | 7.0 (52.9) | 3.3 (31.1) | 9.4 (38.3) | 8.9 (38.5) | 9.0 (38.1) |
| $I/\sigma I$* | 18.9 (3.35) | 31.1 (3.1) | 29.8 (3.8) | 17.0 (4.4) | 14.5 (3.7) | 14.5 (3.6) |
| Completeness (%)* | 96.6 (98.3) | 98.7 (91.0) | 99.4 (98.5) | 99.6 (96.7) | 99.5 (99.3) | 99.3 (96.4) |
| Redundancy* | 4.4 (4.4) | 19.8 (6.5) | 6.1 (5.4) | 5.7 (5.3) | 3.8 (3.7) | 3.8 (3.7) |
| **Refinement** | | | | | | |
| Resolution (Å) | 19.70-2.33 | 69.4-1.80 | 19.60-2.70 | | | |
| No. reflections | 9974 | 43284 | 7798 | | | |
| $R_{work}/R_{free}$ | 0.218/0.247 | 0.184/0.216 | 0.249/0.274 | | | |
| No. atoms | | | | | | |
| Protein | 1319 | 2995 | 1417 | | | |
| RNA | 313 | 642 | 321 | | | |
| Water/ligands | 33 | 413 | 5 | | | |
| B-factors | | | | | | |
| Protein | 50.7 | 29 | 102.2 | | | |
| RNA | 108.5 | 33.3 | 103.8 | | | |
| Water/ligands | 40.9 | 35 | 77.7 | | | |
| R.m.s. deviations | | | | | | |
| Bond lengths (Å) | 0.004 | 0.006 | 0.003 | | | |
| Bond angles (˚) | 0.8 | 1.1 | 0.7 | | | |
| Ramachandran plot (%) | | | | | | |
| Preferred region | 95.6 | 96.1 | 95.4 | | | |
| Allowed region | 4.4 | 3.9 | 4.6 | | | |
| Outliers | 0 | 0 | 0 | | | |

*Values in parantheses denote highest resolution shell

**Table 2.1 Data collection, phasing, and refinement statistics.**

## 2.3 Results

### 2.3.1 Csy4 is an endoribonuclease that processes pre-crRNA

To identify the protein (or proteins) responsible for producing crRNAs from pre-crRNAs in Pa14, we tested the six recombinant Cas proteins from Pa14 for endoribonuclease activity and observed sequence-specific pre-crRNA processing with Csy4 (Fig. 2.3A). Csy4 did not cleave pre-crRNA from *Streptococcus thermophilus*, which has a repeat stem-loop of a distinct sequence from Pa14 (Fig. 2.3A). CRISPR transcript cleavage is a rapid, metal ion–independent reaction, as observed for crRNA processing within two other CRISPR/Cas subtypes (Brouns et al., 2008; Carte et al., 2008).

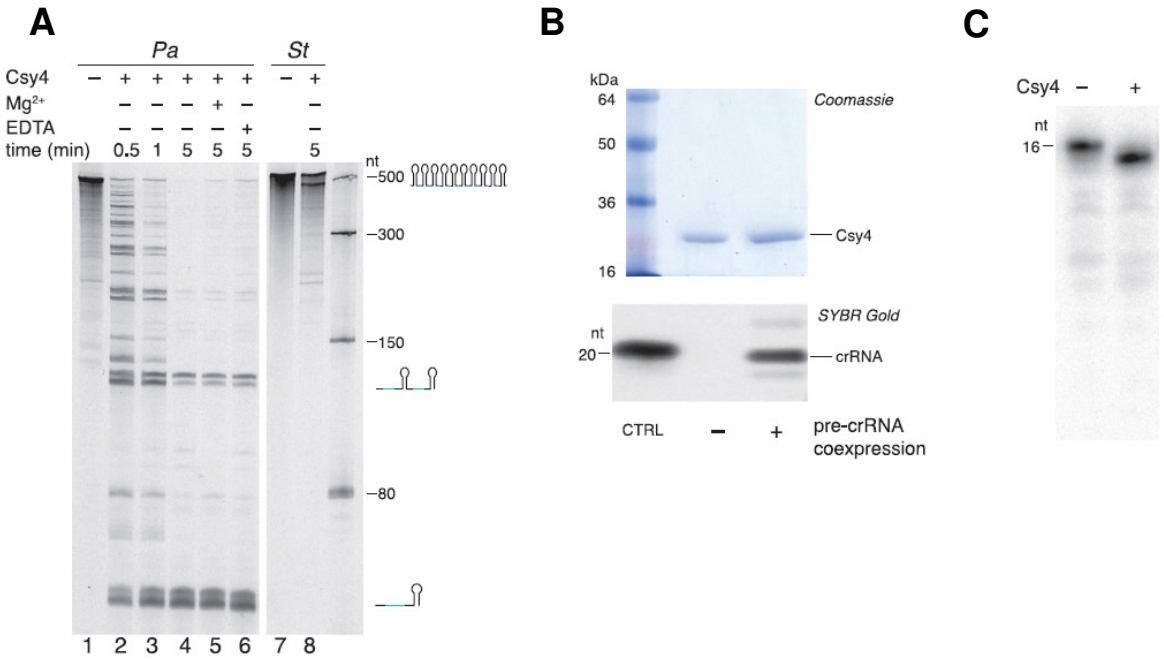



**Figure 2.3 Csy4 specifically cleaves only its cognate pre-crRNA substrate.** (**A**) *In vitro* transcribed Pa14 pre-crRNA (Pa, lanes 1 to 6) was incubated with Csy4 in the absence of exogenous metal ions (lanes 2 to 4) or in the presence of $MgCl_2$ (lane 5) or EDTA (lane 6). *S. thermophilus* (St) pre-crRNA (lanes 7 and 8) served as negative control. Products were separated by means of denaturing polyacrylamide gel electrophoresis (PAGE) and visualized with SYBR Gold staining (Invitrogen, Carlsbad, CA). (**B**) Csy4 was expressed in *E. coli* in the presence (+) or absence (–) of a plasmid expressing a Pa14 CRISPR transcript. Csy4 was affinity purified; copurifying RNA was extracted and analyzed by means of denaturing PAGE and SYBR Gold staining. The ~19-nucleotide RNA corresponds to a protected fragment derived from the CRISPR repeat. (**C**) The Pa14 crRNA stem-loop (C6-G20) was 3′-end labeled by using [5′-32P] pCp, resulting in a minimal Csy4 substrate containing the 32P radiolabel at the position of the scissile phosphate. The RNA was incubated in the presence (+) or absence (–) of Csy4. Products were separated by means of denaturing PAGE and visualized with phosphorimaging.

Csy4 RNA recognition is highly specific for CRISPR-derived transcripts. When expressed in *Escherichia coli* together with a synthetic Pa14 CRISPR RNA consisting of eight repeat sequences (derived from the CRISPR locus proximal to the Cas1 ORF) and seven identical spacer sequences, Csy4 co-purified with a protected ~19-nucleotide fragment derived from the Pa14 crRNA repeat (Fig. 2.3B and Fig. 2.4). To explore the protein/RNA interactions required for Csy4 substrate recognition and cleavage, assays were performed *in vitro* by using RNA oligonucleotides corresponding to different regions of the 28-nucleotide Pa14 CRISPR repeat sequence. A 16-nucleotide minimal RNA fragment, consisting of the repeat-derived stem-loop and one downstream nucleotide, was sufficient for Csy4-catalyzed cleavage (Fig. 2.5A). Csy4-mediated cleavage resulted in products carrying 5′-hydroxyl and 3′-phosphate (or 2′-3′ cyclic phosphate) groups, respectively (Fig. 2.3C). Csy4 activity required the presence of a 2′-hydroxyl group in the nucleotide immediately upstream of the cleavage site because 2′-deoxyribonucleotide substitution at this position abrogated cleavage but did not disrupt Csy4 binding (Fig. 2.5). Additionally, substitution of a 2′-O-methyl nucleotide at the position upstream of cleavage blocked catalysis (Fig. 2.6).

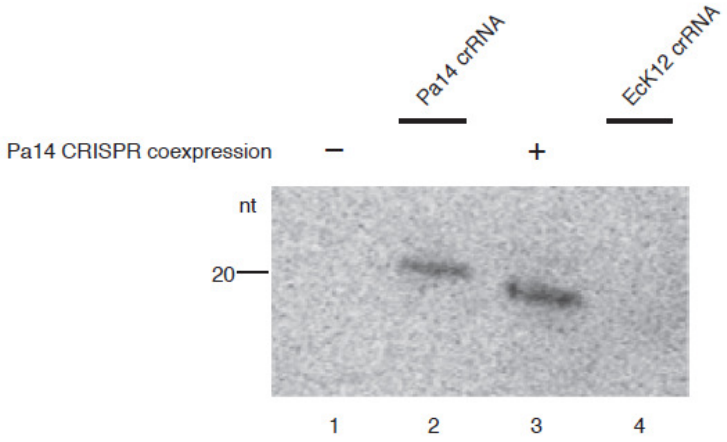



**Figure 2.4 Csy4 co-purifies with a sequence derived from its cognate CRISPR transcript.** Csy4 was heterologously expressed in *E. coli* in the absence (–) or presence (+) of a plasmid expressing a Pa14 CRISPR transcript. Nucleic acids co-purifying with Csy4 were isolated by phenol:chloroform extraction and probed by Northern blotting using an oligonucleotide probe complementary to the crRNA repeat stem-loop sequence. Synthetic RNA oligonucleotides corresponding to the Pa14 (lane 2) and *E. coli* K12 (EcK12, lane 4) crRNA repeat hairpins served as positive and negative controls, respectively.

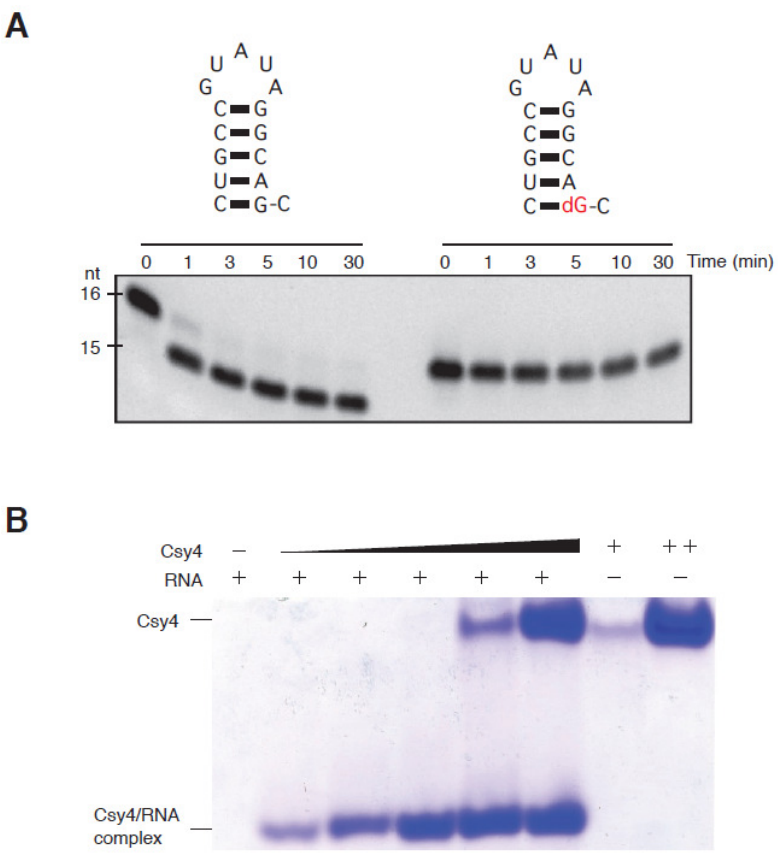



**Figure 2.5 2′-deoxy substitution upstream of the scissile phosphate inhibits Csy4-catalyzed cleavage.** (**A**) Csy4 (200 pmol) was incubated in a 10 µl reaction with ~185 pmol of a 16-nucleotide RNA corresponding to the stem-loop and single downstream nucleotide (left) and a 16-nucleotide RNA with a 2'-deoxyribose at the penultimate nucleotide (right). Products were extracted with acid phenol-chloroform, separated on a denaturing polyacrylamide gel and visualized by SYBR Gold staining. (**B**) Electrophoretic mobility shift assay to evaluate binding of the 2'-deoxyribose-containing substrate from (A) to Csy4. 10 µM RNA was incubated with (left to right) 2.5 µM, 5.0 µM, 10 µM, 20 µM and 40 µM Csy4 and analyzed on a 6% native polyacrylamide gel, followed by Coomassie blue staining. The Csy4-RNA complex migrates rapidly through the gel (lower bands), while unbound Csy4 migrates slower (upper bands). (+) and (++) are 2.5 µM and 40 µM Csy4.


**Figure 2.6 A 2′-O-methyl-substituted RNA nucleotide upstream of the scissile phosphate abrogates cleavage.** Csy4 (75 pmol) was incubated in a 10 µl with 50 pmol of a 16-nt RNA comprising the stem-loop and a single downstream nucleotide with a 2′-O-methyl substituted nucleotide immediately upstream of the cleavage site for 5 min at 25 ˚C. Products were extracted with acid phenol-chloroform, separated on a 15% denaturing polyacrylamide gel and visualized by SYBR Gold staining (lane 2). A no protein control is shown in lane 1.

### 2.3.2 Csy4/substrate RNA co-crystal structure

We co-crystallized Csy4 in complex with the non-cleavable 16-nucleotide minimal RNA substrate in three distinct crystal forms, one containing wild-type Csy4 and two containing a catalytically active point mutant (S22C) of Csy4 (Fig. 2.7 and Fig. 2.8), and

solved their structures to a resolution of 2.3, 2.7, and 1.8 Å, respectively (Table 2.1). In all three structures, the RNA binds to Csy4 in an almost identical manner, in which the protein makes extensive interactions with the single-stranded RNA (ssRNA)–double-stranded RNA (dsRNA) junction at the base of the crRNA stem as well as with the major groove of the RNA hairpin (Fig. 2.9A). The RNA is clamped in a highly basic groove between the main body of the protein and an arginine-rich helix (α3, residues 108 to 120) that inserts into the major groove of the hairpin (Fig. 2.9B).

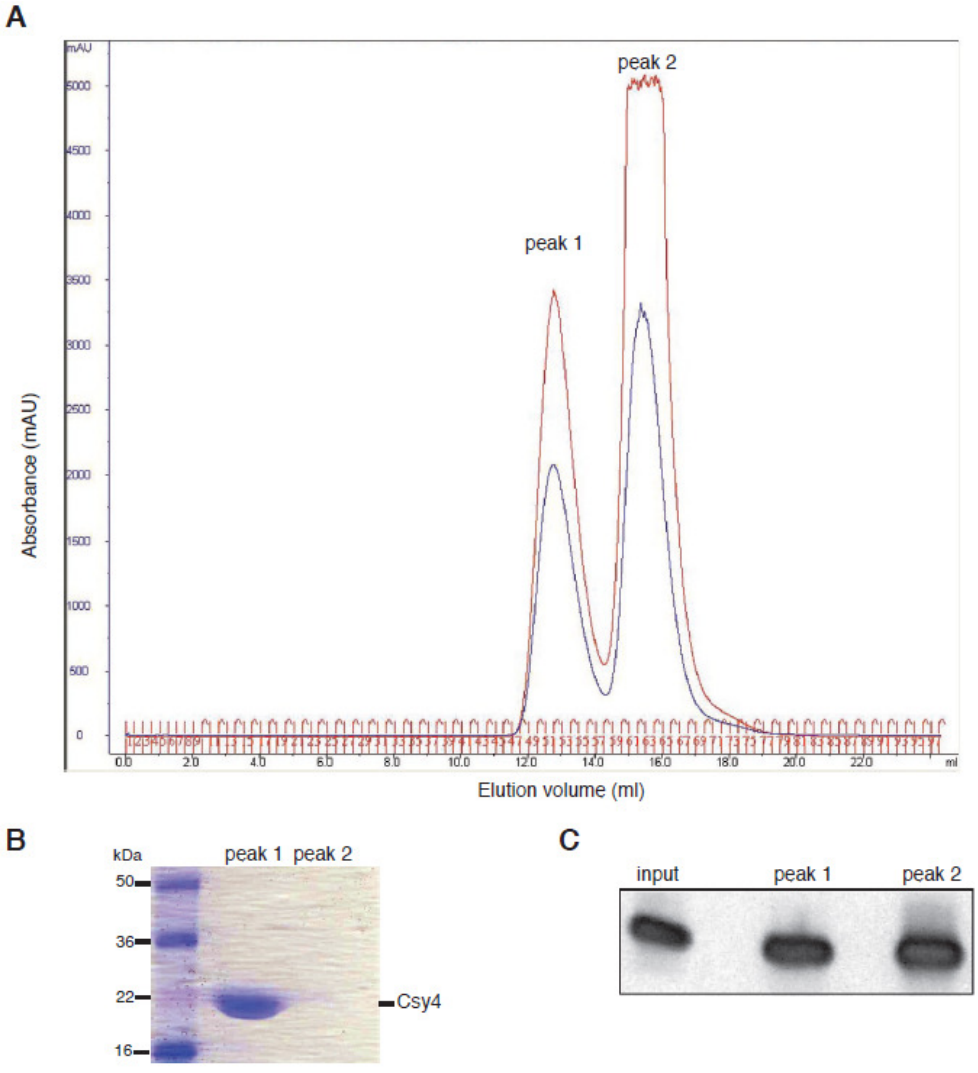



**Figure 2.7 Csy4/RNA complex reconstitution for crystallographic analysis.** (**A**) Csy4 and the minimal 16-nucleotide non-cleavable RNA (Fig. 2.5A, right) were mixed together in a 1:2 molar ratio and incubated at 30 ˚C for 30 min. The complex was separated from free RNA by size exclusion chromatography on a Superdex 75 10/300 column. Absorbance traces at 280 nm and 260 nm are denoted in blue and red, respectively. (**B**) Protein content of the two peaks was analyzed by SDS-PAGE and visualized with Coomassie blue staining. (**C**) Samples from both peaks were phenol- chloroform extracted, separated using urea denaturing PAGE and visualized by staining with SYBR Gold.

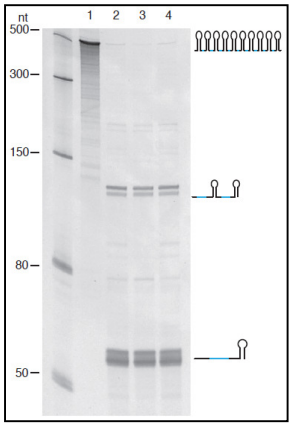



**Figure 2.8 Csy4(S22C) retains wild-type cleavage activity.** 75 pmol of Csy4(S22C) (lanes 2-4) were incubated with 5 pmol of *in vitro* transcribed Pa14 pre-crRNA (lanes 1-4) in a 10 µl reaction for 5 minutes in a buffer containing no exogenous metal ions (lane 2) or buffers supplemented with 2.5 mM $MgCl_2$ (lane 3) or 2.5 mM EDTA (lane 4). Cleavage products were extracted with acid phenol-chloroform, separated by urea denaturing PAGE and visualized by SYBR Gold staining.

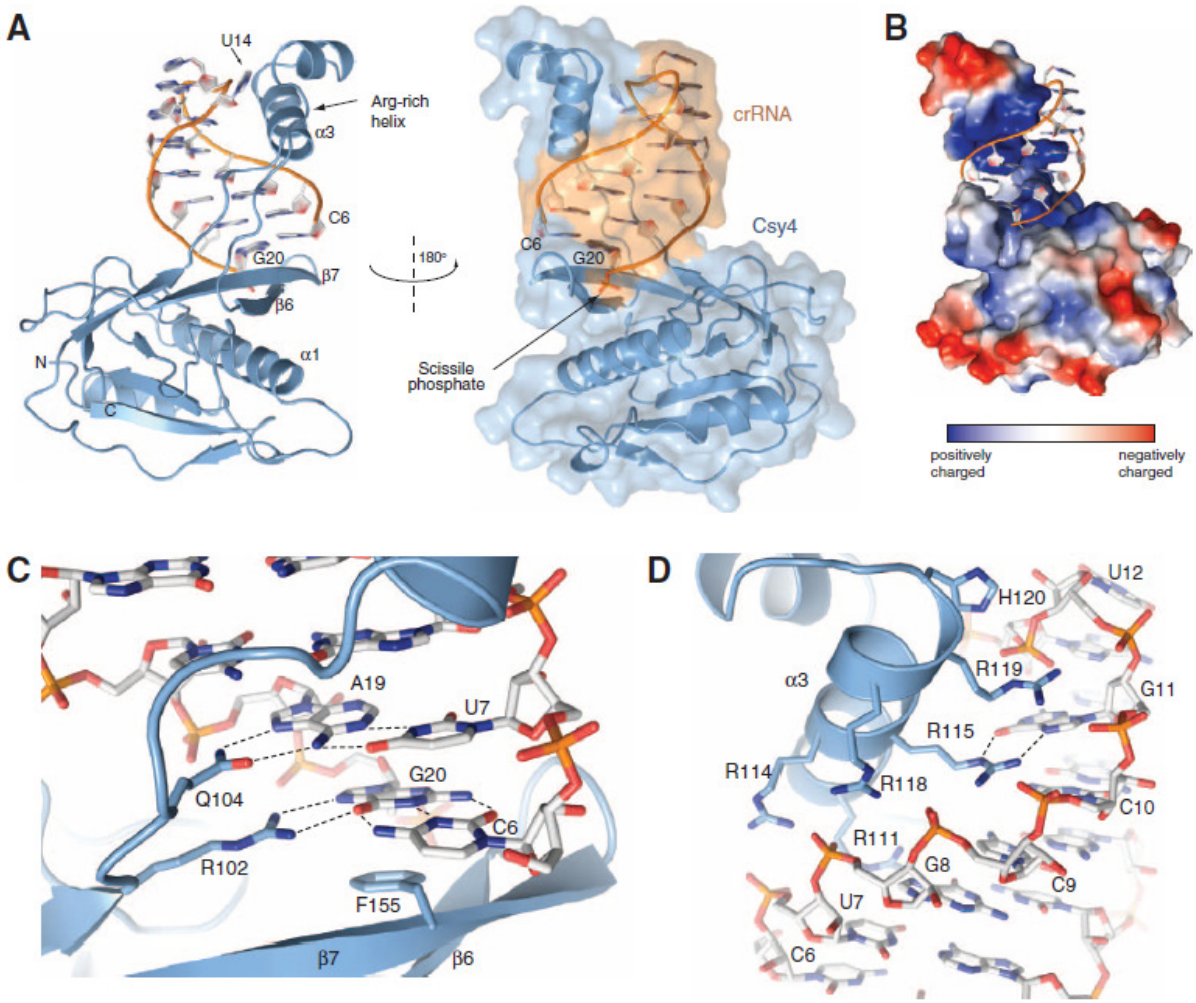



**Figure 2.9 The crystal structure of Csy4 bound to RNA substrate.** (**A**) Front and back views of the complex. Csy4 is colored in blue, and the RNA backbone is colored in orange. (**B**) Csy4 is shown as a surface representation colored according to electrostatic potential [in the same orientation as in (A), right]. The RNA is shown in ribbon representation and colored orange. (**C**) Magnified view of the interactions between Csy4 and the major groove of the RNA hairpin. Hydrogen bonding is depicted with dashed lines. (**D**) Expanded view of the interactions between the arginine-rich helix α3 (blue) and the RNA phosphate backbone (shown in stick format, orange).

In the complex, Csy4 adopts a two-domain architecture consisting of an N-terminal ferredoxin-like domain (residues 1 to 94) and a C-terminal domain (residues 95 to 187) that mediates most of the interactions with the RNA (Fig. 2.10A). At the sequence level, Csy4 shares less than 10% identity with the two other known endoribonucleases involved in crRNA biogenesis, CasE from *Thermus thermophilus* (Ebihara et al., 2006) and Cas6 from *Pyrococcus furiosus* (Carte et al., 2008). The crystal structures of CasE and Cas6 in their nucleic acid–free states showed that these proteins possess a duplicated ferredoxin fold. The N-terminal ferredoxin fold is preserved in Csy4; structural superpositions made by using the DALI server (Holm and Sander, 1993) indicate that Csy4 in its RNA-binding conformation superimposes with CasE and Cas6 with root-mean-square deviation (RMSD) of 3.8 Å (over the N-terminal 111 Cα atoms) and 3.9 Å (over 104 Cα atoms), respectively. Although the C-terminal domain of Csy4 (residues 95 to 187) shares the same secondary structure connectivity as a ferredoxin- like fold, its conformation is markedly different, possibly as a result of RNA binding (Fig 2.10B).

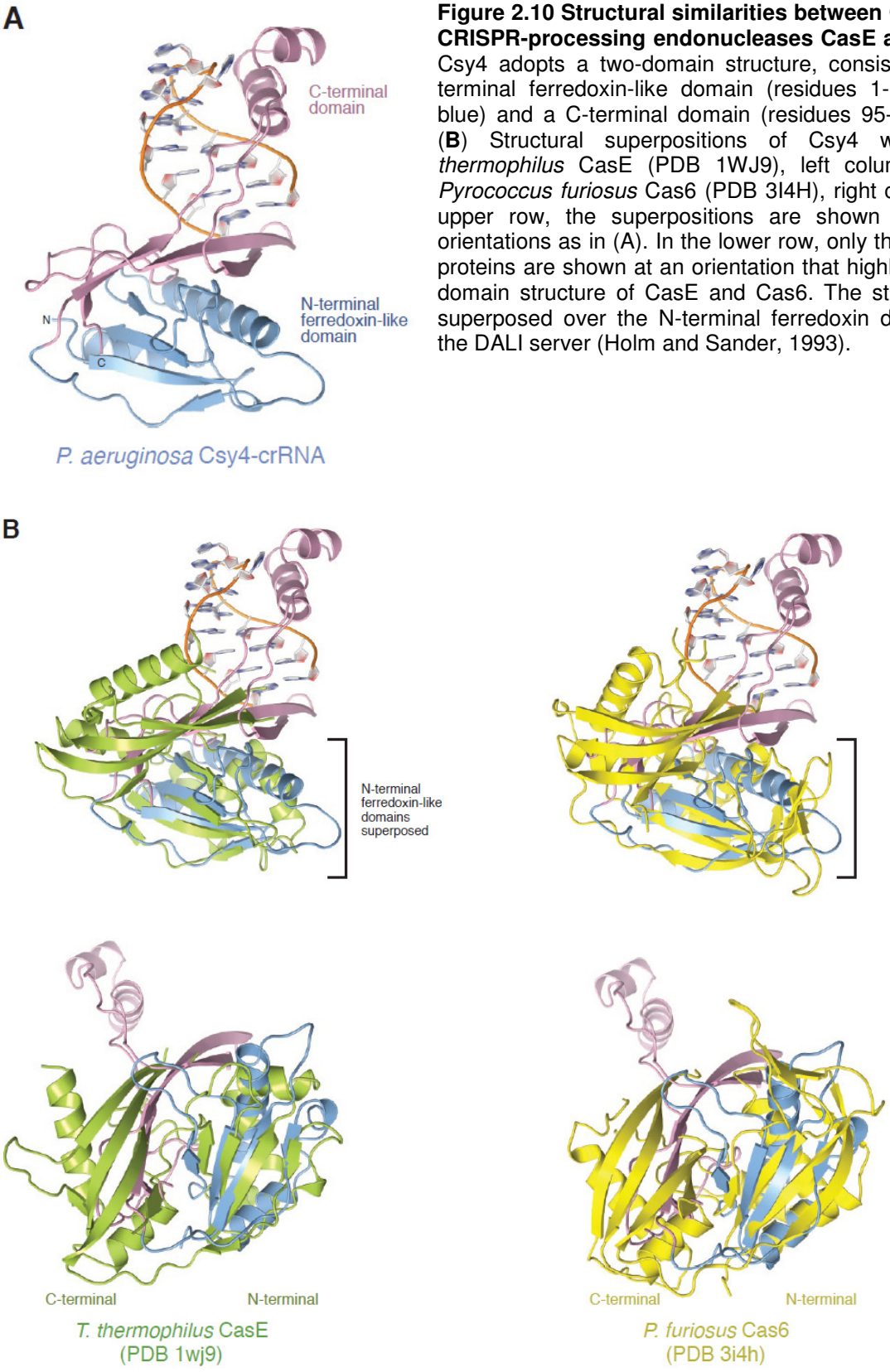



**Figure 2.10 Structural similarities between Csy4 and the CRISPR-processing endonucleases CasE and Cas6.** (**A**) Csy4 adopts a two-domain structure, consisting of an N-terminal ferredoxin-like domain (residues 1-94, shown in blue) and a C-terminal domain (residues 95-187, in pink). (**B**) Structural superpositions of Csy4 with *Thermus thermophilus* CasE (PDB 1WJ9), left column, and with *Pyrococcus furiosus* Cas6 (PDB 3I4H), right column. In the upper row, the superpositions are shown in the same orientations as in (A). In the lower row, only the superposed proteins are shown at an orientation that highlights the two-domain structure of CasE and Cas6. The structures were superposed over the N-terminal ferredoxin domains using the DALI server (Holm and Sander, 1993).

The crRNA substrate forms a stem-loop structure (Kunin et al., 2007). Nucleotides 6 to 10 and 16 to 20 basepair to produce a regular A-form helical stem. The GUAUA pentaloop contains a sheared G11-A15 base pair and an extruded nucleotide U14, which closely resembles the structures adopted by other GNR(N)A pentaloops (Huppler et al., 2002; Legault et al., 1998). In the Csy4-RNA complex, the RNA stem loop straddles the β-hairpin formed by strands β6- β7 of Csy4, with the C6-G20 base pair directly stacking onto the aromatic side chain of Phe155 (Fig. 2.9C). In the context of the full-length CRISPR transcript, this allows Csy4 to recognize the ssRNA-dsRNA junctions in the pre-crRNA and anchor the RNA stem-loop to permit sequence-specific interactions in the major groove.

Arg102 and Gln104, located in a linker segment connecting the body of Csy4 to the arginine-rich helix, make sequence-specific hydrogen-bonding contacts in the major groove of the RNA stem to nucleotides G20 and A19, respectively (Fig. 2.9C). The Csy4-crRNA interaction is further stabilized by the insertion of the arginine-rich helix into the major groove of the RNA hairpin near the pentaloop (Fig. 2.9D). The side chains of Arg114, Arg115, Arg118, Arg119, and His120 contact the phosphate groups of nucleotides 7 to 12. Additionally, the sidechain of Arg115 hydrogen-bonds to the base of G11. The binding of the arginine-rich helix to the major groove of the crRNA hairpin is reminiscent of the N-peptide/boxB RNA interaction in lambdoid phages (Fig. 2.11) (Cai et al., 1998) and of lentiviral Rev-RRE and Tat-TAR complexes (Anand et al., 2008; Ye et al., 1996).

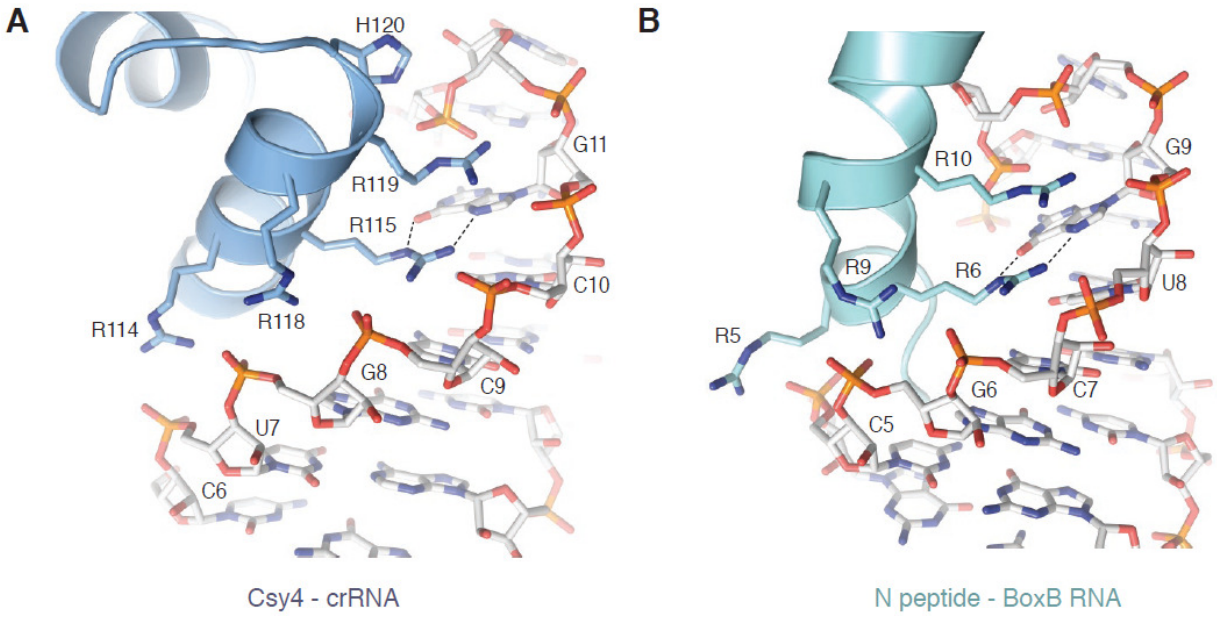



**Figure 2.11 Csy4 and the antiterminator N-peptides of lambdoid bacteriophages utilize similar mechanisms for RNA phosphate backbone recognition.** Comparisons of the arginine-rich helices of (**A**) Csy4 bound to crRNA and (**B**) the antiterminator N-peptide of bacteriophage p22 bound to the boxB hairpin RNA (PDB 1A4T). The structures were superimposed using Pymol (DeLano, 2002) and are shown in identical orientations. Dashed lines indicate hydrogen-bonding interactions.

### 2.3.3 Functional analysis of Csy4 active site

Csy4 recognizes the hairpin element of the CRISPR repeat sequence and cleaves immediately downstream of it. In the Csy4-RNA complex structure, where RNA cleavage is abrogated by a 2′-deoxy modification in nucleotide G20, ordered electron density is only evident for the scissile phosphate between G20 and C21. The ribose and cytosine moieties of C21 are not resolved and presumably disordered. We confirmed that the ribose and cytosine moieties of C21 were still present in the crystal by harvesting ~10 of the rod-shaped crystals, washing them, and phenol:chloroform extracting the nucleic acid. Separation on a denaturing gel confirmed the presence of intact substrate RNA (Fig. 2.12). The scissile phosphate binds in a pocket located between the β6- β7 hairpin turn on one side and helix α1 on the other (Fig. 2.13A), hydrogen-bonding to the backbone amide of Gln149 and the side chain of His29. Ser148 is adjacent to the 2′ ribose carbon atom of nucleotide G20 (4.6 Å) and may make a hydrogen-bonding interaction with the 2′-hydroxyl group of G20 in a bona fide pre-crRNA substrate. Mutation of the strictly conserved His29 or Ser148 (to alanine and cysteine, respectively) abolished cleavage activity without disrupting RNA binding (Fig. 2.13B, Fig. 2.14, and Fig. 2.15), suggesting that these two residues participate in catalysis. Surprisingly, the S148C mutant forms two shifted species with crRNA at higher protein concentrations in an electrophoretic mobility shift assay, potentially indicative of an additional RNA/protein complex morphology (Fig. 2.15). A strongly conserved tyrosine (Tyr176) is also positioned near the scissile phosphate (Fig. 2.13A). However, mutation of Tyr176 to phenylalanine had only a minimal effect on activity (Fig. 2.13B).

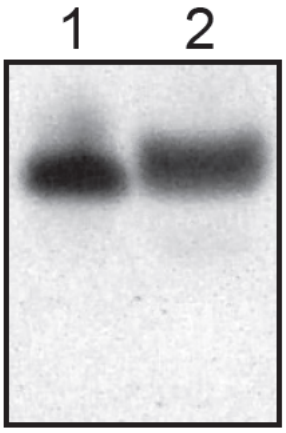



**Figure 2.12 Nucleic acid content of the rod-shaped crystals.** RNA was phenol:chloroform extracted from the rod-shaped crystals and analyzed on a 15% denaturing acrylamide gel (lane 2). The synthetic RNA used to form the substrate complex is shown in lane 1.

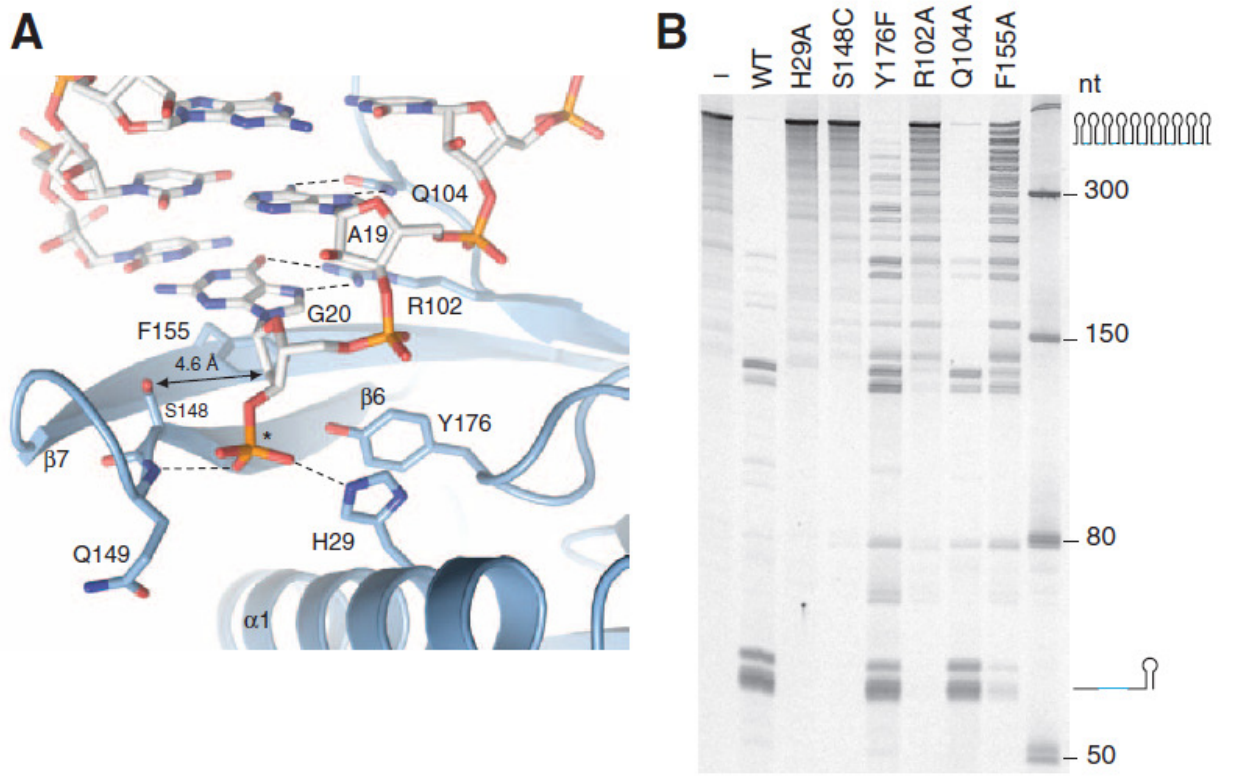



**Figure 2.13 Functional analysis of catalytic residues in Csy4.** (**A**) Detailed view of the catalytic center. Only the phosphate group of nucleotide C21 (the scissile phosphate, indicated with an asterisk) is visible in electron density maps. Strictly conserved residues found in proximity of the scissile phosphate are shown in stick format. The arrow indicates the distance between the hydroxyl group of Ser148 and the 2′ ribose carbon of G20. (**B**) Cleavage activity of Csy4. Wild-type (WT) Csy4 and a series of single-point mutants were incubated with *in vitro* transcribed pre-crRNA for 5 min at 25°C. Products were resolved by means of denaturing PAGE and visualized with SYBR Gold staining.

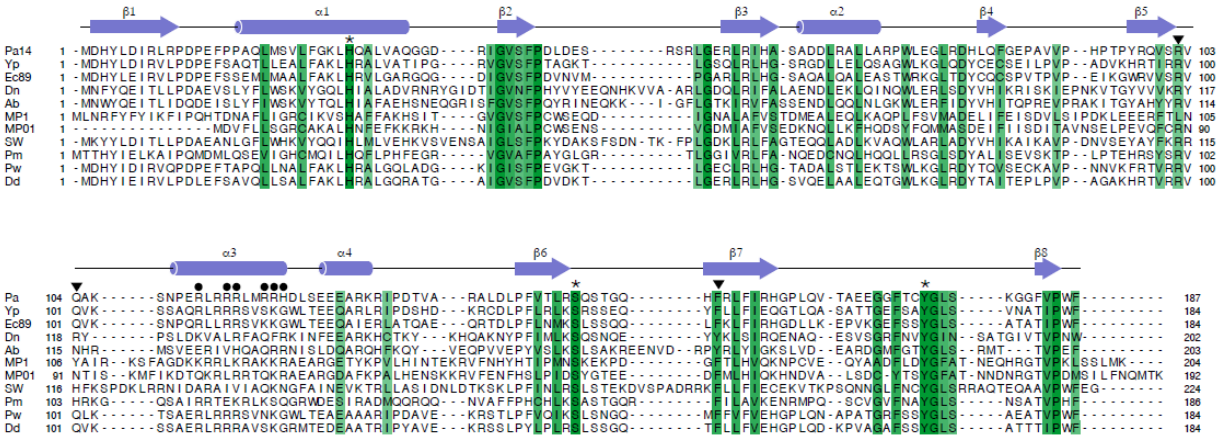


**Figure 2.14 Evolutionary conservation of functional residues in Csy4.** Multiple sequence alignment of 11 Csy4 homologues from *Pseudomonas aeruginosa* UCBPP-PA14 (Pa14), *Yersinia pestis* AAM85295 (Yp), *Escherichia coli* UTI89 (Ec89), *Dichelobacter nodosus* VCS1703A (Dn), *Acinetobacter baumannii* AB0057 (Ab), *Moritella sp.* PE36 (MP1, MP01), *Shewanella sp.* W3-18-1 (SW), *Pasteurella multocida subsp. multocida* Pm70 (Pm), *Pectobacterium wasabiae* (Pw), and *Dickeya dadantii* Ech703 (Dd). The Csy4 candidates were identified using PSI-BLAST (Altschul et al., 1997) and aligned with CLUSTALW (Thompson et al., 1994). The presence of a nearby CRISPR locus in the genome was verified by manual inspection of the CRISPRdb database (Grissa et al., 2007). Invariant and significantly conserved amino acids are highlighted in dark green and light green, respectively. Asterisks indicate active site residues; inverted triangles indicate residues that contact the bases of the RNA; circles indicate residues contacting the phosphate backbone. The secondary structure of Csy4 is shown above the sequence.

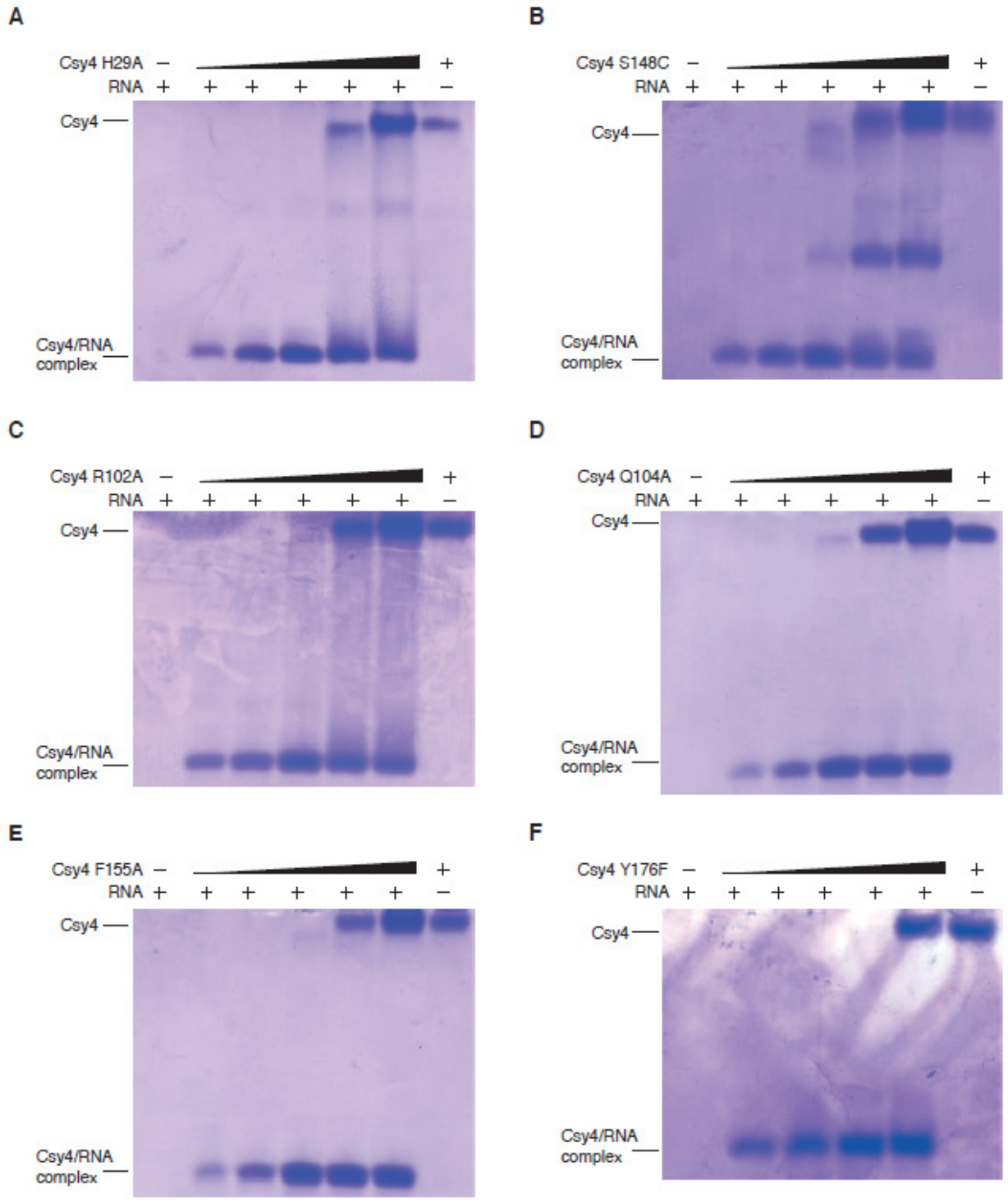



**Figure 2.15 Electrophoretic mobility shift assay to evaluate binding of the six Csy4 point mutants from Fig. 2.13B to a 28-nucleotide oligonucleotide consisting of the Pa14 CRISPR repeat sequence.** 10 uM oligonucleotide was incubated with (left to right) 2.5 uM, 5.0 uM, 10 uM, 20 uM and 40 uM **(A)** Csy4H29A, **(B)** Csy4S148C, **(C)** Csy4R102A, **(D)** Csy4Q104A, **(E)** Csy4F155A or **(F)** Csy4Y176F and analyzed on a 6% native polyacrylamide gel, followed by staining with Coomassie blue. (+) denotes 10 uM Csy4.

The requirement for a 2′ hydroxyl group in the nucleotide immediately preceding the cleavage site suggests that the catalytic mechanism of Csy4 may proceed through a 2′-3′ cyclic intermediate. In this context, the observation of an invariant serine residue (Ser148) adjacent to the 2′ ribose position upstream of the scissile phosphate is unprecedented and points to Ser148 playing a role in activating and/or positioning the 2′-hydroxyl for a nucleophilic attack on the scissile phosphate. The other functionally critical active site residue, His29, may act as a proton donor for the 5′-hydroxyl–leaving group because mutation of His29 to lysine partially preserved catalytic activity (Fig. 2.16).

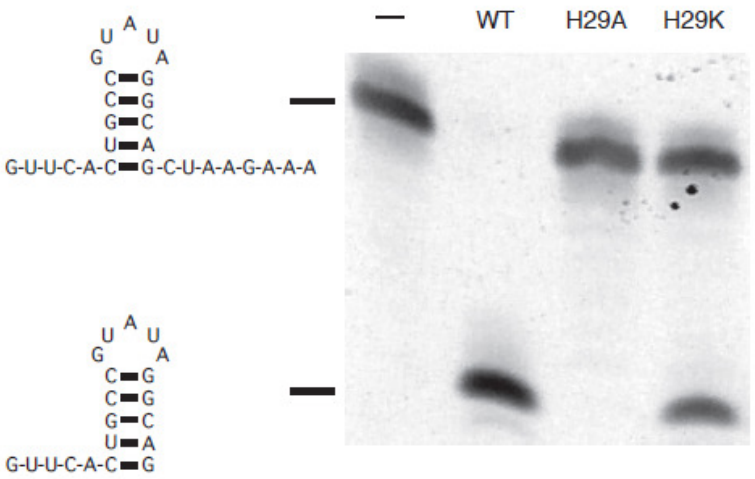



**Figure 2.16 Lysine substitution of the catalytic His29 partially preserves catalytic activity.** WT or mutant Csy4 (200 pmol) were incubated with 100 pmol of a 28-nucleotide RNA substrate corresponding to the Pa14 crRNA repeat element for 10 minutes at 25 ˚C in a 10 µl reaction volume. The products were extracted, separated by denaturing PAGE and visualized by SYBR Gold staining. Only the 20-nucleotide product is shown.

We next tested the functional importance of Csy4 residues involved in crRNA recognition. Alanine substitution of Arg102 abolished pre-crRNA processing *in vitro*, whereas mutation of Gln104 to alanine did not substantially disrupt activity (Fig. 2.13B). Mutation of Phe155 to alanine severely impaired crRNA processing, suggesting that this residue also plays an important role in substrate orientation. However, none of the above mutations severely disrupted crRNA binding, as judged by means of electrophoretic mobility shift assays, indicating that the structural integrity of the mutant proteins was not compromised (Fig. 2.15). Thus, interaction between Csy4 and the closing base pair of the RNA stem is critical for pre-crRNA processing, whereas sequence-specific recognition of the penultimate base pair in the stem is less important. Incubation of Csy4 with a panel of short RNA oligonucleotides containing a variety of mutations in the CRISPR repeat stem-loop sequence further confirmed that Csy4 requires a C-G base pair closing the RNA stem and that Csy4 can accommodate different nucleotides at the penultimate RNA base pair (Fig. 2.17).

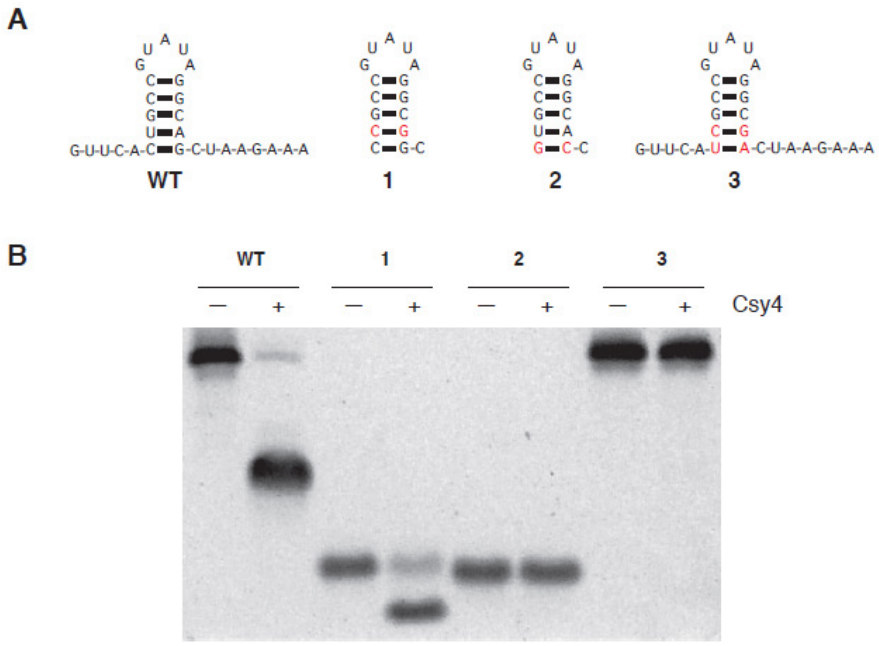



**Figure 2.17 The C6-G20 base pair is critical for pre-crRNA processing by Csy4.** (**A**) Four short (28- or 16-nucleotide) RNA oligonucleotides corresponding to the wild-type (WT) Pa14 CRISPR repeat sequence and three mutant versions. Deviations from WT sequence are marked in red. (**B**) Csy4 (75 pmol) was incubated with each RNA substrate (50 pmol) indicated in (A) for 5 minutes at 25 ˚C in a 10 µl reaction containing 20 mM HEPES pH 7.5 and 100 mM KCl. Cleavage products were acid phenol-chloroform extracted and analyzed by denaturing PAGE and staining with SYBR Gold.

## 2.4 Preliminary results

*Escherichia coli* UTI89 (Ec89), a pathogenic strain of *E. coli* that causes urinary tract infections, contains a CRISPR/Cas system highly related to the CRISPR/Cas locus found in *P. aeruginosa* UCBPP-PA14 (Grissa et al., 2007). Like Pa14, Ec89 has two CRISPR loci that flank the six *cas* genes (Fig. 2.1). The repeat sequences from both loci are identical, and contain only two differences from the Pa14 CRISPR repeat described above. The Ec89 CRISPR repeat: GTTCACTGCCGTA**C**AGGCAGCT**T**AGAAA. The two differences between the Ec89 and Pa14 repeats (bold) are located in single-stranded regions and are unlikely to affect the ability of Csy4 process this substrate. The nucleotides that base pair to form the stem are underlined. The Csy4 protein from Ec89 is 48% identical and 63% similar to Pa14Csy4.

We hypothesized that based on the sequence similarity of the Csy4 protein and the cognate CRISPR repeats, Ec89Csy4 would be able to process pre-crRNA from Pa14. We expressed and purified Ec89Csy4 and incubated it with Pa14 pre-crRNA *in vitro*. Like Pa14Csy4, it is a metal-independent enzyme and it is able to process pre-crRNA into mature ~60-nt crRNAs (Fig. 2.18A). We generated two predicted active site mutants (alanine and lysine substitutions of His29) and tested their activity *in vitro* with Pa14 pre-crRNA (Fig 2.18B). Alanine substitution of His29 abolished activity, whereas lysine substitution partially recovered processing activity.

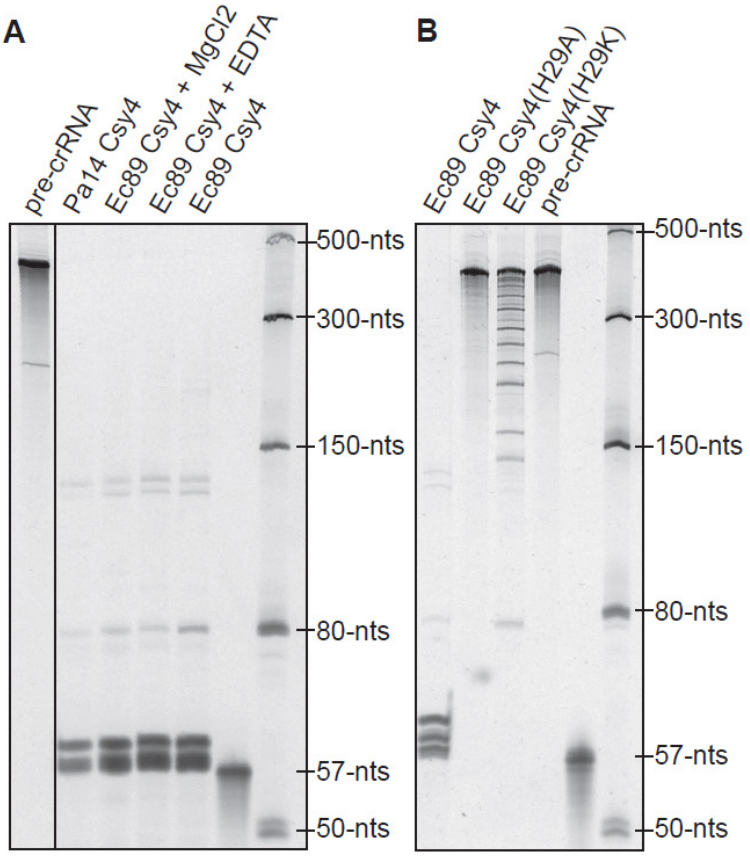



**Figure 2.18 Ec89Csy4 can process pre-crRNA from Pa14.** (**A**) 75pmol of Pa14 or Ec89 Csy4 were incubated with 5 pmol *in vitro* transcribed Pa14 pre-crRNA in a 10 µl reaction for 30 min at 25 ˚C in buffer containing 20mM HEPES pH 7.5, 100mM potassium chloride, and 2.5mM $MgCl_2$ or EDTA, as noted. Products were phenol:chloroform extracted, separated on a 15% denaturing gel, and visualized with SYBR Gold staining. (**B**) 75pmol of wild-type or mutant Ec89Csy4 were incubated with 5 pmol *in vitro* transcribed Pa14 pre-crRNA in a 10 µl reaction for 30 min at 25 ˚C in buffer containing 20mM HEPES pH 7.5 and 100mM potassium chloride. Products were phenol:chloroform extracted, separated on a 15% denaturing gel, and visualized with SYBR Gold staining.

In order to evaluate whether Ec89Csy4 can complement Pa14Csy4 activity *in vivo*, we sent plasmid constructs for Ec89Csy4 and Ec89Csy4(H29A) to the O'Toole lab at Dartmouth Medical School. Kyle Cady, a graduate student there, evaluated mature crRNA levels *in vivo* in a DMS3 lysogen of *P. aeruginosa* UCBPP-PA14 using Northern blotting (Fig. 2.19, (Cady and O'Toole, 2011)). In the wild-type lysogen, mature crRNAs can be detected and knocking out *csy4* abolishes crRNA processing *in vivo*. Complementation of Pa14Csy4, Pa14Csy4(H29K), or Ec89Csy4 on a plasmid restores crRNA processing. As expected, the active site alanine substitution mutants (Pa14H29A and Ec89H29A) cannot restore crRNA processing.

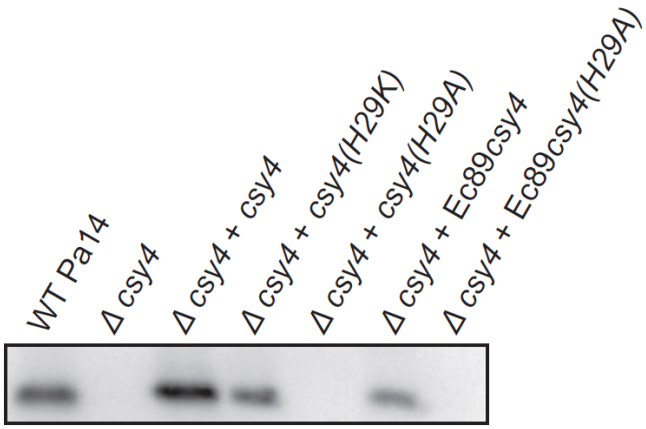



**Figure 2.19 Ec89Csy4 can complement Pa14Csy4 *in vivo*.** Northern blot for mature crRNAs expressed in a DMS3 lysogen of *P. aeruginosa* UCBPP-PA14. Adapted from (Cady and O'Toole, 2011).

## 2.5 Discussion

Phylogenetic analysis of CRISPR loci suggests that CRISPR repeat sequences and structures have co-evolved with the *cas* genes (Kunin et al., 2007). The similarity of Csy4 at the fold level to the CRISPR-processing endonucleases CasE and Cas6 suggests that collectively they are likely to have descended from a single ancestral endoribonuclease enzyme that has diverged throughout evolution. The structure described here reveals how Csy4 and related endonucleases from the same CRISPR/Cas subfamily use an exquisite recognition mechanism to discriminate crRNA substrates from other cellular RNAs. This illustrates the importance of co-evolution in shaping molecular recognition mechanisms in the CRISPR pathway. Furthermore, the ability of Csy4 to form a tight complex with the cleaved crRNA product points to Csy4 having a functional role within the CRISPR pathway that extends beyond pre-crRNA cleavage.

# Chapter 3

# Csy4 cleavage mechanism

*A portion of the work presented in this chapter has been previously published as part of the following paper: Haurwitz, R.E., Sternberg, S.H., Doudna, J.A. (2012). Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.*

*Rachel Haurwitz designed experiments, purified proteins, performed single-turnover cleavage assays, crystallized the complexes, solved the crystal structures, and wrote the manuscript. Samuel Sternberg designed experiments, performed pH-rate profile experiments, and contributed to the manuscript.

## 3.1 Introduction

As described in Chapter 2, Csy4 is a 21.4 kDa protein that recognizes its CRISPR RNA substrate via sequence- and structure-specific contacts. It cleaves cognate RNAs at the 3′ end of a five-base-pair stem-loop, generating crRNAs comprising a unique spacer sequence flanked by 8 and 20 repeat-derived nucleotides on the 5′ and 3′ ends, respectively. Csy4 has equally tight affinity for both its substrate pre-crRNA and product crRNA, binding both with a 50 pM equilibrium dissociation constant (Sternberg et al., 2012). A single mature crRNA and one copy of Csy4 are components of the large ribonucleoprotein (RNP) Csy targeting complex (Wiedenheft et al., 2011b), but the mechanism of Csy complex assembly is currently unknown.

RNA cleavage by Csy4 is divalent metal ion-independent and requires chemical activation of a ribosyl 2′-hydroxyl for internal nucleophilic attack on the phosphodiester bond (Haurwitz et al., 2010). In the crystal structures of Csy4 bound to substrate RNA presented in Chapter 2, we used a construct lacking the 2′-hydroxyl nucleophile upstream of the scissile phosphate to abrogate cleavage. The structures revealed three active site-proximal residues: Ser148, His29, and Tyr176 (Fig. 3.1A). crRNA biogenesis was strongly inhibited by S148C and H29A mutations, while a Y176F mutation exhibited near wild-type activity. This mutational analysis led us to speculate that Ser148 plays a role in activating and/or positioning the 2′-hydroxyl for nucleophilic attack because it is located in close proximity to the 2′ carbon. Based on structural and biochemical evidence, we hypothesized that His29 may act as a proton donor for the 5′-hydroxyl leaving group because mutation of His29 to lysine partially preserved catalytic activity.

We used biochemical and structural methods to investigate the chemical mechanism of Csy4-catalyzed CRISPR RNA cleavage. Three crystal structures of wild-type and mutant Csy4 bound to product RNAs, coupled with kinetic analyses of mutant Csy4 cleavage rates, suggest a substrate positioning and cleavage mechanism in which Ser148 holds the 2′-hydroxyl nucleophile in place and His29 deprotonates it for attack on the scissile phosphate. The lack of both a general acid and positively charged residues in the active site explains the observed rate constants that are $10^3$- to $10^4$-fold slower relative to other metal ion-independent ribonucleases. We additionally demonstrate that CRISPR transcript processing by Csy4 is essential for subsequent formation of the Csy complex *in vivo*. Given the essential role Csy4 plays in formation of this targeting complex, slow cleavage rates in conjunction with highly accurate substrate selection likely ensure that cognate pre-crRNA substrates are cleaved with little to no off-target activity on other cellular RNAs.

## 3.2 Methods

### 3.2.1 Protein expression and purification

Csy4 and single point mutants were expressed and purified as described in Chapter 2 with minor exceptions. Briefly, $His_6$-MBP-Csy4 or $His_6$-Csy4 fusion constructs (vectors pHMGWA and pHGWA, respectively (Busso et al., 2005)) were expressed in either *E. coli* BL21(DE3) cells or *E. coli* Rosetta 2(DE3) cells (Novagen). Following batch nickel resin affinity purification, cleavage with TEV protease, and a second nickel resin step, samples were separated on a single Superdex75 (16/60) size exclusion column (GE Healthcare) in 100mM HEPES pH 7.5, 500mM potassium chloride, 5%

glycerol, and 1mM TCEP. Proteins were then dialyzed against 100mM HEPES pH 7.5, 150mM potassium chloride, 5% glycerol, and 1mM TCEP; concentrated; and stored at -80 ˚C.

### 3.2.2 RNA cleavage assays

Single-turnover cleavage experiments were performed at 24 ˚C in 20mM HEPES, 100mM potassium chloride, pH 7.2. Cleavage reactions were carried out in 60 ul volume containing 500 pM [5′-32P]-crRNA repeat (5′-GUUCACUGGCCGUAUAGGCAGCUAAGAAA-3′), 400 nM Csy4, and 72 units RNasin Plus (Promega). At noted time points, 10 ul of the reaction were removed and quenched with 30 ul of acid phenol:chloroform (Ambion). 5 ul of the aqueous layer were mixed with 5 ul of formamide loading buffer and separated on a 15% denaturing polyacrylamide gel in 1X TBE running buffer. Cleaved and uncleaved RNAs were visualized by phosphorimaging and quantified using ImageQuant (GE Healthcare). For each sample, the percentage of RNA cleaved (intensity of cleaved RNA band divided by the sum of the cleaved and uncleaved bands) was plotted as a function of time. Plots were fit to an exponential decay curve using Kaleidagraph (Synergy Software). Rate constants are reported as $k_{obs}$ because the rate-limiting step for cleavage is unknown. All cleavage assays were done in triplicate.

Cleavage reactions for pH-rate profiles were 55 ul in volume, contained 400 nM Csy4 and 500 pM [5′-32P]-crRNA repeat, and were performed in 20mM buffer, 100mM potassium chloride, and 1mM dithiothreitol (DTT). Buffers used were as follows:

pH 4.0–6.5: citric acid;
pH 7.0–8.5: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES);
pH 9.0–9.5: N-cyclohexyl-2-aminoethanesulfonic acid (CHES);
pH 10.0–11.0: N-cyclohexyl-3-aminopropanesulfonic acid (CAPS).

Cleavage data were collected and analyzed as described above. pH-rate plots were fit to the following equation using Kaleidagraph (Synergy Software):

$$k_{obs} = (k_{obs,MAX} \times K_a) \div (K_a + [H^+]),$$

where $K_a$ is an apparent acid dissociation constant and $[H^+]$ is the proton concentration.

### 3.2.3 Crystallization

When co-expressed with a pre-crRNA in *E. coli* BL21 (DE3) cells, Csy4 co-purifies with a ~19-nt protected fragment of the crRNA repeat (Chapter 2). Given that Csy4 cleaves this RNA substrate after nucleotide G20 and that the minimal hairpin is only 15-nts, Csy4 must protect a further 4-nts of ssRNA upstream of the hairpin. To identify the molecular interactions between Csy4 and this piece of ssRNA, we crystallized this Csy4/product RNA complex in 16% PEG4000, 150mM sodium citrate pH 5.0, and 100 mM $MgCl_2$. We obtained beautiful plate- and box-shaped crystals (Fig. 3.1A), but they yielded highly heterogeneous diffraction patterns. We hypothesized that the lack of clear diffraction pattern may result from the heterogeneity of the RNA component of the complex. In order to minimize this heterogeneity, we generated a complex of Csy4(S22C) and a synthetic 20-nucleotide crRNA fragment (5′-UUCACUGCCGUAUAGGCAGC-3′) *in vitro.*

Csy4/RNA complexes were generated and purified as described in Chapter 2. Briefly, an excess of synthetic crRNA fragment was added to Csy4 and the sample was incubated at 30 ˚C for 30 min. For the product complex, this incubation step permitted full cleavage of the substrate RNA into product RNA. The RNA/protein complex was then separated from free RNA via size exclusion chromatography. All crystals were grown at 18 ˚C using the hanging drop vapor diffusion method by mixing equal volumes (1 ul + 1 ul) of protein/RNA sample and reservoir solution. All complexes yielded plate-shaped crystals (Fig. 3.1B) Csy4S22C/product complex crystals were grown in 22% PEG4000, 120mM sodium citrate pH 5.0, and 50mM magnesium chloride. Csy4S148A/ RNA complex crystals were grown in 20% PEG4000, 150mM sodium citrate pH 5.0, and 100mM magnesium chloride. Minimal complex crystals were grown in 21% PEG4000, 180mM sodium citrate pH 5.0, and 100mM magnesium chloride. Crystals were cryoprotected with reservoir solution containing 25% glycerol and flash frozen in liquid nitrogen. Minimal complex crystals were soaked with mother liquor supplemented with 2mM ammonium metavanadate for 1.5 h prior to cryo-protection and flash freezing.

**Hint:** We attempted to crystallize a transition state analog by trapping vanadate in the active site of Csy4. This method has been useful for a variety of RNase (Ladner et al., 1997) and ribozyme (Rupert et al., 2002) crystal structures. We were unable to see any electron density indicative of vanadate in the minimal crystals.

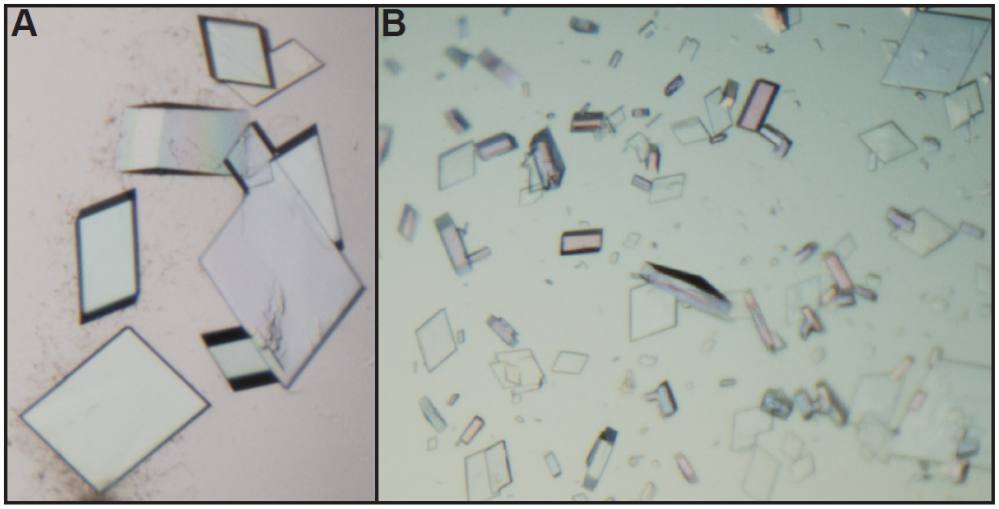



**Figure 3.1** (**A**) Plate- and box-shaped crystals containing wild-type Csy4 that had been co-expressed with pre-crRNA. (**B**) Plate-shaped crystals containing the Csy4(S22C) mutant in complex with product RNA.

### 3.2.4 Structure determination

Diffraction data were collected at beam lines 8.2.1 and 8.3.1 of the Advanced Light Source, Lawrence Berkeley National Laboratory. Datasets were processed in XDS (Kabsch, 2010). All three structures were determined using molecular replacement in Phaser (Collaborative Computational Project, 1994; McCoy et al., 2007). Chains A and C (corresponding to protein and RNA, respectively) from the highest resolution Csy4/substrate complex described in Chapter 2 (PDB ID 2XLK) were used as search

models for the product complex. The Csy4 protein (lacking the arginine-rich helix) and RNA (lacking the A5 nucleotide) models from the product complex were used as search models for the S148A and stem-loop complex structures. The models presented here resulted from iterative rounds of manual rebuilding in COOT (Emsley and Cowtan, 2004) and KiNG (Chen et al., 2009) and refinement in Phenix.refine (Adams et al., 2010). Riding hydrogens were included during refinement. Models were periodically validated using MolProbity (Chen et al., 2010).

All three complexes yielded crystals belonging to the *C*2 monoclinic space group that contained one complex per asymmetric unit. As in one of the substrate structures presented in Chapter 1 (PDB ID 2XLI), the RNA stems from neighboring complexes form coaxially stacked helices via an RNA kissing-loop interaction. The RNA helix and the associated arginine-rich alpha helix sit in a large solvent channel and exhibit elevated B factors. In the 2.0 Å resolution product structure, there is clear density for all amino acids in the arginine-rich helix, whereas in the 2.6 Å S148A structure and the 2.3 Å minimal complex structure, there is no density for the arginine-rich helix.

All structure figures were made using PyMol (DeLano, 2002).

Coordinates and structure factors for the Csy4-crRNA complexes have been deposited in the Protein Data Bank under the accession codes 4AL5, 4AL6, and 4AL7.

### 3.2.5 Csy complex *in vivo* reconstitution

The four Csy proteins (Csy1, Csy2, Csy3, and Csy4) were co-expressed from a polycistronic expression construct in which Csy3 had a $His_6$ fusion tag along with a synthetic CRISPR locus containing eight repeats and seven identical spacers in *E. coli* BL21(DE3) cells as described previously (Wiedenheft et al., 2011b). Site-directed mutagenesis was used to introduce an alanine substitution at position 29 of the *csy4* gene. Briefly, protein expression was induced with addition of 0.5mM IPTG at an optical cell density at 600 nm of ~0.5, followed by shaking at 18 ˚C overnight. Samples were lysed and clarified as previously reported (Wiedenheft et al., 2011b). Samples were affinity purified with nickel NTA resin (Qiagen) and incubated overnight with TEV protease to release the $His_6$ tag. Following a second nickel affinity step, samples were purified on a Superose 6 (10/300) size exclusion column (GE Healthcare) in 20mM HEPES pH 7.5, 100mM potassium chloride, and 1mM TCEP.

### 3.2.6 Csy complex *in vitro* reconstitution

Csy3 was recombinantly expressed as a $His_6$-MBP fusion in *E. coli* BL21(DE3) cells. $His_6$-MBP-Csy1 and untagged Csy2 were co-expressed in *E. coli* BL21(DE3) cells. Both protein samples were subjected to the same purification steps as described above for Csy4. Mature crRNAs were purified from *in vivo* reconstituted Csy complex (see above) by acid phenol:chloroform extraction, chloroform extraction, and ethanol precipitation. Csy1/2, Csy3, Csy4, and crRNA were mixed in 1:6:1:1 molar ratios for a total of 160 ug of sample in 250 ul. Samples were subjected to size exclusion chromatography as described in the previous section.

## 3.3 Results

### 3.3.1 His29 functions as a general base to activate the 2′-hydroxyl nucleophile

The biochemical analysis of Csy4 described in Chapter 2 implicated a serine residue (Ser148) as the general base or as important for substrate positioning and a histidine residue (His29) as a general acid in the transesterification reaction catalyzed by Csy4. In our previous experiments, we conducted single time-point (5 min) reactions. This method may obscure mutants that have severe cleavage defects but nevertheless retain a low level of activity, and so to more accurately investigate the specific involvement of the proposed catalytic dyad and other active site-proximal residues during pre-crRNA cleavage (Fig. 3.2A), we performed quantitative single-turnover cleavage assays with various mutants and determined their corresponding first-order rate constants (Fig. 3.2B and C; Table 3.1). Alanine substitution of the active site histidine abolished all activity, indicating that His29 contributes an essential catalytic function.

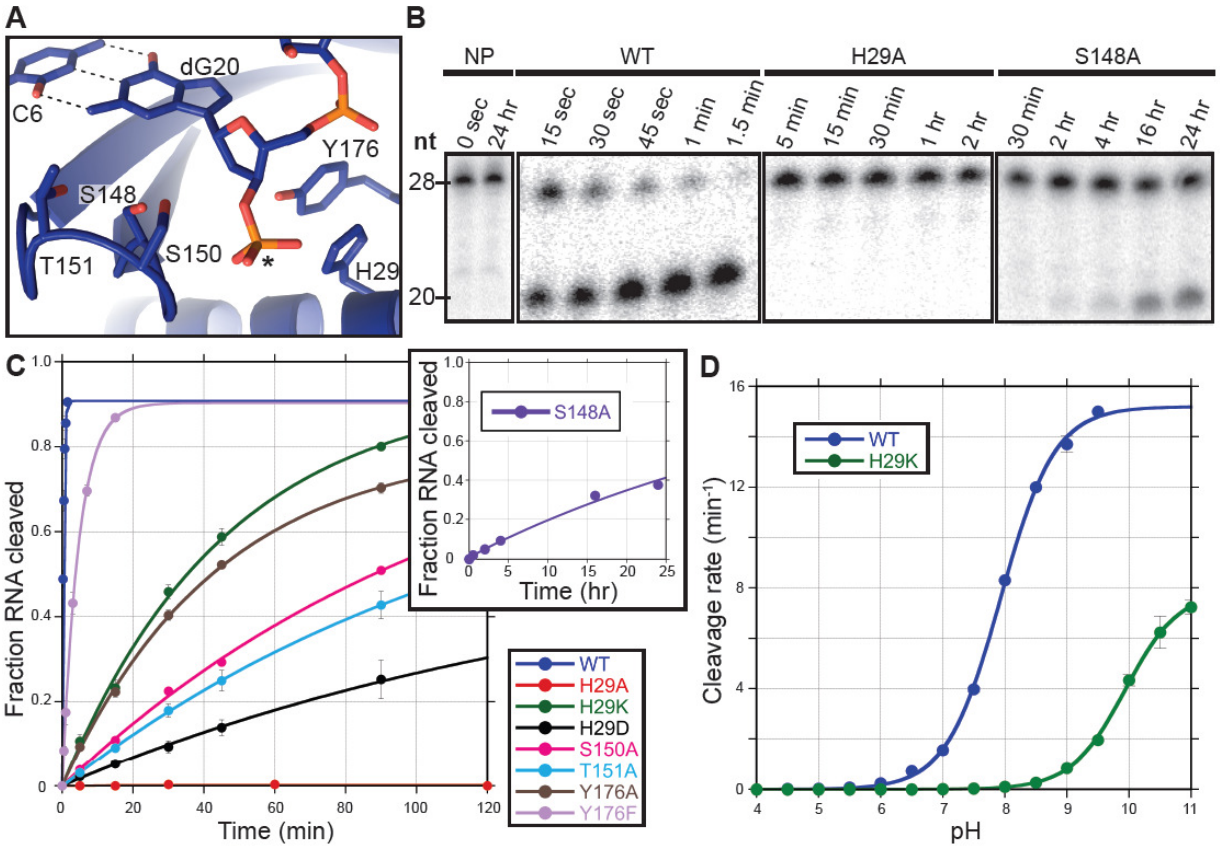



**Figure 3.2 Amino acid contributions to catalysis.** (**A**) Csy4 active site from Csy4/substrate complex (PDB ID 2XLK). Active site residues are shown in stick format and the scissile phosphate is marked with an asterisk. The hydrogen bonds of the base pair between nucleotides C6 and dG20 are shown as dashed lines. (**B**) Representative single-turnover cleavage assays with wild-type and mutant Csy4. No protein (NP) controls shown at left. (**C**) Single-turnover cleavage analysis of wild-type and mutant Csy4. Data plotted are average of triplicate experiments and error bars represent the standard error of the mean (s.e.m.). Solid lines represent fits to an exponential equation. (**D**) pH-rate profile for wild-type and H29K

Csy4. Rapid cleavage kinetics above pH 9.5 for wild-type Csy4 prevented accurate determination of the rate. Each data point is an average of three independent experiments and error bars represent the s.e.m. Data were fit according to the equation described in the Methods.

| | $k_{obs}$ ($min^{-1}$) | Fold defect relative to WT |
|---|---|---|
| WT | 2.100 ± 0.04 | |
| H29A | 0 | — |
| H29K | 0.022 ± 0.0006 | 130 |
| H29D | 0.0032 ± 0.0007 | 910 |
| S148A | 0.00035 ± 0.00003 | 8,300 |
| S150A | 0.0084 ± 0.0007 | 350 |
| T151A | 0.0076 ± 0.0009 | 380 |
| Y176F | 0.22 ± 0.02 | 13 |
| Y176A | 0.023 ± 0.002 | 130 |

**Table 3.1 Observed cleavage rates for WT and mutant Csy4.** These rates are the averages of three independent experiments, and errors represent the standard error of the mean.

To further investigate the role of His29, we evaluated the pH dependence of Csy4-catalyzed RNA cleavage. The resulting pH-rate profile (Fig. 3.2D) exhibits a sigmoidal shape and reveals that cleavage rates increase monotonically with pH. These data are consistent with the catalytic requirement of a single titratable residue having a pKa ≈ 7.9 that is active only in its deprotonated state. Consistent with our previous work, a Csy4 mutant with lysine substitution of His29 retains cleavage activity, albeit with ~130-fold slower kinetics than wild-type (Fig. 3.2C; Table 3.1). The pH-rate profile for RNA cleavage by the H29K mutant has the same shape as wild-type but is shifted to a higher pH (Fig. 3.2D; pK ≈ 9.9), in good agreement with the corresponding shift in pKa of the imidazole and amino side groups of histidine and lysine, respectively. These data strongly suggest that catalytic activity requires His29 to be in its deprotonated form, and that this residue functions as a general base during cleavage by activating the 2′-hydroxyl nucleophile through proton abstraction. Substitution of His29 with aspartate, whose side chain is negatively charged at physiological pH, resulted in a functional enzyme, further supporting the role of His29 as the general base (Fig.e 3.2C).

Direct proton abstraction would require the His29 side chain to be positioned proximally to the G20 2′-hydroxyl, but in the previously described Csy4/substrate structures, the His29 side chain interacts instead with the scissile phosphate and is not within hydrogen bonding distance of the expected location of the 2′-hydroxyl. Those crystals were grown at acidic pH ranges (~4.6–5) where the His29 side chain is likely to be protonated and Csy4 is catalytically defective (Fig. 3.2D). Thus, the previously observed interaction between the scissile phosphate and His29 side chain may result artificially from the acidic pH of the crystallization conditions (see below).

Alanine substitution of Ser148 decreased the cleavage rate ~8000-fold relative to wild-type (Fig. 3.2B and C; Table 3.1), suggesting that this residue plays a critical role in substrate binding, positioning, or cleavage chemistry (see below). Mutation of Tyr176 to phenylalanine or alanine reduced the cleavage rate only ~13-fold and ~130-fold, respectively (Fig. 3.2C; Table 3.1). The side chain of Tyr176 points into the active site and stacks on top of the His29 imidazole group; mutation to phenylalanine likely disrupts any role the phenolic hydroxyl plays in substrate binding, whereas mutation to alanine could also disrupt His29 positioning. Alanine substitution of either Ser150 or Thr151, both located in the active site loop, reduced the cleavage rate ~350-fold, suggesting these residues may play a role in either direct binding of the RNA substrate or by forming a network of hydrogen-bonding interactions that orient the side chain of Ser148.

### 3.3.2 The Csy4 active site constrains the G20 ribose in the C2′-endo sugar pucker

To determine how Csy4 interacts with the 2′-hydroxyl nucleophile, we crystallized a Csy4/RNA product complex comprising Csy4(S22C) and a 19-nucleotide RNA product that was generated by endoribonucleolytic cleavage of a 20-nucleotide substrate RNA (Fig. 3.3A). Csy4(S22C) is a mutant of Csy4 that retains wild-type activity and yields better diffracting crystals as described in Chapter 2. Crystals of this complex diffracted x-rays to 2.0 Å resolution, and the structure was solved by molecular replacement using the previous substrate complex structure (PDB ID 2XLK) as a search model (Table 3.2). The structure of Csy4 in this product complex is similar to that observed in the previously published substrate complex (PDB ID 2XLK; RMSD = 0.431 Å over 811 atoms) (Fig. 3.3B and Fig. 3.4A). Additionally, the crRNA hairpins of the product and substrate RNAs are bound to Csy4 in the same location and align with an RMSD of 0.519 Å over 214 atoms. We observed clear density for a 3′-phosphate (Fig. 3.5), consistent with previous mass spectrometry results that identified the termini of Csy4 cleavage products as a 5′-hydroxyl and 3′-phosphate (Wiedenheft et al., 2011b). Additionally, we observe that nucleotide A5, a single-stranded nucleotide immediately upstream of the stem-loop, makes two hydrogen-bonding contacts in a base-specific fashion with the peptide backbone of Leu139 (see Chapter 4; (Sternberg et al., 2012)).

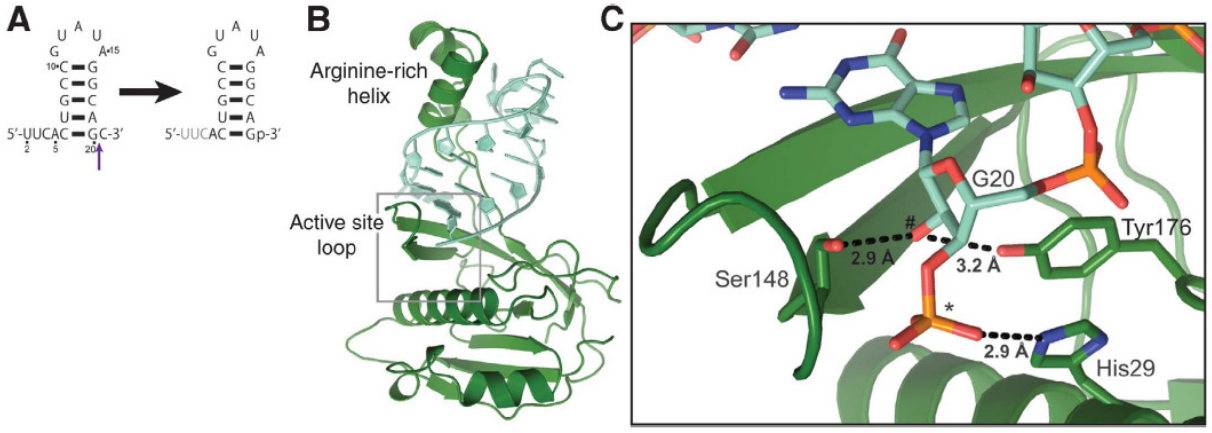



**Figure 3.3 Crystal structure of Csy4/product RNA complex at 2.0 Å resolution.** (**A**) Shown at left is the substrate RNA used to generate the protein/RNA complex. Cleavage by Csy4 (purple arrow)

produces the product RNA (right) present in the crystal structure. Gray lettering denotes nucleotides for which there was no corresponding electron density and therefore could not be modeled. (**B**) Overall structure of Csy4S22C (dark green) bound to product RNA (light green). Electron density was well-defined for all 187 amino acids of Csy4 and 16 of the 19 nucleotides in the product RNA. (**C**) Detailed view of the Csy4 active site (gray box, in (B)). The 2′-hydroxyl nucleophile is marked with a pound sign and the scissile phosphate is marked with an asterisk. RNA/protein hydrogen-bonding interactions are marked with dashes.

| | **Product** | **S148A** | **Minimal** |
|---|---|---|---|
| **Data collection** | | | |
| Space group | *C*2 | *C*2 | *C*2 |
| Cell dimensions | | | |
| *a*, *b*, *c* (Å) | 60.79, 47.80, 86.57 | 62.39, 46.88, 87.39 | 62.84, 47.28, 87.93 |
| α, β, γ (˚) | 90.0, 109.7, 90.0 | 90.0, 107.2, 90.0 | 90.0, 106.7, 90.0 |
| Resolution (Å) | 81.51-2.00 (2.05-2.00) | 41.73-2.73 (2.8-2.73) | 36.48-2.32 (2.38-2.32) |
| $R_{sym}$ (%)* | 9.3 (81.3) | 10.5 (65.3) | 8.4 (48.0) |
| $I/\sigma I$* | 15.58 (2.06) | 13.44 (2.54) | 13.5 (2.49) |
| Completeness (%)* | 99.7 (99.0) | 99.7 (100) | 99.4 (100) |
| Redundancy* | 6.5 (5.2) | 5.1 (5.2) | 4.10 (3.3) |
| | | | |
| **Refinement** | | | |
| Resolution (Å) | 81.51-2.00 | 41.73-2.73 | 36.48-2.32 |
| No. reflections | 15934 | 7275 | 10773 |
| $R_{work}/R_{free}$ | 0.183, 0.238 | 0.200, 0.252 | 0.202, 0.256 |
| No. atoms | | | |
| Protein | 1458 | 1172 | 1182 |
| RNA | 348 | 314 | 306 |
| Water/ligands | 134 | 30 | 45 |
| B-factors | | | |
| Protein | 31.2 | 57.6 | 49.5 |
| RNA | 47.8 | 135.7 | 119.1 |
| Water/ligands | 36.2 | 47.3 | 40.4 |
| R.m.s. deviations | | | |
| Bond lengths (Å) | 0.012 | 0.016 | 0.015 |
| Bond angles (˚) | 1.328 | 1.477 | 1.540 |
| Ramachandran plot (%) | | | |
| Preferred region | 95.74 | 95.42 | 95.33 |
| Allowed region | 4.26 | 4.58 | 4.67 |
| Outliers | 0 | 0 | 0 |

* Values in parentheses denote highest resolution shell

**Table 3.2 Data collection and refinement statistics.**

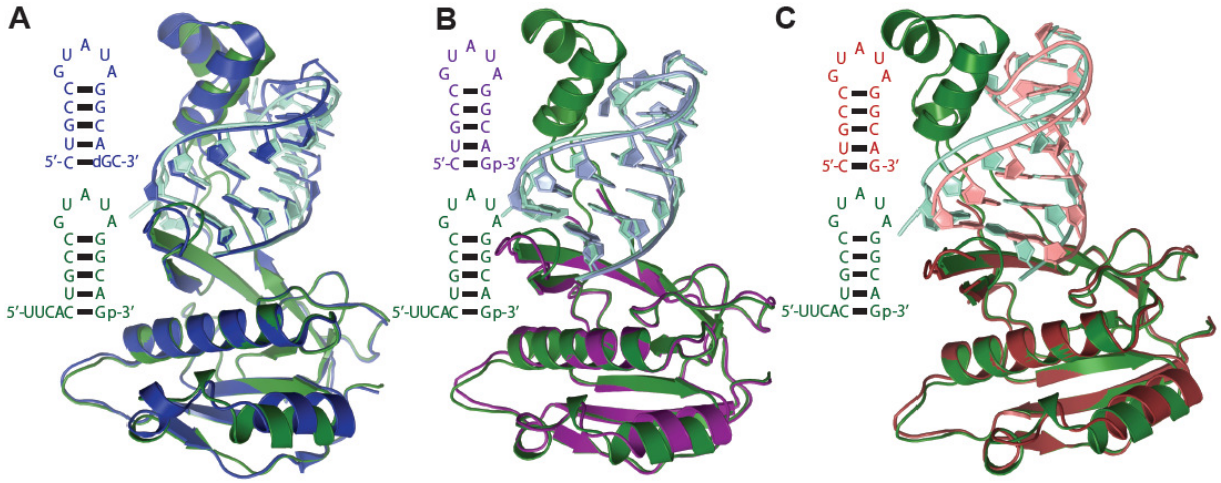



**Figure 3.4 The overall folds of the Csy4/product complexes are highly similar to each other and the previously published Csy4/substrate complex.** (**A**) Chains A and C from the substrate complex (dark blue, PDB ID 2XLK) align with the protein and RNA molecules from the product complex (dark and light green) with an RMSD of 0.431 Å and 0.519 Å over 811 and 214 atoms, respectively. Depicted at left is the RNA content of the crystal structures. (**B**) The protein and RNA molecules from the product and S148A mutant structures (dark and light purple) align with an RMSD of 0.309 Å and 0.526 Å over 815 and 270 atoms, respectively. Depicted at left is the RNA content of the crystal structures. (**C**) The protein and RNA molecules from the product and minimal complex structures (dark red and pink) align with an RMSD of 0.346 Å and 0.499 Å over 843 and 263 atoms, respectively. The RNA stem from the minimal structure exhibits a rigid body rotation with respect to the Csy4 molecule as compared to the other structures shown. Depicted at left is the RNA content of the crystal structures.

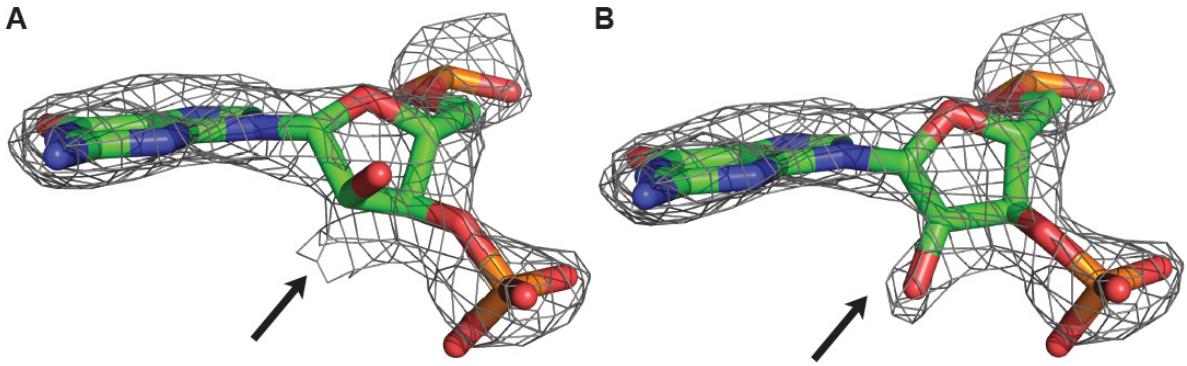



**Figure 3.5 The G20 ribose adopts the C2′-endo conformation in the active site of the product complex.** We performed a molecular replacement experiment on the product complex native dataset using the A and C chains (protein and RNA, respectively) from the previously determined Csy4-substrate complex (PDB ID 2XLK) as search models. We removed the G20 nucleotide and C21 phosphate from the RNA search model in order to minimize model bias. A $2F_o$-$F_c$ electron density map calculated from the molecular replacement solution phases contoured at 1 σ is displayed in gray mesh. Density for the G20 nucleotide and C21 phosphate is readily apparent. We manually built the G20 nucleotide into the electron density using a model with either a C3′-endo (**A**) or C2′-endo sugar pucker (**B**). The nucleotide modeled with a C3′-endo sugar pucker does not accurately account for all of the observed density (black arrow), whereas the nucleotide modeled with a C2′-endo sugar pucker agrees well with the observed electron density (black arrow).

Unique to the product complex structure is the presence of the 2′-hydroxyl nucleophile in the active site (Fig. 3.3C), which was readily apparent in the molecular replacement solution (Fig. 3.5). Upon modeling a ribonucleotide into the active site, we observed that the electron density was inconsistent with a ribose in the C3′-endo conformation but was fit well with a ribose in the C2′-endo form (Fig. 3.5). The 2′-hydroxyl nucleophile is positioned between the side chains of Ser148 and Tyr176, both of which are within hydrogen-bonding distance (2.10 Å and 3.2 Å) (Figure 3.3C), suggesting that these interactions may force the G20 ribose to adopt the C2′-endo sugar pucker observed in the crystal structure. In-line attack of a 2′-hydroxyl nucleophile on the adjacent scissile phosphate requires a locally extended RNA backbone (Yang, 2011) and does not proceed when the sugar pucker is C3′-endo. The observation of a C2′-endo sugar pucker in the Csy4 active site is therefore representative of the extended conformation that would be required for cleavage to proceed.

### 3.3.3 Ser148 positions the RNA for cleavage

Our cleavage assays demonstrated that the S148A mutation is far more deleterious to catalysis than the Y176A mutation, suggesting that Ser148 is the primary residue responsible for positioning the 2′-hydroxyl and maintaining the requisite extended phosphate backbone conformation. The Tyr176 side chain likely plays a redundant role in stabilization of the C2′-endo conformation and may be more important for positioning His29. To test this hypothesis, we crystallized a complex of Csy4(S148A) and a 16-nucleotide substrate RNA (Fig. 3.6A). The resulting 2.6 Å structure (Fig. 3.6B, Table 3.2), solved by molecular replacement, likely contained a mixture of substrate and product RNAs (16- and 15-nucleotides in length, respectively) due to the slow rate of Csy4(S148A)-catalyzed cleavage. The C21 nucleotide, immediately downstream of the scissile phosphate, is disordered when present and electron density for this nucleotide is therefore not observed (Haurwitz et al, 2010). The Csy4(S148A) protein structure is similar to that of wild-type Csy4 (RMSD = 0.309 Å over 815 atoms), and the RNA hairpin is bound to the S148A mutant in the same location as observed in the product structure (RMSD = 0.526 Å over 270 atoms; Fig. 3.4B). However, the active site ribose adopts a C3′-endo sugar pucker in this case, thereby repositioning the 2′-hydroxyl nucleophile 5.5 Å away from the Tyr176 side chain (Fig. 3.6C). We conclude that the Tyr176 side chain is insufficient to maintain the C2′-endo sugar pucker in the absence of Ser148, suggesting that the large catalytic defect for the S148A mutant may result from the Csy4 active site relying on the inherent sugar pucker interconversion rate in order for the substrate phosphate backbone to be properly extended for cleavage.

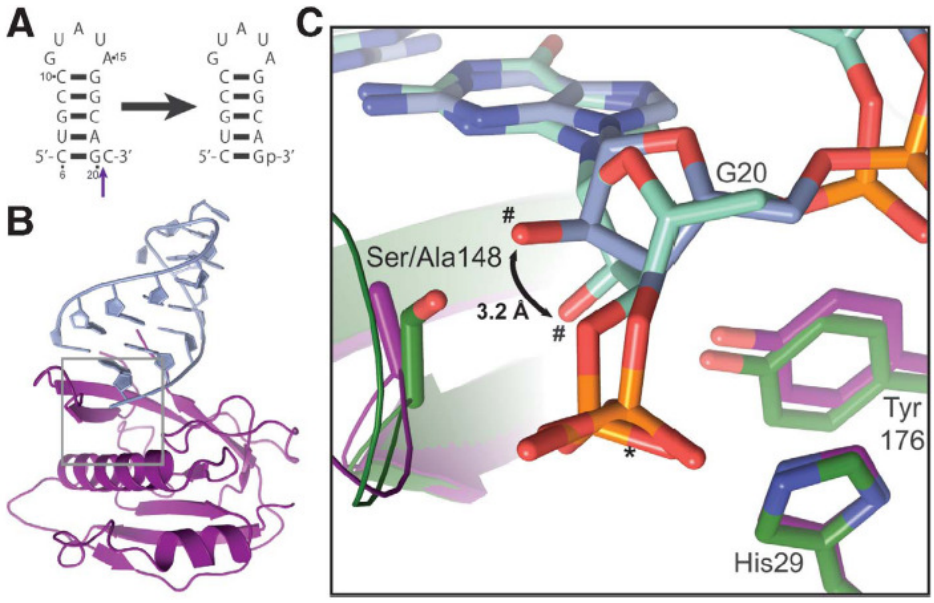



**Figure 3.6 Crystal structure of the Csy4S148A/RNA complex at 2.6 Å resolution.** (**A**) Shown at left is the substrate RNA used to generate the protein/RNA complex. Cleavage by Csy4 (purple arrow) produces product RNA (right). Because of the slow cleavage rate of the S148A mutant, crystals likely contained a mixed population of substrate and product RNAs. (**B**) Overall structure of Csy4S148A (dark purple) and RNA (light purple). 153/187 amino acids and 14/ 15 nucleotides could be modeled into the electron density. The amino acids composing the arginine-rich helix are among those for which there is little to no electron density. (**C**) Superposition and close-up of product complex (green) and S148A complex (purple) active sites (gray box, in (B)). The double-headed black arrow highlights the 3.2 Å change in 2′-hydroxyl location between the two structures. The two 2′-hydroxyl nucleophiles are labeled with pound signs and the scissile phosphates are indicated with an asterisk.

### 3.3.4 His29 may interact directly with the 2′-hydroxyl nucleophile

As described above, all of the Csy4/RNA crystal structures result from crystals grown at pH 4.6–5. To determine what interactions His29 may make in the absence of the potentially pH-induced interaction with the scissile phosphate, we crystallized a complex of Csy4 and a 15-nucleotide RNA composed of only the crRNA hairpin with a 3′-hydroxyl terminus (Fig. 3.7A). The 2.3 Å resolution structure of this complex (hereafter called the minimal structure) once again revealed a Csy4 conformation similar to that observed previously (RMSD = 0.346 Å over 843 atoms; RNA superposition RMSD = 0.499 Å over 263 atoms) (Fig. 3.7B and Fig. 3.4C; Table 3.2). While the locations of the Tyr176 and His29 side chains are nearly identical between the product and minimal structures, the G20 nucleotide and the active site loop that contains Ser148 shift 3.4 Å and 2.5 Å between the two structures, respectively (Fig. 3.7C). The G20 ribose is in the C2′-endo conformation, and the 2′-hydroxyl nucleophile is 3.6 Å and 3.7 Å away from the His29 and Tyr176 side chains, respectively (Fig. 3.7D). The lack of a 3′-phosphate results in significant disorder in the active site loop as is evidenced by a lack of density for residue 149 and for the side chains of nearly all of the active site loop residues (Fig. 3.7D). This structure provides evidence that there is flexibility in the location of RNA within the Csy4 active site because in previous structures, the His29 sidechain is greater than 5 Å from the G20 2′-hydroxyl. This flexibility likely facilitates His29 activating the 2′-hydroxyl nucleophile via proton abstraction.

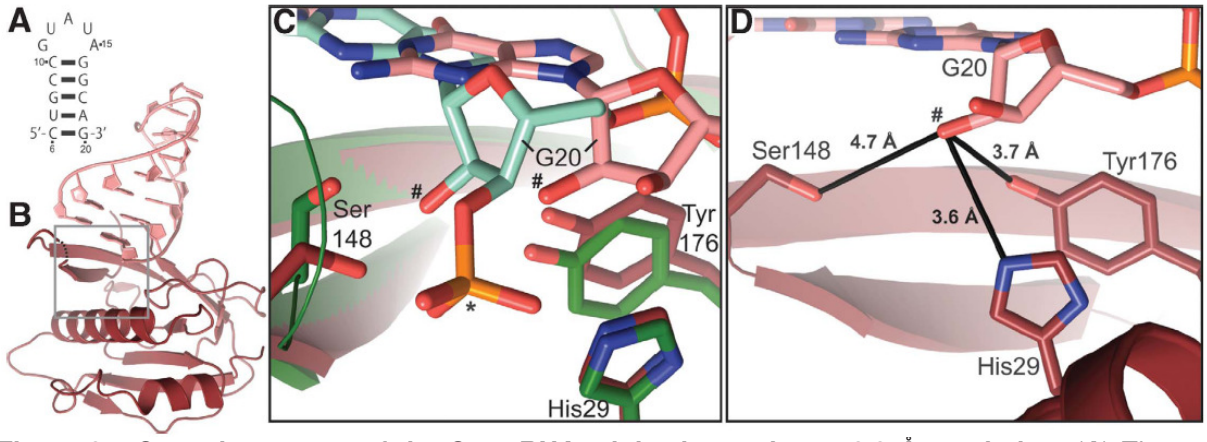



**Figure 3.7 Crystal structure of the Csy4/RNA minimal complex at 2.3 Å resolution.** (**A**) The stem-loop RNA used for co-crystallography lacks a 3′-phosphate. (**B**) Overall structure of Csy4 (dark red) and stem-loop RNA (pink). 151/187 amino acids and all 15 RNA nucleotides could be modeled into the electron density. Electron density for the active site loop is severely broken, and a dashed line indicates its approximation location. There is no electron density for the arginine-rich helix. (**C**) Superposition and detailed view of product complex (green) and minimal complex (red) active sites (gray box, in (B)). The scissile phosphate belonging to the product complex is marked with an asterisk and the two 2′-hydroxyl nucleophiles are marked with pound signs. (**D**) Magnified view of the minimal complex active site. Black lines indicate the distances between active site residues and the 2′-hydroxyl nucleophile.

### 3.3.5 Csy complex formation requires Csy4-catalyzed cleavage of CRISPR transcripts

Recent work has demonstrated that Csy4 associates with three other Cas proteins (Csy1-3) and a single copy of crRNA to form the Csy complex, which targets complementary nucleic acids (Wiedenheft et al., 2011b). To determine whether pre-crRNA cleavage by Csy4 is necessary for complex formation, we co-expressed Csy1-3 and a pre-crRNA with either wild-type Csy4 or the catalytically inactive mutant, Csy4(H29A), in *E. coli* BL21(DE3) cells. The Csy complex was affinity purified via a 6 X histidine tag appended to the N-terminus of Csy3, followed by size exclusion chromatography. Co-expression of the wild-type proteins and pre-crRNA yielded an RNP with an estimated molecular mass of ~350 kilodaltons (Fig. 3.8), in agreement with previous work (Wiedenheft et al., 2011b). Substitution of catalytically inactive Csy4 in the co-expression experiment resulted in the purification of only Csy3, which was not associated with a crRNA (Fig. 3.8). Csy3 over-expressed on its own in *E. coli* BL21(DE3) cells purifies as both a large oligomeric complex containing non-specific RNA and as a nucleic acid-free monomer (unpublished observations), similar to the two peaks observed for Csy3 co-expressed with mutant Csy4. To ensure that Csy4(H29A) is defective only in catalysis and not in its ability to interact with other Csy complex components, we mixed together Csy complex components that were individually purified and evaluated the mixtures by size exclusion chromatography. Adding either wild-type or H29A Csy4 to Csy1-3 and a mature crRNA resulted in Csy complex formation (Fig. 3.9), suggesting that the Csy4(H29A) mutant is defective only for

catalysis and not for interaction with other Csy complex components, and that catalysis is a necessary precursor to complex formation.

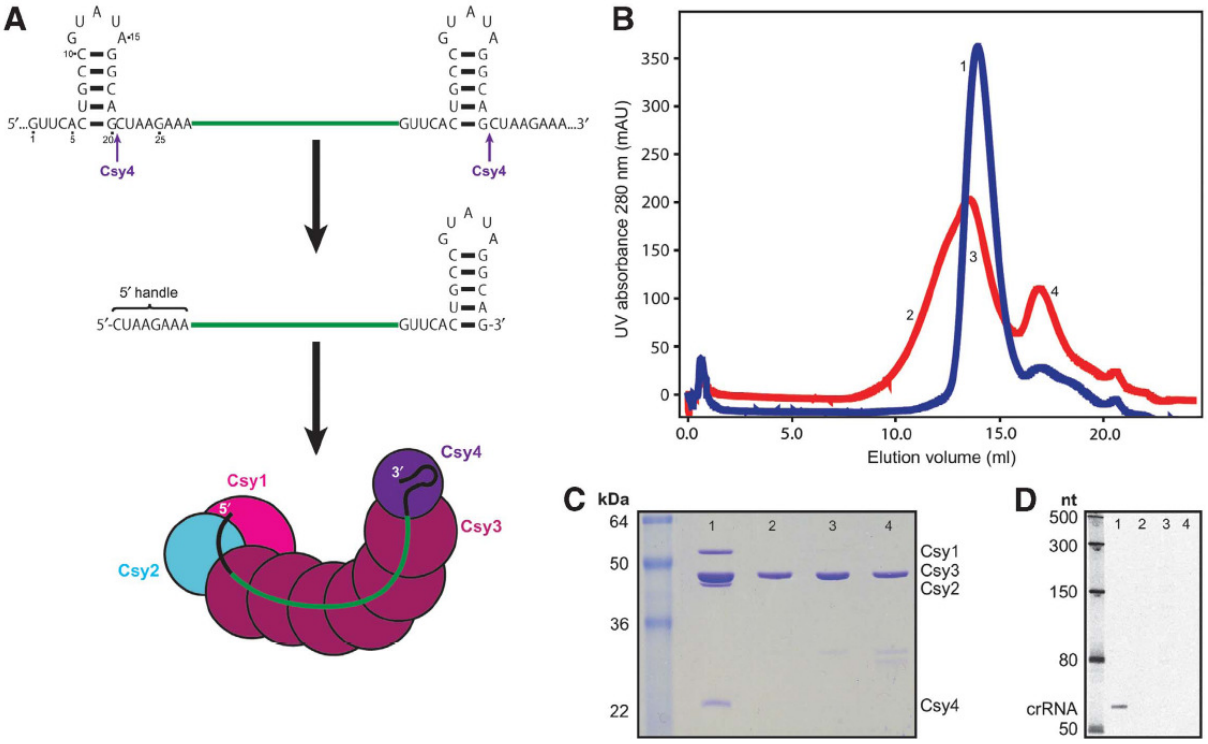



**Figure 3.8 Csy4 cleavage of pre-crRNA is required for Csy complex formation.** (**A**) Schematic depicting pre-crRNA cleavage by Csy4 and formation of the Csy CRISPR ribonucleoprotein (crRNP) complex. The CRISPR repeat and spacer sequence are in black and green, respectively. Cleavage sites are denoted with purple arrows. (**B**) Superose 6 gel filtration column elution profiles of affinity-purified Csy1, Csy2, $His_6$-Csy3, and pre-crRNA co-expressed with wild-type (blue) or H29A (red) Csy4. (**C**) Coomassie blue-stained 12% SDS–PAGE showing protein components of the superose 6 fractions for wild-type (lane 1) and H29A (lanes 2–4, as noted in (B)) Csy4 co-expression assays. (**D**) SYBR Gold-stained 15% denaturing PAGE showing phenol:chloroform extracted nucleic acids from superose 6 fractions (from (B)).

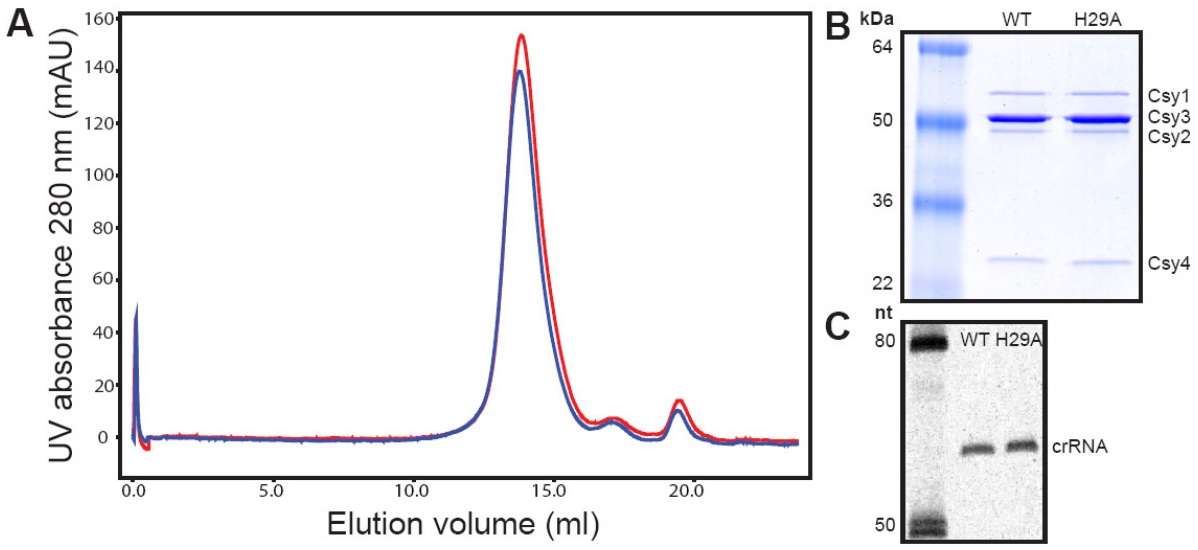



**Figure 3.9 Csy4(H29A) is competent for assembly into the Csy complex.** (**A**) Superose 6 gel filtration column elution profiles of recombinantly assembled Csy complex containing individually purified Csy1/Csy2 heterodimer, Csy3 monomer, mature crRNA, and wild-type (blue) or H29A (red) Csy4. (**B**) Coomassie blue-stained 12% SDS PAGE showing protein components of the superose 6 fractions for wild-type (WT) and H29A mutant (H29A) Csy4 *in vitro* assembly assays (as noted in (A)). (**C**) SYBR Gold stained 15% denaturing PAGE showing phenol:chloroform extracted nucleic acids from superose 6 fractions (from (A)).

Taken together with previous work demonstrating that Csy complex assembly does not proceed in the absence of RNA (Wiedenheft et al., 2011b), we conclude that Csy4-catalyzed biogenesis of mature crRNAs with fully processed termini is necessary for stable Csy complex formation.

## 3.4 Discussion

The production of crRNAs is central to CRISPR-mediated adaptive immunity in prokaryotes. The three crystal structures of Csy4/RNA complexes and quantitative cleavage assays presented here reveal an unexpected endoribonuclease active site in which a serine residue constrains the nucleophile-containing ribose in the C2′-endo sugar pucker and a histidine residue serves as the general base to activate the 2′-hydroxyl nucleophile. Unlike RNase A and other well-studied metal ion-independent nucleases, the Csy4 active site lacks a general acid and positively charged residues near the active site that would lower the energetic barrier to the transition state, resulting in correspondingly slow cleavage rates. We propose that upon binding a pre-crRNA substrate, the Ser148 residue rearranges the G20 ribose into the C2′-endo conformation, providing the correct geometry for His29 to abstract a proton from the 2′-hydroxyl nucleophile and enable nucleophilic attack of the scissile phosphate. The resulting 2′,3′-cyclic phosphate terminus is likely opened to a 3′-phosphate via hydrolysis by a water. Csy4 then retains its crRNA product (Sternberg et al., 2012) and serves as the nucleation point for Csy complex formation.

We observe that the G20 ribose in the wild-type Csy4 active site adopts the C2′-endo sugar pucker. The C2′-endo conformation is generally rare in double-stranded RNA but is overrepresented in catalytic active sites and RNA tertiary interactions (Cantor and Schimmel, 1980; Mortimer and Weeks, 2009). In the Csy4 active site, Ser148 and Tyr176 likely interact directly with the 2′-hydroxyl nucleophile via hydrogen bonding, restraining the ribose ring in the C2′-endo conformation. Mutation of Ser148 to alanine slows cleavage nearly 8000-fold and allows the G20 ribose to retain the C3′-endo conformation. We propose that this significant cleavage rate defect may arise from a particularly slow rate of C2′-endo/C3′-endo interconversion at the G20 ribose in the absence of the Ser148 side chain. While most RNA sugars interconvert between the C2′- and C3′-endo conformations on a microsecond to millisecond time scale (Johnson and Hoogstraten, 2008), a discrete set of C2′-endo nucleotides has been observed to experience local dynamics with half-lives on the order of 10–100 seconds, significantly slower than other local RNA conformational changes (Gherghe et al., 2008; Mortimer and Weeks, 2009). For example, the folding rate of bacterial RNase P RNA is limited by the sugar pucker interconversion of a single RNA nucleotide from C3′-endo to C2′-endo, which occurs at a rate of ~0.24 $min^{-1}$ (Mortimer and Weeks, 2009). Consistent with the observation that members of this class of slow interconverting C2′-endo containing ribonucleotides are partially constrained by hydrogen-bonding or base-stacking interactions (Gherghe et al., 2008), the G20 nucleotide base pairs with C6, hydrogen-bonds with Arg102 on the major groove face, and stacks below A19 and above Phe155. We hypothesize that G20 belongs to this unusual class of C2′-endo containing nucleotides and propose that the ~8000-fold defect in observed cleavage rate of the S148A mutant is due in large part to the extremely slow sugar pucker interconversion dynamics of the G20 nucleotide. However, we cannot rule out that the hydrogen bonding interaction between Ser148 and the 2′-hydroxyl also contributes to nucleophile activation.

The observed rate of cleavage for wild-type Csy4 (~3 $min^{-1}$ at pH 7.2) is orders of magnitude slower than that of other well-characterized RNases. For example, RNase A enzymes from a variety of organisms cleave RNA substrates with apparent single-turnover rate constants of 910 to 40 500 $min^{-1}$ (Katoh et al., 1986), and the colicin E5 ribonuclease from *E. coli* cleaves minimal substrates with a $k_{cat}$ of ~5000 $min^{-1}$ (Ogawa et al., 2006). In fact, Csy4 has an observed cleavage rate similar to ribozyme-catalyzed RNA cleavage rate constants, which are typically < 2 $min^{-1}$ (Zamel et al., 2004). Ribozymes perform the same transesterification reaction as protein RNases (Cochrane and Strobel, 2008), but are thought to be significantly slower because they typically lack general acids and bases with p$K_a$ values close to neutral pH (Yang, 2011). The well characterized metal-independent RNase families of RNase A, RNase T1, and RNase T2 contain catalytic cores composed of a histidine pair; a glutamate and histidine; and a glutamate, lysine, and three histidines, respectively (Yang, 2011). Like many of these protein RNases, the Csy4 active site contains a histidine general base, but it appears to lack a general acid as there is no chemically appropriate residue positioned proximal to the 5′-hydroxyl leaving group. Consistent with this observation is the sigmoidal shape of the Csy4 pH-rate profile (Fig. 3.2D). Whereas RNase A exhibits a bell-shaped pH-rate profile indicative of a cleavage mechanism that relies on two titratable residues (Raines,

1998), the Csy4 pH-rate profile is consistent with only a single titratable residue that is likely to be His29.

An additional hallmark of metal ion-independent RNases is stabilization of the pentacovalent transition state by one or more positively charged residues (Cochrane and Strobel, 2008). Like ribozymes, which lack functional groups that are positively charged at a neutral pH, Csy4 does not have any positively charged residues in or surrounding the active site. We hypothesize that Csy4 compensates for a lack of stabilizing positive charges by making additional hydrogen bonds to the transition state, analogous to the hairpin ribozyme, which makes 2–3 more contacts to the transition state than precursor or product RNAs (Cochrane and Strobel, 2008; Rupert and Ferre-D'Amare, 2001; Rupert et al., 2002). This is consistent with the ~350-fold effect on cleavage observed for alanine substitution of Ser150 or Thr151, which lie in the active site loop and participate in a hydrogen bonding network that can include Ser148 and the scissile phosphate (Fig. 3.10). Through this network, Ser150 and Thr151 may aid in the stabilization of the pentacovalent transition state.

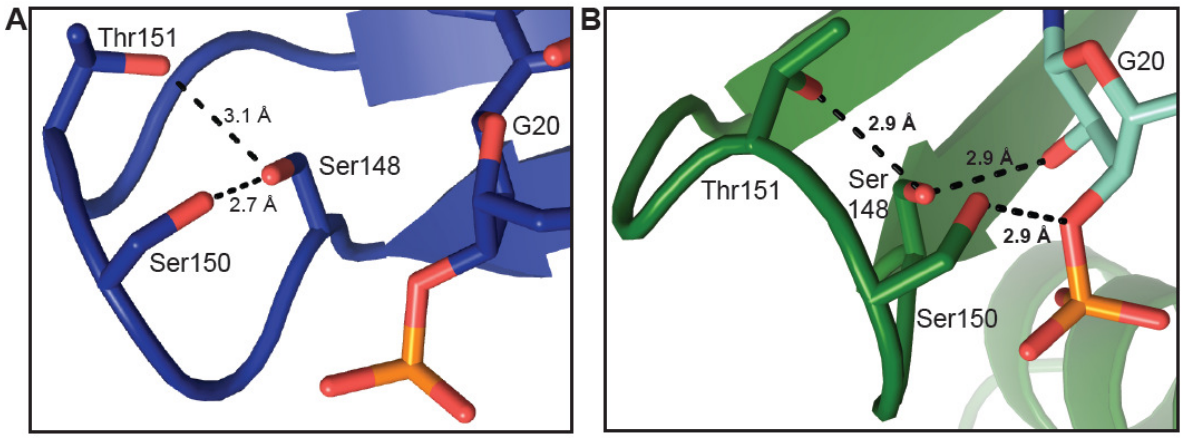



**Figure 3.10 Active site loop residues have the potential to form a hydrogen bonding network with one another and the bound RNA.** Detailed view of the active site loops from the (**A**) substrate (PDB ID 2XLK) and (**B**) product complexes. Dashed lines indicate hydrogen bonding interactions.

Using an *in vivo* assembly assay, we found that crRNA processing by the endoribonuclease Csy4 is essential to the stable formation of crRNA-containing targeting complexes that bind to complementary nucleic acids and trigger their degradation. Because Csy complexes do not stably form on unprocessed pre-crRNA, we hypothesize that the formation of the mature Csy crRNP requires a free 5′ terminus generated by Csy4-catalyzed cleavage. Mature crRNAs across multiple CRISPR types contain 8-nucleotides of repeat-derived sequence at the 5′ end (Brouns et al., 2008; Carte et al., 2008; Hale et al., 2009; Marraffini and Sontheimer, 2008), and it has been proposed that these sequences, termed the 5′ handle, may serve as Cas protein binding sites (Terns and Terns, 2011; Wiedenheft et al., 2012). For example, the 5′ handle forms a hook-like structure in the crRNP from *E. coli* K12 (Cascade) that correlates with termination of the ribonucleoprotein filament (Wiedenheft et al., 2011a). We speculate

that the 5′ handle of the mature crRNA in *P. aeruginosa* recruits one or more Csy proteins to the nascent RNP. The requirement for a free crRNA 5′ terminus during complex formation would therefore point to specific recognition of the 5′ handle in the assembly of Cas protein complexes.

These observations, along with our recent work demonstrating a very tight crRNA binding affinity by Csy4 (50 pM) (Chapter 4; (Sternberg et al., 2012)), have led us to conclude that Csy4 evolved as a finely tuned RNA binding protein while retaining only modest cleavage kinetics. Similarly, the CRISPR type I-E endoribonuclease (referred to as Cas6e, Cse3, or CasE) exhibits relatively slow cleavage kinetics (~5 $min^{-1}$) and tight substrate and product binding ($K_d$≈3nM) (Sashital et al., 2011). Both Csy4 and Cse3 retain their crRNA products and are members of the crRNPs that target invading nucleic acids. These two CRISPR systems have likely evolved CRISPR endoribonucleases whose highly accurate substrate selection ensures incorporation of the appropriate RNA into the targeting complex, while the lack of a substrate turnover requirement has not contributed selective pressure for rapid cleavage kinetics.

# Chapter 4

# Mechanism of Csy4 substrate selection

*A portion of the work presented in this chapter has been previously published as part of the following paper: Sternberg, S.H., Haurwitz, R.E., Doudna, J.A. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* 18, 661-672.

*Samuel Sternberg performed most of the Csy4 binding and cleavage assays. Rachel Haurwitz performed the cleavage assay with the nicked RNA substrate, solved the crystal structure, and performed the Northern blot analysis.

## 4.1 Introduction

Bioinformatic analyses of Csy4-related Cas proteins together with existing CRISPR databases (Grissa et al., 2007) have revealed a potentially large number of enzyme variants whose substrate specificities have co-evolved with the RNAs encoded by CRISPR repeats. Gaining a thorough understanding of the selection mechanism by which Csy4 faithfully binds and cleaves its substrate should inform future work aimed at expanding the toolbox of these sequence-specific endoribonucleases. Furthermore, the propensity of many pre-crRNA repeat sequences to form small, stable stem–loops (Kunin et al., 2007) suggests that general principles of substrate recognition employed by Csy4 will be broadly applicable to other Cas6 family members that associate with structured repeats. To determine the importance of sequence- and shape-specific RNA recognition during pre-crRNA processing, we investigated the relative contributions of substrate base-pair composition and geometry to binding and cleavage by Csy4.

Here we show that Csy4 binds its substrate RNA with extremely high affinity ($K_d$ ≈ 50 pM) and functions as a single-turnover enzyme. Single-stranded RNA (ssRNA) nucleotides that flank the stem–loop contribute negligibly to binding energy, but base-pair changes throughout the double-stranded stem and mutations to the loop sequence result in substantially weaker binding. We find that substrate recognition also involves the precise length of the stem, such that small base-pair insertions cause severe binding and/or cleavage defects due to their effects on helical geometry and substrate positioning. These findings reveal how Csy4 employs a unique set of molecular interactions to achieve highly specific selection of its pre-crRNA substrate while discriminating against similar, non-cognate stem–loop structures.

## 4.2 Methods

### 4.2.1 Protein expression and purification

R102A, Q104A, F155A, and H29A Csy4 mutants were purified as described in Chapter 2. R114A/R118A, R118A/R115A, R115A/R119A, and H120A Csy4 mutants were generated using site-directed mutagenesis and purified essentially as described in Chapter 2, with the following exceptions. Protein genes encoded by the pHGWA vector (Busso et al., 2005) were overexpressed in BL21(DE3) cells. Following the second Ni-NTA affinity purification step, Csy4 mutants were purified by size exclusion chromatography using a single Superdex 75 (16/60) column (GE Healthcare) in 100 mM HEPES ($pH_{RT}$ 7.5), 500 mM KCl, 5% glycerol, 1 mM TCEP. Proteins were then concentrated and buffer-exchanged into 100 mM HEPES ($pH_{RT}$ 7.5), 150 mM KCl, 5% glycerol, 1 mM TCEP; snap-frozen in liquid nitrogen; and stored at -80 °C.

### 4.2.2 Northern blot analysis

Total RNA was extracted from cultures of *P. aeruginosa* PAO1, *P. aeruginosa* UCBPP-PA14, and a csy4 deletion strain of *P. aeruginosa* UCBPP-PA14 (SMC3894) (Zegans et al., 2009) grown to exponential phase using the mirVana kit (Ambion). Duplicate samples of each RNA preparation (6 mg) were separated on adjacent lanes of a 15% denaturing polyacrylamide gel and subsequently transferred to a nylon membrane (Hybond-N+, GE Healthcare) using a semi-dry transfer cell (Bio-Rad). The

single membrane was then cut in half to yield two membranes with identical samples. The membranes were pretreated with ULTRAHyb-Oligo Hybridization Buffer (Ambion) and probed with 5′-[32P]-radiolabeled DNA oligonucleotides corresponding to either the crRNA repeat sequence (5′-GTTCACTGCCGTATAGGCAGCTAAGAAA-3′) or the reverse complement of the crRNA repeat (5′-TTTCTTAGCTGCCTATA CGGCAGTGAAC-3′). Membranes were washed twice with 2X saline-sodium citrate (SSC) buffer containing 0.5% SDS and visualized by phosphorimaging.

### 4.2.3 RNA transcription, purification, and 5′ radiolabeling

The following RNAs were synthesized by Integrated DNA Technologies: the non-cleavable substrate, product RNA (Δ21–28), 5′ truncation constructs (Δ1–5, Δ1–4), the 5′-strand (nucleotides 1–12) and 3′-strand (nucleotides 13–28) used to generate the nicked substrate, the G20A mismatched substrate, and three substrates containing base-pair substitutions at the bottom of the stem (C6U/G20A, C6G/G20C, U7A/A19U). All other RNAs were transcribed *in vitro* using T7 polymerase and purified using denaturating polyacrylamide gel electrophoresis, according to the following protocol. Synthetic single-stranded DNA templates (Integrated DNA Technologies) containing the reverse complement of the desired crRNA repeat construct were annealed to a 1.5-fold molar excess of an oligonucleotide corresponding to the T7 promoter sequence (5′-TAATACGACTCACTATA-3′). Templates encoded an extra guanosine at the 5′ end of all constructs in order to ensure optimal transcription by T7 polymerase. This had no effect on binding affinities but did lead to a slight (~20%) increase in $k_{obs}$ for cleavage of the WT-crRNA repeat substrate. Transcription reactions (100 µl) were incubated at 37 °C for 3–5 h and contained 1 mM template DNA, 100 mg/mL T7 polymerase, 1 mg/ml pyrophosphatase (Roche), 5 mM NTPs, 30 mM Tris-Cl ($pH_{RT}$ 8.1), 25 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 2 mM spermidine, and 0.01% Triton X-100. Reactions were then treated with 5 units of DNase (Promega) and incubated for an additional 30 min at 37 °C before being loaded on a 15% urea-polyacrylamide gel. RNAs were excised from the gel and eluted into DEPC $H_2O$ overnight at 4° C. 5′ triphosphates were removed by incubating RNAs at 37 °C for 1 h with 10 units of calf intestinal phosphate (New England Biolabs) in 1X NEBuffer 3, followed by phenol-chloroform extraction and ethanol precipitation. RNAs were resuspended in DEPC $H_2O$ and stored at -20°C. For biochemical experiments, 10 pmol RNA were 5′ radiolabeled by incubating with 5 units T4 polynucleotide kinase (New England Biolabs) and ~3–6 pmol (~20–40 µCi) [γ-32P]-ATP (Promega) in 1X T4 polynucleotide kinase reaction buffer at 37 °C for 30 min, in a 25 µl reaction. After heat inactivation (65°C for 20 min), reactions were spun through an illustra MicroSpin G-25 column (GE Healthcare) to remove ATP. Radiolabeled RNAs were diluted to ~100 nM stock concentrations with DEPC $H_2O$ and stored at -20°C.

### 4.2.4 Electrophoretic mobility shift assays

Protein concentrations were determined by taking multiple absorbance spectra using a NanoDrop spectrophotometer (Thermo Scientific), averaging absorbance values at 280 nm and converting to molar concentrations using the calculated Csy4 extinction coefficient (15,470 $M^{-1}$ $cm^{-1}$). Spectra were also recorded under denaturing conditions (6 M guanidine hydrochloride, 20 mM potassium phosphate buffer, pH 6.5), and absorbance values were within error of those taken under native conditions. Binding

experiments were conducted in the following buffer: 20 mM HEPES ($pH_{RT}$ 7.5), 100 mM KCl, 5% glycerol, 0.01% Igepal-630, 1 mM DTT, and 0.1 mg/mL yeast tRNA (Sigma-Aldrich) to prevent nonspecific binding. After diluting concentrated 5′-[32P]-labeled RNA and Csy4 stock solutions into 1X binding buffer, trace amounts of RNA (≤0.05–0.2 nM, depending on construct and specific activity) were incubated with increasing concentrations of Csy4 in a 15 µl reaction at room temperature (~24 ℃) for one hour. Twelve microliters of each reaction were then loaded on a 10% native polyacrylamide gel containing 0.5X TBE buffer and resolved by running at 12 W for 90–120 min at 4℃ in 0.5X TBE running buffer. Phosphor screens were exposed to dried gels and scanned with a Storm imager (GE Healthcare), and the intensities of unbound and Csy4-bound RNA were quantified using Image- Quant (GE Healthcare). The fraction of RNA bound at each Csy4 concentration was plotted as a function of Csy4 concentration, and binding data were fit with a standard binding isotherm using Kaleidagraph (Synergy Software), according to the equation:

$$\text{fraction bound} = A \times [\text{Csy4}] \div (K_d + [\text{Csy4}]),$$

where A is the amplitude of the binding curve.

Binding experiments with the substrate nicked between U12 and A13 contained ~1 nM radiolabeled 3′-strand (nucleotides 13–28) and a 1000-fold excess (1 mM) of cold 5′-strand (nucleotides 1–12). For experiments with $K_d$ values in the low pM range, binding data were also fit with the solution of a quadratic equation describing a bimolecular dissociation reaction, as described previously (Maag and Lorsch, 2003), out of concern that [RNA] in these experiments was not sufficiently below the $K_d$ to approximate $[\text{Csy4}]_{total} = [\text{Csy4}]_{free}$. This analysis returned values that agreed well with equilibrium dissociation constants determined from the standard binding isotherm equation, so these original values are reported. When fitting binding data with the rc-crRNA repeat, the amplitude was set equal to one because saturation could not be reached. Binding data with the RNA substrate containing a five G–C basepair insertion showed apparent cooperativity and were fit with a modified binding equation using a variable Hill coefficient (n ≈ 1.5) and an amplitude fixed at one.

At least one binding experiment for each RNA or Csy4 mutant titrated Csy4 across a concentration range of three orders of magnitude centered around the $K_d$. Additional replicates typically tested five concentration points centered around the $K_d$ and returned values in excellent agreement with those derived from a more complete titration. $K_d$ values presented represent the average and standard error of the mean from at least three independent experiments. The average percent error for all reported $K_d$ values is 10%. ΔΔG values for Csy4 or RNA mutants were calculated according to the equation:

$$\Delta\Delta G = -RT\ln(K_{d,WT}/K_{d,mutant}),$$

where R is the gas constant, T is temperature (set to 298 K), and $K_{d,WT}/K_{d,mutant}$ is the ratio of $K_d$ values for the WT and mutant construct.

### 4.2.5 RNA cleavage assays

Cleavage assays were conducted at room temperature (~24°C) in the following buffer: 20 mM HEPES, 100 mM KCl, 1 mM DTT at $pH_{RT}$ 7.5. Single-turnover cleavage experiments were 55 µl in volume and contained 0.5 nM 5'-[32P]-labeled RNA and a saturating concentration of Csy4 (typically 500 nM). At each desired time point, a 10 µl aliquot was removed and quenched by mixing it with 50 µl phenol:chloroform:isoamyl alcohol 25:24:1 at pH 8.0 (Sigma-Aldrich). The aqueous layer was mixed with an equal volume of formamide loading dye, heated to ~80°C for ~2 min, and separated on a 15% urea-polyacrylamide gel in 0.5X TBE running buffer. RNA was visualized by phosphorimaging, and the intensities of uncleaved and cleaved RNA were quantified using ImageQuant (GE Healthcare). The fraction of RNA cleaved at each time point was plotted as a function of time, and these data were fit with a single exponential decay curve using Kaleidagraph (Synergy Software), according to the equation:

$$\text{fraction cleaved} = A \times (1 - \exp(-k \times t)),$$

where A is the amplitude of the curve, $k$ is the first-order rate constant, and t is time. In order to avoid overestimating $k$ in cases where the RNA was not quantitatively cleaved, the amplitude was fixed at one. Cleavage of the WT-crRNA repeat by Csy4-R118A/R115A and Csy4-R115A/R119A revealed biphasic kinetics, and the data were fit to a double exponential decay. The slower kinetic process may reflect a rate-limiting conformational change.

To ensure that Csy4 concentrations were saturating and that the on-rate for Csy4:RNA binding was not rate-limiting, cleavage experiments were repeated at five-fold higher enzyme concentrations and analyzed similarly. This analysis frequently returned slightly larger rate constants for RNAs with fast cleavage kinetics, which we attribute to slower quenching rates in the presence of more enzyme. Overall, rate constants for these experiments were generally within ~30% of those measured at the lower enzyme concentration. The precise nature of the rate-limiting step in our single-turnover cleavage assays is not known, and so first-order rate constants are reported as $k_{obs}$. $k_{obs}$ values presented in this chapter represent the average and standard error of the mean from three independent experiments. The average percent error for all reported $k_{obs}$ values is 4%.

Cleavage experiments with WT-Csy4 and WT-crRNA repeat at variable molar ratios were conducted at a constant RNA concentration of 10 nM (0.25 nM 5'-radiolabeled RNA, 9.75 nM unlabeled RNA) and varying Csy4 concentrations (40, 20, 10, 5, 2.5 nM) in a final volume of 88 µl. Ten-microliter aliquots were removed and quenched at 0.25, 0.5, 1, 2, 5, 10, 30, and 60 min, and analyzed as described above. In determining the concentration of unlabeled RNA, hypochromicity of the stem–loop was corrected for by first hydrolyzing the RNA to nucleotides by incubating in 3 M NaOH at 50°C for one hour. Then, absorbance spectra were recorded using a NanoDrop spectrophotometer (Thermo Scientific), and absorbance values at 260 nm were averaged and converted to molar concentrations using the calculated extinction coefficient (295,900 $M^{-1}$ $cm^{-1}$). The 50% yield observed at an enzyme:substrate molar ratio of 1:1 may reflect Csy4 dimerization (Przybilski et al., 2011) or partial specific activity of purified WT-Csy4.

## 4.3 Results

### 4.3.1 Csy4 binds the crRNA repeat stem–loop with high affinity and functions as a single-turnover catalyst

Csy4 is a specialized endoribonuclease that selects CRISPR transcripts from the cellular milieu for binding and cleavage. To determine the basis for this selectivity, we first examined the thermodynamic stability of the Csy4:RNA complex and the energetic contributions of protein:RNA interactions observed crystallographically (Fig. 4.1B). Using modified RNA substrates and/or Csy4 mutants, equilibrium dissociation constants ($K_d$) were measured using electrophoretic mobility shift assays (EMSA). The RNA substrates we tested derive from the invariant 28-nt repeat sequence found within pre-crRNAs generated from *P. aeruginosa* UCBPP-PA14 CRISPR locus 2 (Grissa et al., 2007), herein referred to as the crRNA repeat (Fig. 4.1A). We used the catalytically inactive Csy4(H29A) mutant (Haurwitz et al., 2010) for experiments focused on analyzing the effects of changes to the RNA substrate, enabling investigation of RNA binding independent of cleavage. Wild-type (WT) Csy4 and Csy4(H29A) bind a non-cleavable RNA substrate with affinities that are within three-fold of each other (Fig. 4.2A).

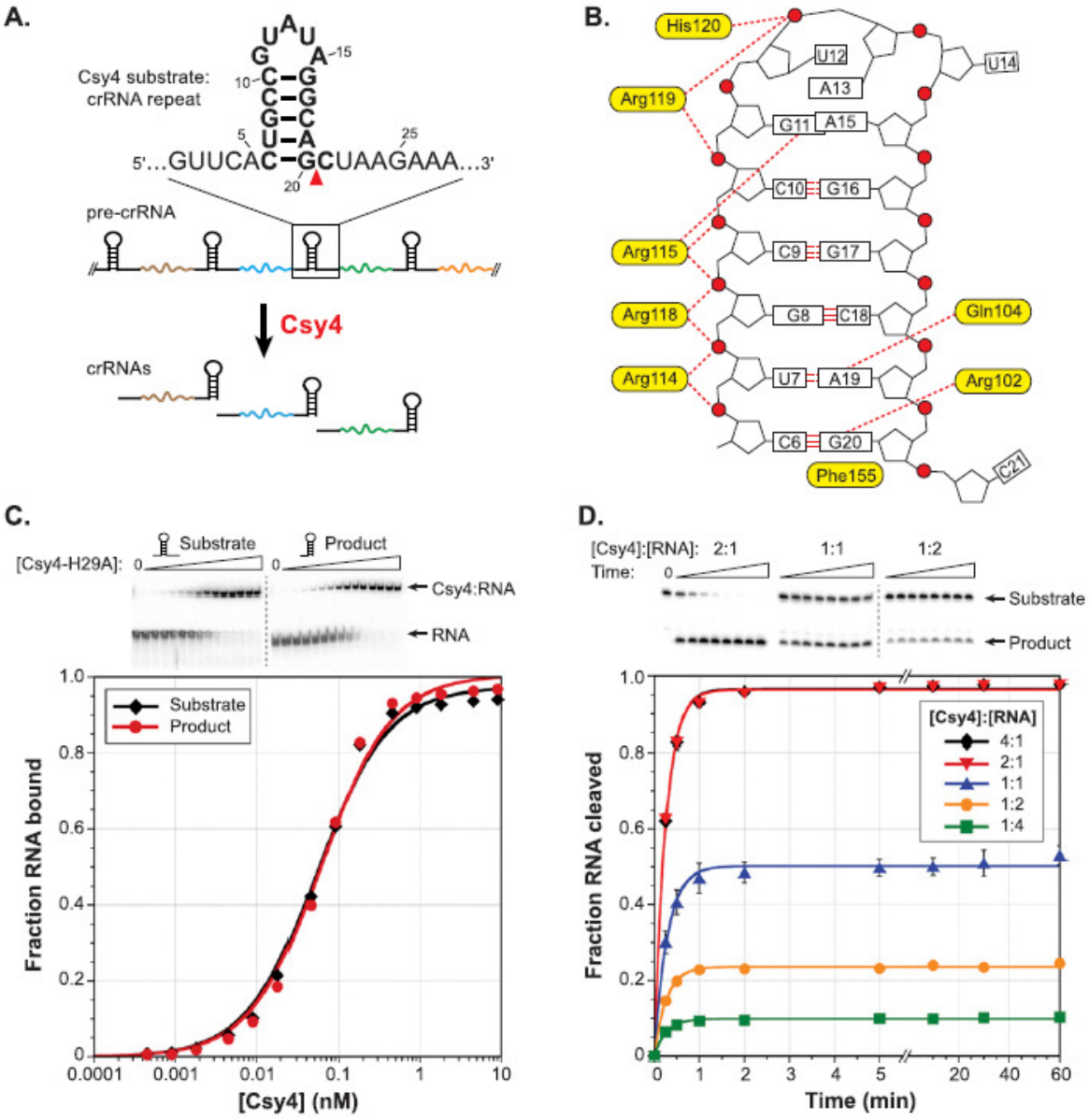



**Figure 4.1 Csy4 binds its substrate and product with high affinity and functions as a single-turnover enzyme.** (**A**) Csy4 cleaves within pre-crRNA repeat sequences (black) to generate mature crRNAs that contain a spacer sequence (colored line) flanked by fragments of the repeat. The substrate sequence and cleavage site (red triangle) are indicated above, with the crRNA substrate construct previously used for crystallography shown in bold. (**B**) A schematic depicts protein:RNA contacts revealed by a previously solved co-crystal structure of Csy4 bound to a fragment of the crRNA repeat (PDB ID: 2XLK). Important amino acid residues are shown in yellow, and RNA nucleotides are numbered as in (A). Red circles, pentagons, boxes, and red dotted lines denote phosphates, ribose groups, bases, and hydrogen-bonding interactions, respectively. (**C**) EMSAs (top) were performed with Csy4(H29A) and the substrate and product of the cleavage reaction. The resulting data for these and all subsequent binding assays were fit with a standard binding isotherm to yield equilibrium dissociation constants (solid lines; see Methods), and average $K_d$ and standard error of the mean (SEM) values from at least three independent experiments are reported in Table 4.1. (**D**) RNA cleavage assays were conducted at five different enzyme:substrate molar ratios, and the extent of the reaction at various time points was assessed by denaturing PAGE (top). The resulting data for these and all subsequent cleavage assays

were fit with a single exponential to yield first-order rate constants (solid lines; see Methods), and average $k_{obs}$ and SEM values from three independent experiments are reported in Table 4.1. Error bars for each time point represent the standard deviation and are not always visible.

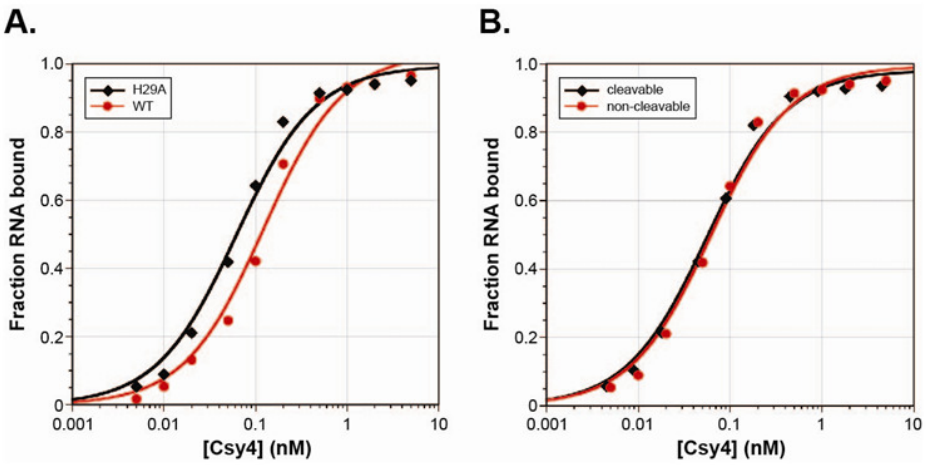



**Figure 4.2 Binding controls with Csy4(H29A) and a non-cleavable RNA substrate. (A)** EMSAs were performed with WT-Csy4 or Csy4(H29A) and a non-cleavable crRNA repeat substrate containing a deoxyribonucleotide substitution at G20. WT-Csy4 exhibits an apparent binding affinity ~3-fold lower than Csy4(H29A). **(B)** To confirm that the non-cleavable and cleavable crRNA repeat substrates are bound similarly, EMSAs were performed with Csy4(H29A) and both RNAs. The data were fit with a standard binding isotherm to yield equilibrium dissociation constants (solid lines), and average $K_d$ and (SEM) values from at least three independent experiments are reported in Table 4.1.

Strikingly, Csy4 binds the full-length, WT-crRNA repeat substrate with extremely high affinity, characterized by an equilibrium dissociation constant of ~50 pM (Fig. 4.1C; Table 4.1). Because Csy4 and the mature crRNA form part of the large Csy ribonucleoprotein complex responsible for target recognition (Wiedenheft et al., 2011b), we wondered whether Csy4 also retains high-affinity binding to the cleaved crRNA. Using a synthetic RNA corresponding to the 5' product stem–loop structure, we found that Csy4 binds this RNA indistinguishably from the substrate (Fig. 4.1C). Thus, all protein:RNA interactions contributing favorably to binding energy occur upstream of the scissile phosphate. Analysis of substrates truncated in the 5' ssRNA region allowed us to further demonstrate that nucleotides 1–4 of the crRNA repeat are completely dispensable for binding (Fig. 4.3A), indicating that the high affinity interaction we observe requires only the 15-nt stem–loop and one upstream nucleotide. We observed binding defects when A5 was mutated, suggesting that it might be specifically recognized. Indeed, a crystal structure of a Csy4:product RNA complex containing nucleotides 2–20 of the crRNA repeat sequence revealed base-specific hydrogen bonds between the Watson-Crick face of A5 and the peptide backbone of Leu139 (Fig. 4.3B, PDB ID 4AL5).

| RNA | $K_d$ (nM)[a] | $K_{d,rel}$[b] | ΔΔG (kcal/mol)[c] | $k_{obs}$ ($min^{-1}$)[d] | $k_{obs,rel}$[e] |
|---|---|---|---|---|---|
| WT (synthetic) | 0.050 ± 0.006 | 1.1 | 0.1 ± 0.1 | 3.8 ± 0.1 | 1.2 |
| WT (transcribed)[†] | 0.045 ± 0.009 | 1.0 | NA | 4.50 ± 0.06 (500 nM) | 1.0 |
| | | | | 3.98 ± 0.03 (40 nM) | 1.1 |
| | | | | 4.0 ± 0.1 (20 nM) | 1.1 |
| | | | | 3.5 ± 0.2 (10 nM) | 1.3 |
| | | | | 3.8 ± 0.2 (5 nM) | 1.2 |
| | | | | 3.90 ± 0.06 (2.5 nM) | 1.2 |
| Δ21-28 (product) | 0.049 ± 0.006 | 1.1 | 0.0 ± 0.1 | NA | NA |
| Δ1-5 | 0.09 ± 0.01 | 2.0 | 0.4 ± 0.1 | 2.93 ± 0.07 | 1.6 |
| Δ1-4 | 0.047 ± 0.005 | 1.0 | 0.0 ± 0.1 | 2.95 ± 0.08 | 1.6 |
| A5T | 0.216 ± 0.009 | 4.8 | 0.9 ± 0.1 | 2.4 ± 0.1 | 1.9 |
| Reverse complement (rc) | 5600 ± 400 | 120,000 | 6.9 ± 0.1 | 0.0057 ± 0.0004 | 790 |
| GUGUA loop (A13G) | 0.05 ± 0.01 | 1.2 | 0.1 ± 0.2 | 4.6 ± 0.2 | 0.97 |
| UAUAC loop (G11U,U12A,A13U,U14A,A15C) | 337 ± 3 | 7,400 | 5.3 ± 0.1 | 2.57 ± 0.08 | 1.8 |
| UUCG loop (G11U,A13C,U14G,ΔA15) | 2000 ± 400 | 45,000 | 6.3 ± 0.2 | 1.90 ± 0.09 | 2.4 |
| AAAAA loop (G11A,U12A,U14A) | 700 ± 100 | 15,000 | 5.7 ± 0.2 | 2.9 ± 0.2 | 1.6 |
| Nicked (between U12 and A13) | 108 ± 4 | 2,400 | 4.6 ± 0.1 | ND | ND |
| G–C, 1st base pair (C6G,G20C) | 54 ± 3 | 1,200 | 4.2 ± 0.1 | 0.00060 ± 0.00002 | 7,500 |
| U–A, 1st base pair (C6U,G20A) | 0.9 ± 0.1 | 20 | 1.8 ± 0.1 | 0.0272 ± 0.0005 | 170 |
| A–U, 1st base pair (C6A,G20U) | 0.211 ± 0.005 | 4.7 | 0.9 ± 0.1 | 0.037 ± 0.002 | 120 |
| C–G, 2nd base pair (U7C,A19G) | 35 ± 2 | 760 | 3.9 ± 0.1 | 1.64 ± 0.06 | 2.8 |
| G–C, 2nd base pair (U7G,A19C) | 1.5 ± 0.1 | 33 | 2.1 ± 0.1 | 1.17 ± 0.02 | 3.8 |
| A–U, 2nd base pair (U7A,A19U) | 0.60 ± 0.09 | 13 | 1.5 ± 0.2 | 1.20 ± 0.09 | 3.8 |

| | | | | | |
|---|---|---|---|---|---|
| C–G, 3rd base pair (G8C,C18G) | 2.10 ± 0.3 | 64 | 1.5 ± 0.1 | 1.50 ± 0.01 | 3.0 |
| A–U, 3rd base pair (G8A,C18U) | 0.12 ± 0.03 | 2.7 | 0.6 ± 0.2 | 4.37 ± 0.07 | 1.0 |
| U–A, 3rd base pair (G8U,C18A) | 0.103 ± 0.002 | 2.3 | 0.5 ± 0.1 | 1.58 ± 0.04 | 2.9 |
| G–C, 4th base pair (C9G,G17C) | 0.63 ± 0.02 | 14 | 1.6 ± 0.1 | 3.87 ± 0.07 | 1.2 |
| A–U, 4th base pair (C9A,G17U) | 0.25 ± 0.01 | 5.5 | 1.0 ± 0.1 | 4.40 ± 0.06 | 1.0 |
| U–A, 4th base pair (C9U,G17A) | 0.15 ± 0.03 | 3.3 | 0.7 ± 0.2 | 4.17 ± 0.09 | 1.1 |
| G–C, 5th base pair (C10G,G16C) | 3.9 ± 0.5 | 85 | 2.7 ± 0.1 | 3.2 ± 0.1 | 1.4 |
| A–U, 5th base pair (C10A,G16U) | 0.40 ± 0.03 | 8.8 | 1.3 ± 0.1 | 2.97 ± 0.09 | 1.6 |
| U–A, 5th base pair (C10U,G16A) | 0.09 ± 0.02 | 2.0 | 0.4 ± 0.2 | 2.05 ± 0.04 | 2.2 |
| Mutate 3 base pairs, #1 (G8A,C9U,C10U,G16A,G17A,C18U) | 0.31 ± 0.02 | 6.9 | 1.2 ± 0.1 | 2.4 ± 0.1 | 1.9 |
| Mutate 3 base pairs, #2 (G8U,C9A,C10A,G16U,G17U,C18A) | 7.0 ± 0.3 | 160 | 3.0 ± 0.1 | 0.54 ± 0.02 | 8.3 |
| Mutate 3 base pairs, #3 (G8C,C9G,C10G,G16C,G17C,C18G) | 217 ± 9 | 4,800 | 5.0 ± 0.1 | 0.312 ± 0.007 | 14 |
| C6G mismatch | 3.8 ± 0.6 | 85 | 2.7 ± 0.2 | 0.128 ± 0.003 | 35 |
| C6A mismatch | 0.057 ± 0.008 | 1.3 | 0.1 ± 0.1 | 0.456 ± 0.002 | 9.9 |
| G20A mismatch | 1.9 ± 0.3 | 41 | 2.2 ± 0.2 | 0.0003 ± 0.0001 | 14,000 |
| G20C mismatch | 3.67 ± 0.09 | 81 | 2.7 ± 0.1 | 0.00020 ± 0.00002 | 23,000 |
| 1 extra G–C, top of stem | 70 ± 10 | 1,600 | 4.4 ± 0.2 | 2.84 ± 0.06 | 1.6 |
| 2 extra G–C, top of stem | 2200 ± 200 | 49,000 | 6.4 ± 0.1 | 0.28 ± 0.01 | 16 |
| 5 extra G–C, top of stem | 4000 ± 1000 | 91,000 | 6.8 ± 0.2 | 0.083 ± 0.001 | 54 |
| 5 extra G–C, top of stem 3' A bulge | 253 ± 3 | 5,600 | 5.1 ± 0.1 | ND | ND |
| 5 extra G–C, top of stem 3' AA bulge | 26 ± 2 | 570 | 3.8 ± 0.1 | ND | ND |

| | | | | | |
|---|---|---|---|---|---|
| 5 extra G–C, top of stem<br>3' AAA bulge | 10.5 ± 0.8 | 230 | 3.2 ± 0.1 | 2.88 ± 0.06 | 1.6 |
| 1 extra A–U, bottom of stem | 0.061 ± 0.003 | 1.3 | 0.2 ± 0.1 | 2.25 ± 0.02 | 2.0 |
| 2 extra A–U, bottom of stem | 0.57 ± 0.02 | 13 | 1.5 ± 0.1 | 2.913 ± 0.009 | 1.6 |
| 1 extra G–C, bottom of stem | 9.6 ± 0.8 | 210 | 3.2 ± 0.1 | 0.102 ± 0.002 | 44 |
| 2 extra G–C, bottom of stem | 36 ± 1 | 790 | 4.0 ± 0.1 | 0.0028 ± 0.0002 | 1,600 |

**Table 4.1 Binding and cleavage data for mutant crRNA repeat substrates.**

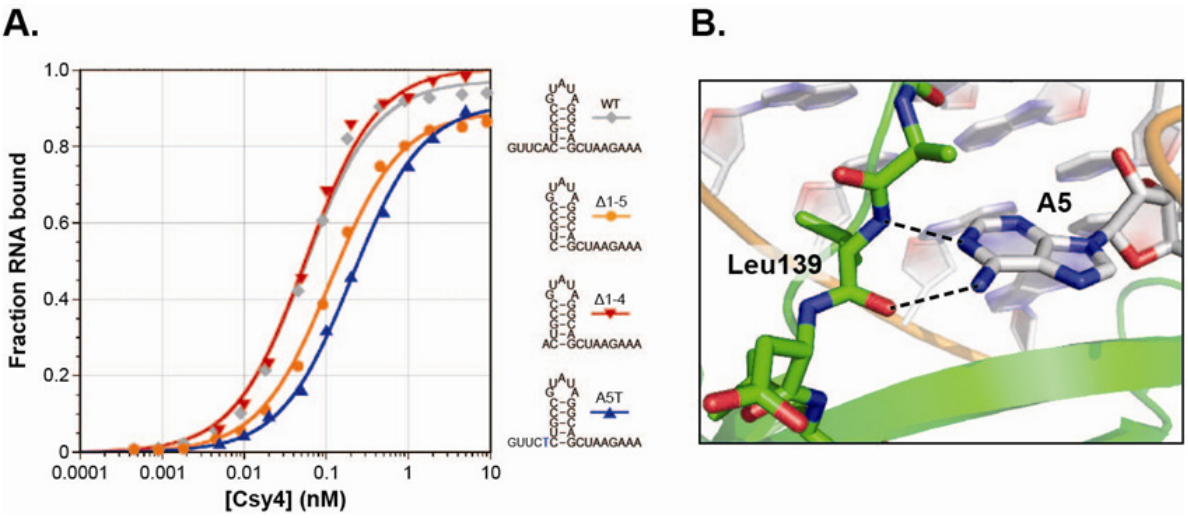



**Figure 4.3 Sequence-specific recognition of A5 by Csy4.** (**A**) To determine the contributions of single-stranded RNA (ssRNA) nucleotides upstream of the stem-loop to overall binding energy, EMSAs were performed with Csy4(H29A) and RNA substrates containing either deletions (Δ1-4, Δ1-5) or a mutation (A5T) in the first 5 ssRNA nucleotides. The 2-fold binding defect observed with Δ1-5 RNA relative to the WT-crRNA repeat was abolished when A5 was reintroduced (Δ1-4), indicating that A5 is the only ssRNA nucleotide bound by Csy4. The even larger magnitude binding defect (~5-fold) with a substrate mutated at this position (A5T) suggests that this binding is sequence-specific. **(B)** Csy4(S22C) was crystallized bound to a product RNA containing nucleotides 2-20 of the WT-crRNA repeat (PDB ID 4AL5). While nucleotides 2-4 are disordered and not visible in the electron density, A5 interacts with the peptide backbone of Leu139 via two base-specific hydrogen bonds (dashed lines) to N1 and N6 at the Watson-Crick edge.

Considering the retention of Csy4 and crRNA in the Csy complex (Wiedenheft et al., 2011b), we speculated that tight association of Csy4 with its product may be an intrinsic mechanistic feature of Csy4 during crRNA biogenesis in type I-F CRISPR systems. To test this hypothesis, we carried out cleavage assays at a range of enzyme:substrate molar ratios and monitored both the rate and yield of product formation. As seen in Figure 4.1D, Csy4 completely lacks the ability to engage in multiple-turnover catalysis. The overall yield of the cleavage reaction remained directly proportional to the Csy4 concentration when present in sub-stoichiometric amounts relative to substrate, even with incubation times >200-fold longer than the reaction time constant. All time courses fit well to a single exponential decay and yielded uniform, first-order observed rate constants ($k_{obs}$; Table 4.1), which would only be the case in the absence of multiple-turnover behavior under conditions where the on-rate is not rate-limiting. These observations indicate that Csy4 remains product-bound after the reaction and is thereby strongly inhibited from performing additional rounds of RNA cleavage. crRNA repeat cleavage reached only 50% completion at an enzyme:substrate molar ratio of 1:1. A recent study used a two-hybrid system to demonstrate that Csy4 can interact with itself, but this result could not be repeated for all fusion constructs (Przybilski et al., 2011). While we cannot formally exclude the possibility that Csy4 might function as a dimer with one inactive subunit, our gel filtration experiments are consistent with purified Csy4 existing as a monomer (data not shown). Therefore, we speculate that the incomplete cleavage we observe reflects partial specific activity of purified WT-Csy4.

### 4.3.2 Protein determinants of high-affinity crRNA repeat binding and cleavage

The high-affinity interaction between Csy4 and the crRNA repeat substrate is tighter than many protein:RNA complexes studied to date. We were therefore interested in gaining a detailed understanding of the primary sources of binding energy, as informed by interactions identified from our crystal structure. We began by focusing on the bottom of the RNA stem, where the side-chains of Arg102 and Gln104 are each involved in two sequence-specific hydrogen bonds with the major groove faces of G20 and A19, respectively. Using a synthetic, non-cleavable substrate that is bound indistinguishably from the WT-crRNA repeat (Fig. 4.2B), EMSAs with Csy4(R102A) and Csy4(Q104A) mutants revealed that the binding energies contributed by these amino acids are quite distinct. The crRNA repeat binds >2000-fold more weakly to Csy4(R102A), representing a $\Delta\Delta G$ of 4.6 kcal/mol, whereas RNA binding by Csy4(Q104A) is destabilized by only 1.4 kcal/mol relative to WT (Fig. 4.4A; Table 4.2). This difference may be explained in part by the expected +1 charge on the arginine's guanidinium group at physiological pH. Whereas deletion of an uncharged hydrogen bond typically weakens binding between enzyme and substrate by 0.5–1.8 kcal/mol, charged hydrogen bonds generally contribute some 3–6 kcal/mol binding energy (Fersht, 1987), in good agreement with our data.

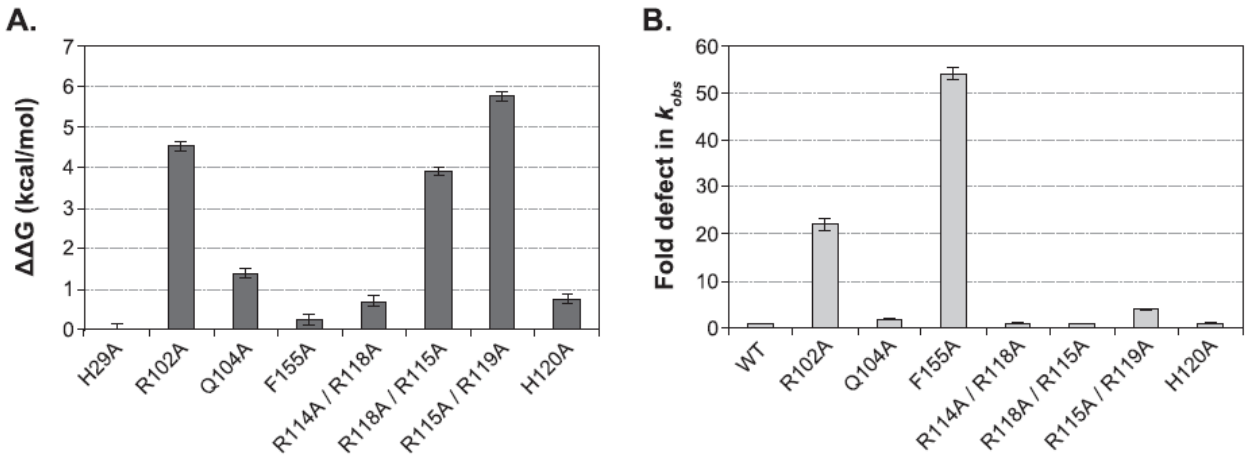



**Figure 4.4 Amino acid contributions to binding energy and cleavage kinetics.** (**A**) Csy4 residues involved in base-pair recognition and phosphate backbone contacts were mutated to alanine in order to assess their energetic contributions to binding. EMSAs were performed with a non-cleavable crRNA repeat substrate containing a deoxyribonucleotide substitution at G20, and binding defects relative to Csy4(H29A) were determined and converted to ΔΔG values (T = 298 K). Plotted are the average and SEM from at least three independent experiments. (**B**) First-order rate constants ($k_{obs}$) for WT-crRNA repeat cleavage by each Csy4 mutant were determined. Cleavage data for R118A/R115A and R115A/R119A mutants showed biphasic kinetics and were fit with a double exponential decay to yield two rate constants (Table 4.2), the faster of which is shown. Plotted are the average fold defects (relative to WT-Csy4) and SEM from three independent experiments. Average $K_d$, $k_{obs}$, and SEM values are reported in Table 4.2.

| Csy4 | $K_d$ (nM)[a] | $K_{d,rel}$[b] | ΔΔG (kcal/mol)[c] | $k_{obs}$ ($min^{-1}$)[d] | $k_{obs,rel}$[e] |
|---|---|---|---|---|---|
| **WT** | 0.132 ± 0.009 | 2.10 | 0.6 ± 0.1 | 4.50 ± 0.06 | 1.0 |
| **H29A** | 0.045 ± 0.009 | 1.0 | NA | NA | NA |
| **R102A** | 98 ± 5 | 2,200 | 4.6 ± 0.1 | 0.20 ± 0.01 | 22 |
| **Q104A** | 0.48 ± 0.02 | 11 | 1.4 ± 0.1 | 2.170 ± 0.006 | 2.1 |
| **F155A** | 0.068 ± 0.006 | 1.5 | 0.2 ± 0.1 | 0.083 ± 0.002 | 54 |
| **R114A/R118A** | 0.15 ± 0.01 | 3.2 | 0.7 ± 0.1 | 3.70 ± 0.06 | 1.2 |
| **R118A/R115A**† | 34 ± 2 | 740 | 3.9 ± 0.1 | 3.9 ± 0.2 (0.31) | 1.2 |
| | | | | 0.008 ± 0.001 (0.52) | 560 |
| **R115A/R119A**† | 780 ± 50 | 17,000 | 5.8 ± 0.1 | 1.14 ± 0.03 (0.78) | 3.9 |
| | | | | 0.035 ± 0.004 (0.20) | 130 |
| **H120A** | 0.16 ± 0.01 | 3.6 | 0.8 ± 0.1 | 3.83 ± 0.03 | 1.2 |

**Table 4.2 Binding and cleavage data for Csy4 mutants.**

[a]Reported as the average and standard error of the mean (SEM) from at least three independent experiments. All binding experiments were performed with the non-cleavable substrate containing a deoxyribonucleotide substitution at G20.

[b]Calculated by dividing each $K_d$ value by the $K_d$ for Csy4(H29A) (0.045 nM).

[c]Reported as the average and SEM, and calculated according to the equation:
$\Delta\Delta G = -RT\ln(K_{d,H29A}/K_{d,mutant})$

[d]Reported as the average and SEM from three independent experiments. All cleavage experiments were performed with the *in vitro* transcribed WT-crRNA repeat substrate.

[e]Calculated by dividing $k_{obs}$ for WT-Csy4 (4.50 $min^{-1}$) by the $k_{obs}$ value for each Csy4 mutant.

†Cleavage data showed biphasic kinetics and were fit with a double exponential decay. Both rate constants and their respective amplitudes (in parentheses) are reported.

NA, not applicable.

In addition to its interaction with Arg102, G20 of the crRNA repeat stacks onto the aromatic side-chain of Phe155. Stacking interactions between aromatic amino acids and nucleotides can contribute up to 5.5 kcal/mol of binding energy (Auweter et al., 2006; Nolan et al., 1999), but we were surprised to observe a negligible 1.5-fold binding defect ($\Delta\Delta G$ = 0.2 kcal/mol) with a Csy4(F155A) mutant (Fig. 4.4A). Given the pre-crRNA processing defects we observed previously with Csy4(F155A) (Haurwitz et al. 2010), these data suggest that Phe155 instead plays a role in achieving rapid cleavage kinetics. Indeed, under single-turnover conditions with saturating enzyme concentrations (see Methods), the F155A mutant led to a ~50-fold reduction in the observed cleavage rate constant (Fig. 4.4B). Csy4(R102A) also exhibited a ~20-fold defect in cleavage kinetics, whereas the rate of cleavage by Csy4(Q104A) was within

2.5-fold of WT (Fig. 4.4B). Collectively, these data suggest that, independent of their effects on binding energy, Phe155 and Arg102 are important for anchoring the G20 guanine in the active site and may thereby assist in positioning the ribose for subsequent activation of its 2'-OH nucleophile.

Moving up the crRNA repeat stem, we next focused on interactions observed in the crystal structure between the RNA and residues found in the alpha helix that inserts into the major groove of the double-stranded stem (Fig. 4.1B). The guanidinium groups of Arg114, Arg115, Arg118, and Arg119 each present ≥2 hydrogen-bond donors within 3 Å of acceptors in the RNA phosphate backbone, yet their contributions to overall binding energy differ widely, as assessed through double R→A mutations. In particular, Arg114 and Arg118, which contact adjacent phosphates, contribute only 0.7 kcal/mol of binding energy, whereas alanine mutations at Arg115 and Arg119 led to a >15,000-fold binding defect ($\Delta\Delta G$ = 5.8 kcal/mol) (Fig. 4.4A). While all four residues are positioned to act as arginine forks, in that each side-chain contacts adjacent phosphates (Calnan et al., 1991), only Arg115 and Arg119 may simultaneously utilize all three nitrogen atoms of the guanidinium group as hydrogen bond donors. Arg115 hydrogen bonds to two phosphates in addition to the major groove face of G11, which forms part of the G·A sheared base pair at the bottom of the GUAUA pentaloop, and Arg119 is situated in a unique pocket of the loop where it interacts with phosphates separated by two nucleotides. His120 also interacts with a phosphate at the apex of the loop and contributes 0.8 kcal/mol of binding energy (Fig. 4.4A). The specific network of multidentate contacts between the arginine-rich helix and the RNA stem–loop suggests that high-affinity binding to the crRNA repeat is highly shape-specific, especially with regard to the tertiary structure of the loop. The large magnitude of the binding energy contributed by this protein helix enables Csy4 to maintain a tight grip on the substrate and product, but this interaction is not required for catalytic activity. Cleavage rates for the H120A and R→A mutants under saturating conditions were within five-fold of WT-Csy4 (Fig. 4.4B).

### 4.3.3 High-affinity crRNA repeat binding is sensitive to the loop structure

The direction of CRISPR loci transcription in *P. aeruginosa* has not been directly analyzed, and a recent report that detected mature crRNAs by Northern blot analysis used dsDNA probes that were not strand-specific (Cady and O'Toole, 2011). Transcription in a direction opposite to that of our own predictions would generate pre-crRNAs containing the reverse complement of the crRNA repeat sequence. To determine whether Csy4 also recognizes and cleaves this potential substrate, we generated the reverse complement crRNA (rc-crRNA) repeat by *in vitro* transcription and tested its affinity for Csy4(H29A). We found that the rc-crRNA repeat binds Csy4 $>10^5$-fold weaker than the WT-crRNA repeat (Fig. 4.5A) and is cleaved >750-fold slower (Fig. 4.6A), strongly suggesting that the genuine Csy4 substrate *in vivo* is pre-crRNA transcribed in an orientation consistent with our previous work (Haurwitz et al., 2010). Northern blot analysis using single-stranded probes indeed confirmed the presence of crRNAs in *P. aeruginosa* UCBPP-PA14 with the repeat sequence we define in Figure 4.1A, but failed to detect transcripts from the opposite strand (Fig. 4.7).

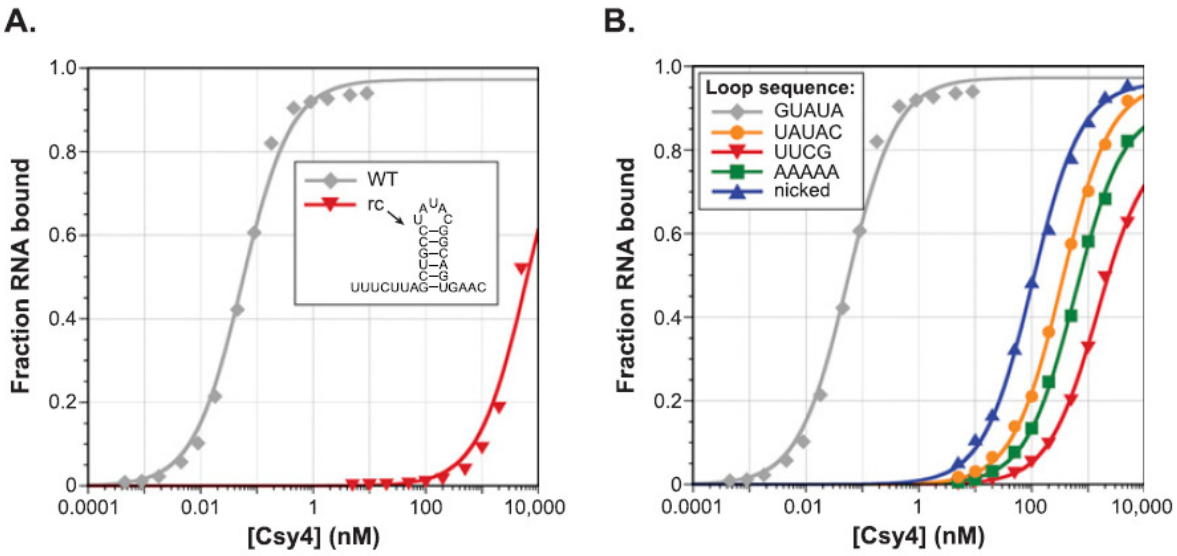



**Figure 4.5 Importance of the loop sequence for high-affinity RNA binding.** (**A**) EMSAs demonstrate that Csy4 binds the reverse complement of the crRNA repeat (rc) >$10^5$-fold more weakly than the WT-crRNA repeat. (**B**) Mutant RNA substrates were generated by changing the WT loop sequence (GUAUA) to a quintuple mutant (UAUAC), the highly stable UUCG tetraloop, or a poly(A) pentaloop, or by removing the loop through use of a substrate nicked between U12 and A13. EMSAs reveal substantial defects associated with binding these mutant RNAs.

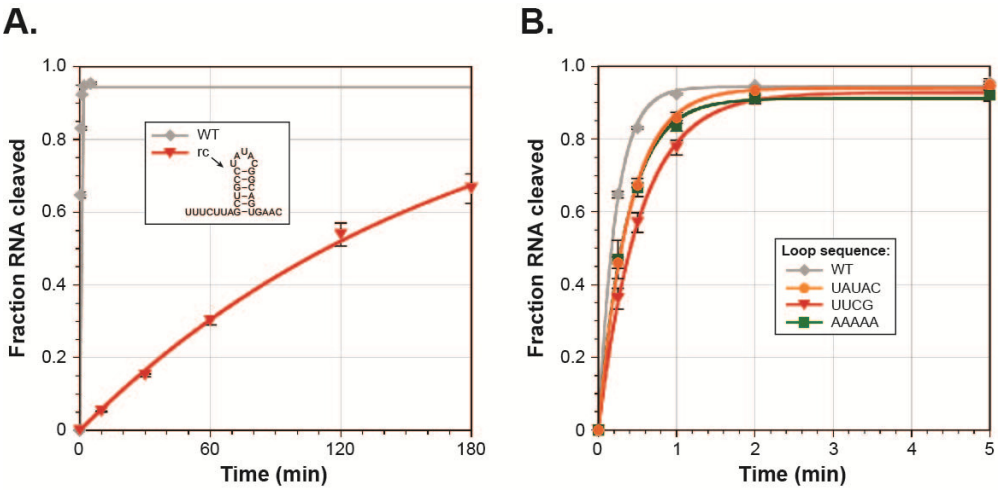



**Figure 4.6 Cleavage of rc-crRNA repeat and loop mutant substrates.** Cleavage assays were performed with the reverse complement (rc) crRNA repeat (**A**) and RNA substrates containing mutated loop sequences (**B**). The rc-crRNA repeat substrate was cleaved >750-fold slower than the WT-crRNA repeat substrate, whereas substrates containing mutations only in the loop sequence were cleaved at rates within 2.5-fold of WT. For these and all subsequent cleavage assays, the data were fit with a single exponential to yield first-order rate constants (solid lines), and average $k_{obs}$ and SEM values from three independent experiments are reported in Table 4.1.

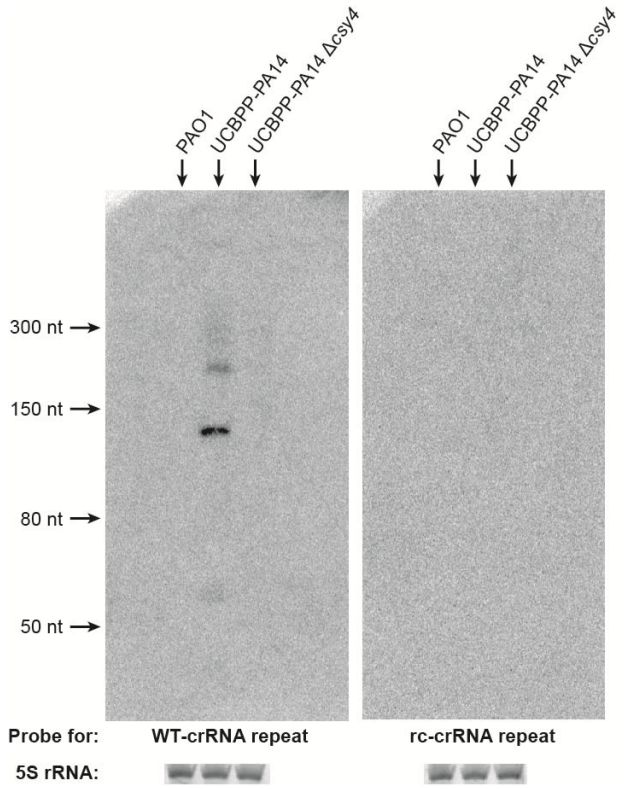



**Figure 4.7 Northern blot analysis of crRNAs in *P. aeruginosa*.** Total RNA was extracted from strains of *P. aeruginosa* without CRISPRs (PAO1), with a complete CRISPR-Cas locus (UCBPP-PA14), or with a CRISPR-Cas locus harboring a Csy4 gene deletion. Duplicates of each RNA preparation were separated by 15% denaturing PAGE, transferred to nylon membranes, and probed with DNA oligonucleotides complementary to either the WT-crRNA repeat (left) or the rc-crRNA repeat (right). The gel was stained with SYBR Gold before transfer, and the 5S RNA band is shown as a loading control (bottom). RNAs containing the WT-crRNA repeat but not the reverse complement were detected in *P. aeruginosa* UCBPP-PA14. The laddering pattern is consistent with precursor transcripts that were incompletely processed, thereby yielding multiples of the length of a mature crRNA (60 nucleotides). Hybridization to the mature crRNA is likely to be less efficient than to partially processed species because the probe is complementary to only 20 out of 28 nucleotides. The precursor transcript may be prone to rapid degradation in the absence of Csy4, explaining why the pre-crRNA band in the Δ*csy4* strain is not more prominent.

When comparing the two RNA sequences, the rc-crRNA repeat contains the identical five-base-pair stem sequence as the WT-crRNA repeat but with an additional predicted G–U wobble base pair below and different loop and flanking ssRNA sequences, indicating that one or more of these regions are specifically recognized by Csy4. Having already demonstrated the negligible binding defects resulting from deletion of flanking ssRNA nucleotides, we suspected that destabilized binding of the rc-crRNA repeat resulted primarily from the inability of Csy4 to interact productively with the UAUAC loop sequence and/or the unique tertiary structure it would impose on the RNA substrate. The GUAUA loop encoded by CRISPR locus 2 in *P. aeruginosa*

UCBPP-PA14 forms a GNR(N)A pentaloop structure (Legault et al., 1998), in which U14 flips out of the loop to enable a GNRA tetraloop fold that involves sequential stacking of U12, A13, and A15 on the 3' strand of the stem (Haurwitz et al., 2010). The CRISPR 3 locus encodes a GUGUA loop in the repeat sequence that is predicted to form the same pentaloop structure, and this crRNA structure is bound and cleaved indistinguishably from the substrate with a GUAUA loop (Fig. 4.8). We hypothesized that Csy4 specifically recognizes this loop motif, and that other loop sequences unable to conform to a GNRA tetraloop fold would bind much more weakly.

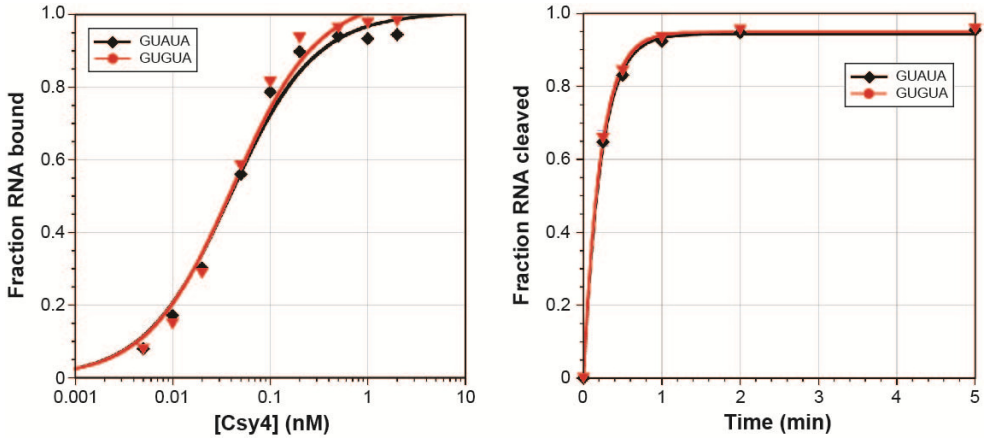



**Figure 4.8 Recognition of a crRNA repeat containing a GUGUA loop.** The CRISPR 2 locus in *P. aeruginosa* UCBPP-PA14 encodes a crRNA repeat hairpin with a GUAUA loop, whereas the CRISPR locus 3 encodes a hairpin with a GUGUA loop. Because both are likely to adopt GNR(N)A pentaloop folds, we suspected that Csy4 would not discriminate between the two substrates. Indeed, binding (left) and cleavage (right) assays of crRNA repeat substrates containing either loop sequence reveal indistinguishable biochemical behaviors.

To test this, we generated a panel of RNA substrates containing mutated loop sequences and tested their affinity for Csy4(H29A). In agreement with our hypothesis, Csy4 bound to each RNA at least 7000-fold more weakly than WT (Fig. 4.5B). To test whether a loop was strictly required, we formed a nicked RNA substrate from two RNA oligonucleotides annealed in trans. Csy4 is able to cleave to cleave the nicked substrate (Fig. 4.9) and this substrate interacted more favorably with Csy4 than those containing a non-GNRA-like loop (Fig. 4.5B; Fig. 4.10). These experiments confirm that high-affinity Csy4 binding relies in part on a precise substrate tertiary structure in the loop region, independently of base-specific contacts, and that the absence of a loop altogether is less detrimental to binding than the presence of a nonnative loop. It is interesting to note that, despite their weakened binding, RNAs with mutated loops were cleaved at rates within 2.5-fold of the WT-crRNA repeat at saturating Csy4 concentrations (Fig. 4.6B). This was true even for a substrate containing the same loop (UAUAC) as the rc-crRNA

repeat, which had a >750-fold defect in $k_{obs}$. Since the stacking interaction between the terminal C–G base pair and the aromatic side chain of Phe155 is important for cleavage (Fig. 4.4B), we suspected that the additional base pair below the WT stem in the rc-crRNA repeat might impede Csy4 activity (see below).

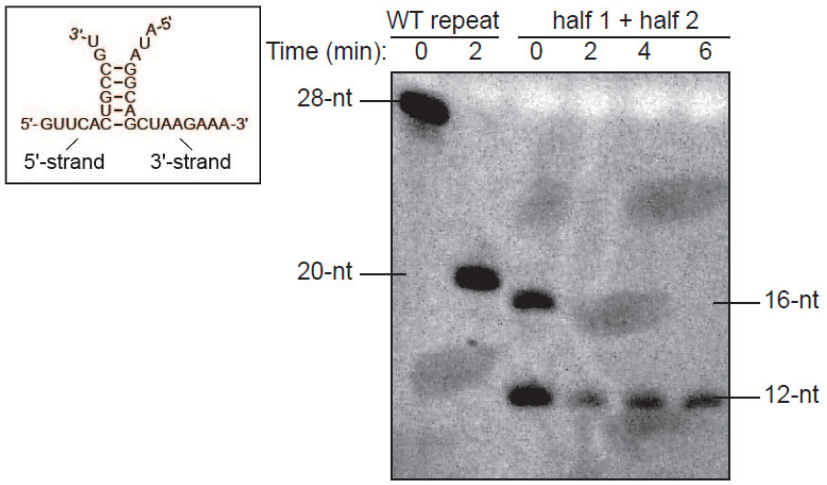



**Figure 4.9 Csy4 can cleave a nicked RNA substrate.** We generated a nicked RNA substrate by annealing two synthetic oligo nucleotides comprising the left and right halves of the wild-type crRNA repeat (left box). We incubated 75 pmol of Csy4 with 50 pmol of either WT repeat or the annealed oligonucleotides for the indicated times, extracted the products with acid phenol:chloroform, and visualized the products with SYBR Gold staining on a 20% denaturing gel.

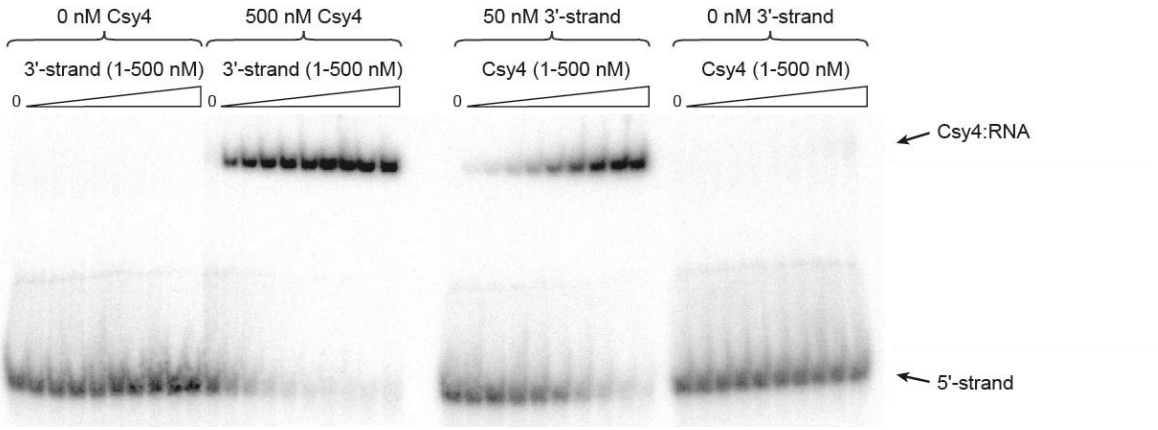



**Figure 4.10 Binding controls with a nicked crRNA repeat substrate.** An RNA substrate was generated using two synthetic oligonucleotides constituting nucleotides 1-12 (5'-strand) and 13-28 (3'-strand) of the WT-crRNA repeat substrate (see boxed inset, right), and EMSAs were performed with Csy4(H29A). The 5'-strand was [$^{32}$P]-radiolabeled for these experiments and present at 0.5 nM in all binding reactions. To confirm that the observed electrophoretic mobility gel shift represented Csy4 bound to the hybridized two-strand duplex, experiments were performed that increased the 5'/3'-strand molar ratio with and without Csy4 present (left), or that increased the Csy4 concentration with and without the 3'-strand present (right). No shift was observed in the absence of Csy4, indicating either that the hybridized substrate does not stably form without Csy4 or that it does not shift relative to the 5'-strand alone. Additionally, Csy4 does not bind the 5'-strand alone at the highest concentrations tested (500 nM). These data indicate that binding requires the double-stranded substrate and that Csy4 may trap a hybridized duplex that is thermodynamically unstable under these experimental conditions.

### 4.3.4 Specificity within the crRNA repeat stem sequence during binding and cleavage

We were particularly interested in investigating the ability of Csy4 to discriminate between substrates containing the cognate five base pairs in the stem and those with similar but noncognate sequences. We therefore made all individual Watson-Crick base-pair substitutions at each position in the double-stranded stem and determined the energetic costs associated with binding each mutant RNA substrate relative to the WT-crRNA repeat using EMSAs (Fig. 4.11A). The data reveal that base-pair changes throughout the stem result in varying degrees of Csy4:RNA complex destabilization, ranging from 0.4 to 4.2 kcal/mol. The largest defects result from G–C and C–G substitutions at the ultimate and penultimate base pair, respectively, where Arg102 and Gln104 provide a direct readout mechanism of recognition and confer similar degrees of discrimination in spite of their unequal contributions to binding energy. To confirm this, we repeated binding experiments with RNA substrates containing substitutions at the bottom two base pairs using either Csy4(R102A) or Csy4(Q104A) (Fig. 4.11A). As expected, the overall specificity for particular base pairs at either position is lost when the amino acid specificity determinant is absent. The Csy4:RNA co-crystal structure did not reveal sequence- specific contacts with Watson- Crick base pairs in the upper part of the double-stranded stem (Haurwitz et al., 2010), but we observed substantial energetic penalties for binding substrates with base-pair substitutions in this region (Fig. 4.11A). Furthermore, the magnitude of these binding defects was highly sequence-dependent; when multiple base-pair substitutions were made in the top three base pairs simultaneously, binding defects ranged from seven- to almost 5000-fold (Fig. 4.12), with the largest destabilization occurring when each C–G pair was mutated to its complement. These results reveal that substrate sequence specificity is mediated by Csy4 via a mechanism that does not rely exclusively on base-specific interactions.

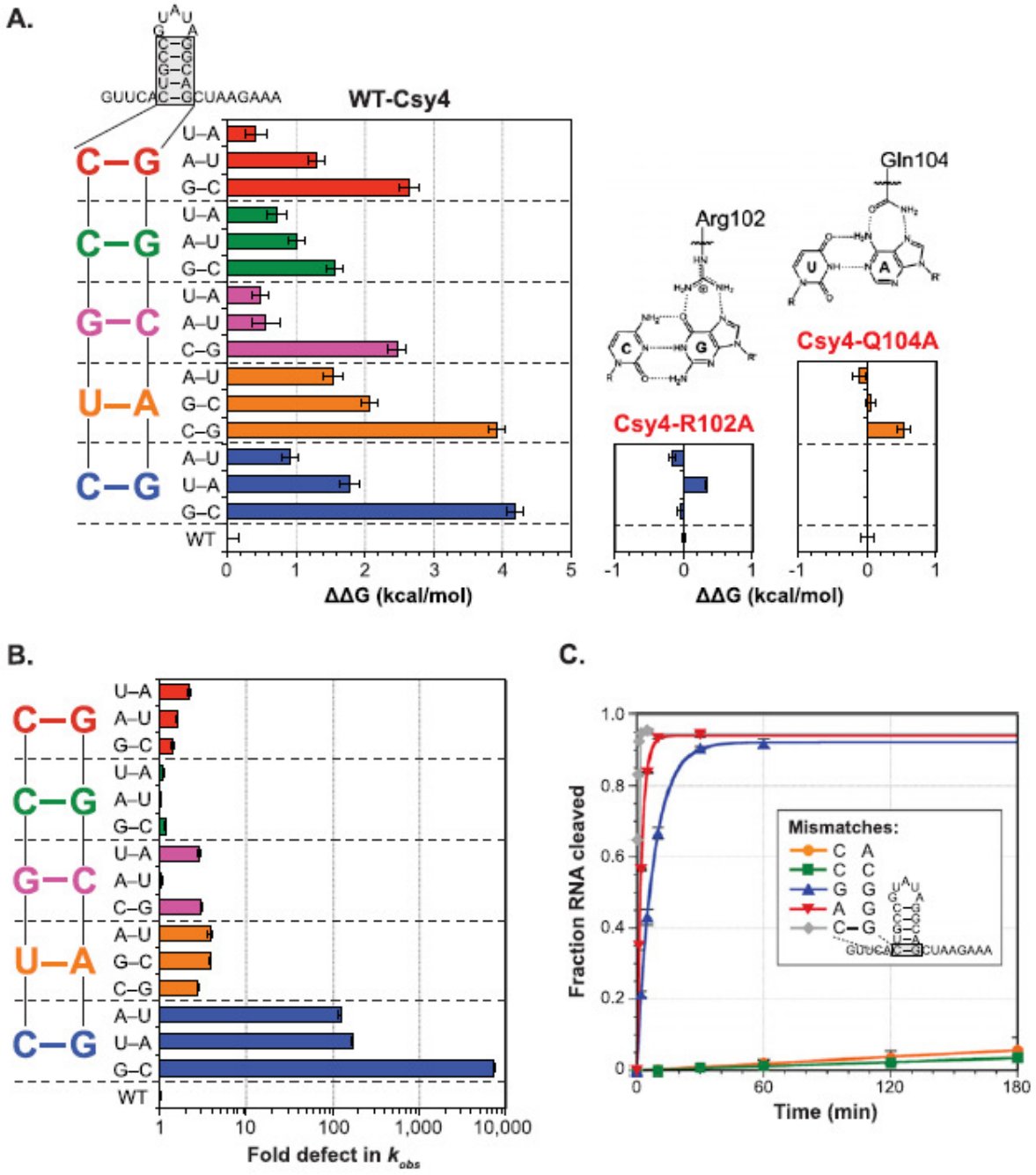



**Figure 4.11 Substrate specificity within the crRNA repeat stem.** (**A**) A library of mutated crRNA repeat substrates was generated containing all possible Watson-Crick base-pair substitutions at each position in the double-stranded stem. EMSAs were performed with Csy4(H29A) and these RNA substrates, and the resulting binding defects relative to WT-crRNA repeat were determined and converted to ΔΔG values (T = 298 K). The WT stem sequence is shown at the left, with data for base-pair substitutions at each position color-coded similarly. Binding experiments with RNAs mutated at the bottom C–G or U–A base pair were repeated with Csy4(R102A) (middle) or Csy4(Q104A) (right), respectively; ΔΔG values were calculated relative to WT-crRNA repeat binding by each Csy4 mutant. Shown above are chemical structures of the interactions made by Arg102 and Gln104 with the WT base pairs. (**B**) Single-turnover cleavage assays were performed with the same library of RNA mutants as in A, and the resulting defects in $k_{obs}$ relative to

WT-crRNA repeat were determined. The data are plotted as in A. (**C**) To investigate the importance of the terminal C–G base pair during cleavage, mismatched substrates were generated by mutating C6 or G20 individually and single-turnover cleavage assays were performed.

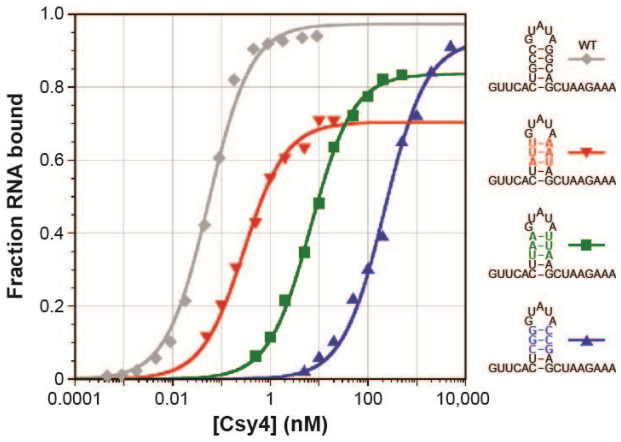



**Figure 4.12 Recognition of base pairs at the top of the stem.** In order to investigate the sequence specificity in the upper part of the stem-loop, we generated mutant crRNA repeat substrates (right) that contained consecutive substitutions in the top three base-pairs and performed EMSAs with Csy4(H29A). A construct that maintained the same purine-pyrimidine pattern (red) showed the mildest binding defect, whereas mutating C–G base pairs to their complement resulted in a ~5,000-fold defect.

We next investigated whether these specificity determinants also influence the chemical cleavage reaction. To test this, we conducted single-turnover cleavage experiments with WT-Csy4 at saturating concentrations using the same library of RNAs as in Figure 4.11A and determined the first-order rate constants for RNA cleavage ($k_{obs}$) relative to WT. In stark contrast to the observed binding specificity, rate constants governing the cleavage of RNA substrates with base pair substitutions at any position other than the terminal position were within four-fold of WT (Fig. 4.11B). However, any mutation of the terminal C–G base pair in the stem–loop was detrimental for cleavage of the crRNA repeat, with kinetic defects ranging from ~100- to 7500-fold. To further dissect the importance of the terminal C–G base pair, we generated a series of RNA substrates containing mismatches at this position by mutating either C6 or G20 independently. Cleavage time courses with these substrates (Fig. 4.11C) clearly demonstrate the importance of G20, regardless of whether or not a base pair can form at the terminal position. RNA substrates containing C6A or C6G mutations were cleaved at rates within 40-fold of WT, whereas mutation of G20 to either adenosine or cytosine led to >10,000-fold defects.

### 4.3.5 Csy4 is highly selective for stem–loops of defined length

Having interrogated Csy4 for sequence specificity throughout the crRNA repeat, we also wondered whether Csy4 is sensitive to the length of the crRNA repeat stem. To test this, we inserted one or two base pairs at the top of the duplex region and tested these substrates for binding. Strikingly, just one or two additional G–C base pairs led to 1600- and 49,000-fold weaker binding affinities, respectively (Fig. 4.13A). This was particularly surprising because the crystal structure did not immediately suggest any obvious steric clashes that would result from insertions at the top of the stem. However,

given the large energetic contribution of the arginine-rich helix to binding (Fig. 4.4C), we suspected that additional base pairs would disrupt protein–loop interactions and prevent stable docking of this helix into the major groove of the double-stranded stem. A-form dsRNA helices have deep and narrow major grooves that are generally inaccessible to proteins (Draper, 1995), but exceptions occur in proximity to helix termini or asymmetric bulges, where the major groove can widen considerably (Weeks and Crothers, 1993). We hypothesized that base-pair insertions cause narrowing of the major groove and thereby disrupt high-affinity interactions between the arginine-rich helix and crRNA repeat.

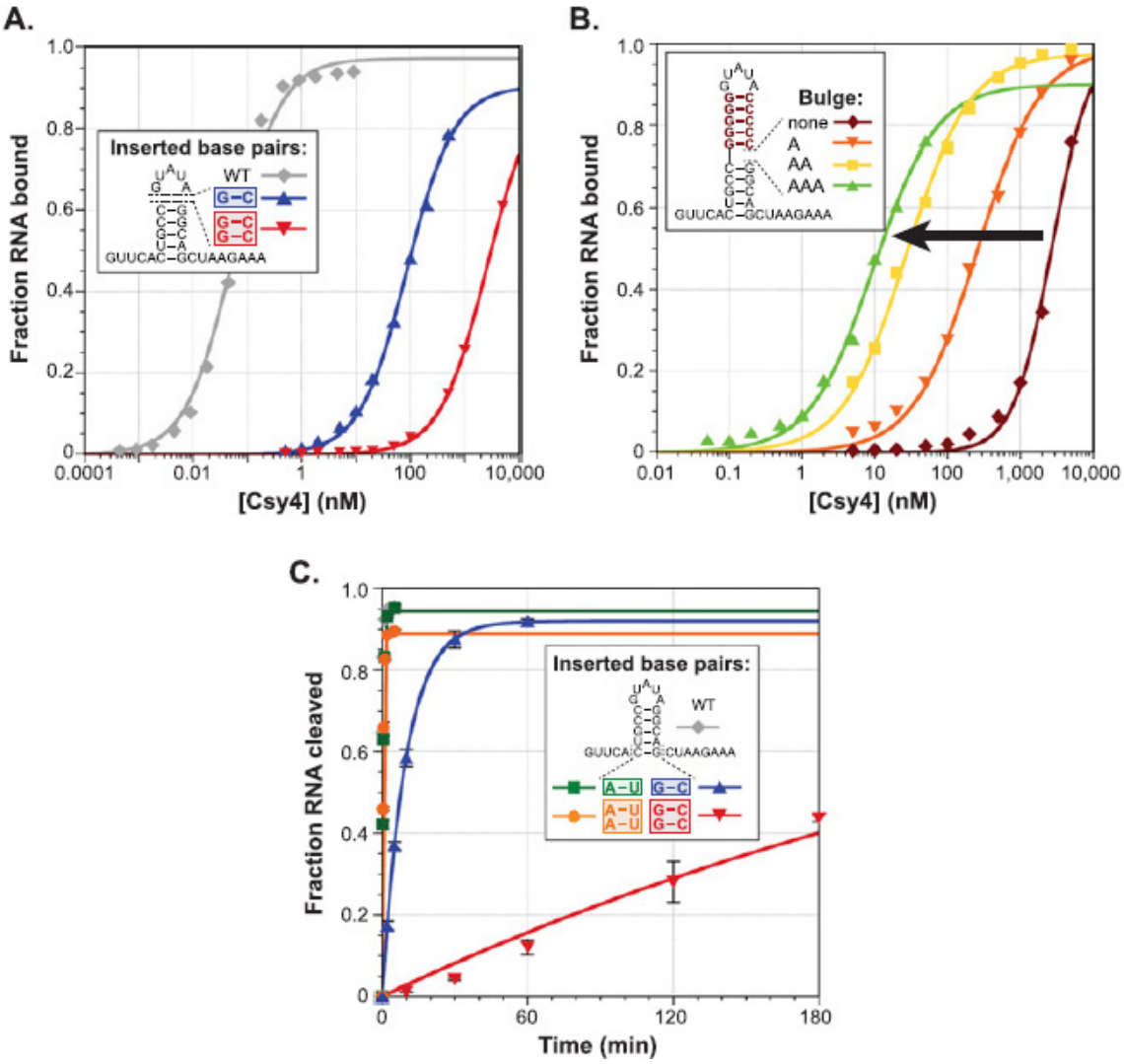



**Figure 4.13 Stem length dependence during substrate binding and cleavage.** (**A**) One or two G–C base pairs were inserted at the top of the stem between the closing C–G base pair and the GUAUA pentaloop, and EMSAs were performed. (**B**) To test the hypothesis that longer stems prevent stable binding of the arginine-rich helix via their effect on major groove accessibility, a substrate was generated that contains five G–C base pairs inserted above the WT stem. Subsequently, asymmetric adenosine bulges were inserted on the 3′ side of the duplex between the five-base-pair WT stem and the five-base-pair insertion. EMSAs reveal that binding affinities increase monotonically (black arrow) with bulges of

increasing size. (**C**) One or two G–C or A–U base pairs were inserted below the terminal C–G base pair, and cleavage time courses were performed. Additional A–U base pairs have negligible effects on $k_{obs}$, whereas two additional G–C base pairs result in ~1500-fold slower kinetics.

To test this idea, we generated an RNA construct that contains five G–C base pairs inserted atop the WT stem sequence while retaining the GUAUA pentaloop. This RNA was bound with an equilibrium dissociation constant of 4 μM (Fig. 4.13B), representing nearly a $10^5$-fold defect relative to WT. We then introduced adenosine bulges of varying size on the 3' side of the stem, at the junction between the WT five-base-pair stem sequence and the five G–C base-pair insertion. These types of asymmetric bulges within perfectly base-paired dsRNA helices have been shown previously to increase major groove accessibility progressively as a function of bulge size, as probed using diethylpyrocarbonate (DEPC) reactivity (Weeks and Crothers, 1993). In excellent agreement with our hypothesis, we found that the binding affinity of Csy4 for these bulged substrates increased in concert with bulges of increasing size (Fig. 4.13B), suggesting that major groove widening enables stable docking of the arginine-rich helix. The inability to form favorable protein–loop interactions likely explains why bulged substrates are still bound >200-fold more weakly than the WT-crRNA repeat.

We also investigated the effects of inserting one or two base pairs at the bottom of the stem–loop below the terminal C–G base pair. We observed a range of binding defects, although these were milder than those resulting from insertions at the top of the stem (Fig. 4.14A). Cleavage defects at saturating enzyme concentrations were highly dependent on sequence. Whereas substrates containing one or two A–U base-pair insertions were cleaved at rates within twofold of the WT substrate, one or two G–C base-pair insertions resulted in ~50- and ~1500-fold lower $k_{obs}$ values, respectively (Fig. 4.13C). Partial RNase T1 digestions and RNA hydrolysis ladders revealed that these RNA constructs were cleaved above the inserted base pair(s) and just below the WT C–G base pair (Fig. 4.14B). Thus, Csy4-catalyzed cleavage likely requires prior melting of any additional secondary structures below the five-base-pair stem, such that the WT stem is correctly positioned in the binding pocket and the guanosine containing the 2'-OH nucleophile can productively interact with Arg102 and Phe155. In support of this interpretation, A–U and U–A base pairs are thermodynamically less stable than G–C and C–G base pairs at the termini of RNA duplexes (Xia et al., 1998) and are likely to be more susceptible to transient fraying (Snoussi and Leroy, 2001), explaining the large magnitude of $k_{obs}$ differences for these distinct insertions.

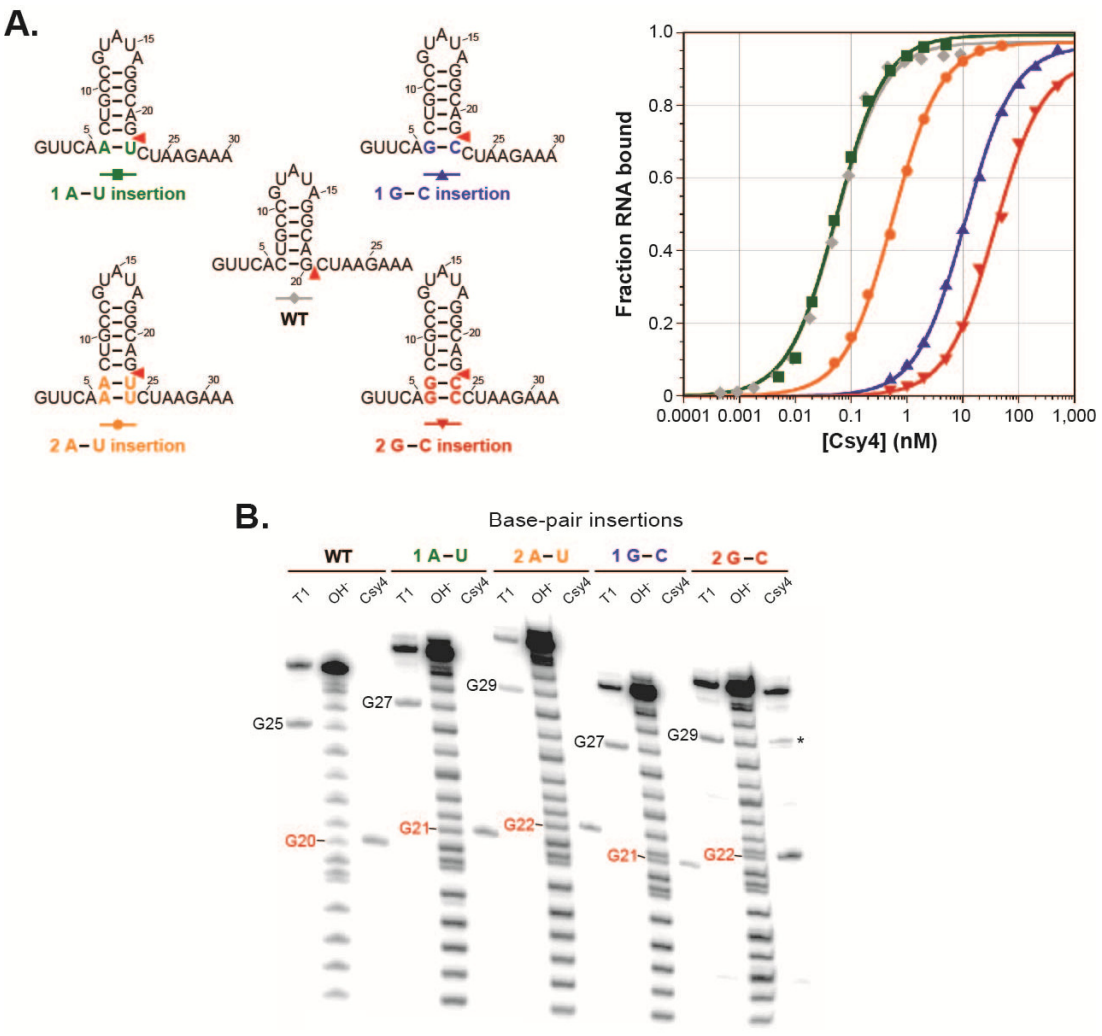



**Figure 4.14 Binding data and cleavage site mapping for base-pair insertion constructs.** (**A**) RNA substrates containing base-pair insertions below the terminal C–G base pair were generated (left), and EMSAs were performed with Csy4(H29A) (right). Binding defects were mildest for one or two A–U base-pair insertions (~1- and ~10-fold) and increased to ~200- and ~800-fold for one or two G–C base pairs, respectively. The cleavage site for each RNA substrate is indicated with a red triangle. (**B**) To experimentally determine cleavage sites, partial RNase T1 digestions and hydrolysis ladders were conducted and resolved by denaturing PAGE adjacent to Csy4 cleavage products. Nucleotides are numbered as in (A), and the guanosine residue directly upstream of the scissile phosphate is shown in red. In all cases, cleavage by Csy4 occurs just below the C–G base pair at the bottom of the WT five base-pair stem, above the base-pair insertion(s). * denotes a minor side product.

Collectively, these data indicate that beyond sequence-specific recognition of its crRNA repeat substrate, Csy4 is finely tuned to bind and cleave stem–loop substrates containing just five base pairs within the dsRNA region, through at least two distinct mechanisms. First, binding energy contributed by the arginine-rich helix requires an

accessible major groove, which depends on the double-stranded stem being properly spaced between interaction sites at its base (e.g., with Arg102) and the loop sequence. Second, rapid cleavage requires the positioning of a terminal C–G base pair within the active site and prior disruption of any additional secondary structures below.

**4.4 Discussion**

The CRISPR-Cas adaptive immune system has evolved a sophisticated strategy for generating large libraries of short effector RNAs that target invasive genetic elements for destruction. Rather than requiring each crRNA to be individually transcribed, the repetitive CRISPR architecture allows large precursor transcripts to be successively processed by Cas endoribonucleases (in type I and III CRISPR systems) that are precisely tailored for specific recognition and cleavage of the invariant repeat sequence. Here we have defined the various molecular strategies employed by one such Cas enzyme—Csy4 (Cas6f) from *P. aeruginosa* UCBPP-PA14—to enable an impressive degree of affinity and specificity for its crRNA repeat substrate.

The Csy4:RNA complex is characterized by an ~50 pM equilibrium dissociation constant ($K_d$) and requires only a 16-nt stem–loop motif for tight binding. For comparison, U1A protein, MS2 coat protein, and the Nλ protein bind their RNA substrates with $K_d$ values of 50 pM, 2.7 nM, and 5 nM, respectively (Cilley and Williamson, 1997; LeCuyer et al., 1995; van Gelder et al., 1993). High-energy interactions are mediated almost exclusively within the major groove of a double-stranded RNA stem–loop, a region of A-form helices that is generally refractory to protein contacts because of its inaccessibility. Prior work used chemical probing to demonstrate that the termini of dsRNA contain uncharacteristically wide major grooves (Weeks and Crothers, 1993), which explains how direct readout of A19 and G20 at their major groove edge is possible. Our data reveal that stable binding of the arginine-rich helix further up the stem is also highly sensitive to major groove accessibility, and that this requirement enables up to ~50,000-fold discrimination against hairpin substrates containing slightly longer stems. Four arginines within this alpha helix are precisely positioned to contact multiple phosphates within the RNA backbone and adopt conformations reminiscent of the arginine fork first described for HIV-1 Tat protein by Frankel and colleagues (Calnan et al., 1991). This mode of multi-dentate interaction requires precise interatomic P–P distances, indicating that the network of hydrogen bonds formed by the arginine-rich helix depends on a very specific substrate conformation. Indeed, changes to the loop sequence or to the identity of base pairs in the upper part of the stem result in substantial binding defects, despite the general lack of base-specific contacts in this region. Substrate selection thus proceeds in large part via an indirect readout mechanism, whereby a particular RNA tertiary structure is recognized that is contingent on both primary sequence and the distinct helical geometry it imposes. Similar modes of substrate recognition have been described for a number of dsDNA-binding proteins (Otwinowski et al., 1988; Rohs et al., 2009).

Csy4 retains the same tight binding for both its substrate and product, and functions as a single-turnover catalyst due to potent product inhibition. These data strongly suggest that crRNA biogenesis in *P. aeruginosa* UCBPP-PA14 requires stoichiometric amounts of the processing endoribonuclease. Cleavage of the crRNA repeat substrate depends critically on the presence of a guanosine upstream of the

scissile phosphate, independently of whether or not this nucleotide is base-paired, and is inhibited when additional secondary structure forms below the five-base-pairstem. The $k_{obs}$ defects we observed with Csy4(R102A) and Csy4(F155A) mutants indicate that the G20 base must be tightly locked in place within the enzyme active site in order to rapidly achieve chemical activation of the ribosyl 2'-OH.

We recently reported that, together with six copies of Csy3 and single copies of both Csy1 and Csy2, Csy4 and the mature crRNA assemble into a large ribonucleoprotein complex (Csy complex) that is responsible for target recognition during the interference stage of the CRISPR pathway (Wiedenheft et al., 2011b). Our data are consistent with a model where the Csy4-bound crRNA serves as a nucleation point for assembling the remainder of the complex, which does not form independently of RNA (Wiedenheft et al., 2011b). Interestingly, Cse3 (Cas6e), the CRISPR-specific endoribonuclease from type I-E CRISPR systems, also acts as a single-turnover enzyme (Sashital et al., 2011) and forms part of the downstream target recognition effector complex (Cascade) (Brouns et al., 2008; Jore et al., 2011; Wiedenheft et al., 2011a). It is tempting to speculate that these related enzymes evolved to react stoichiometrically during pre-crRNA cleavage in order to ensure that the mature crRNA is not prematurely released into the cytoplasm but instead remains tightly sequestered by the Cas machinery. While this mechanistic feature may be intrinsic to certain Cas6 family members, it is not generalizable. Cas6 in type III-B CRISPR systems is not a component of the downstream effector complex (Cmr complex) (Hale et al., 2009), and Cas6 from type I-A CRISPR systems remains only loosely associated with the downstream effector complex (archaeal Cascade) (Lintner et al., 2011). Intriguingly, these differences correlate with the thermodynamic stability of hairpin structures encoded by CRISPR repeats typical of each subtype; repeats clustered based on sequence similarity that associate with type I-E and type I-F CRISPR systems encode highly stable RNA secondary structures, whereas those that associate with type I-A and type III-B systems encode RNAs predicted to be unstructured (Kunin et al., 2007). 2007).

CRISPR-specific endoribonucleases are unusual in that their biological function involves cleavage of a single, invariant substrate. As such, these enzymes have likely coevolved with their target crRNA repeats to retain a high degree of substrate specificity, which serves to avoid spurious binding and/or cleavage of noncognate RNAs inside the cell. The work presented in this chapter highlights the diverse molecular strategies exploited by *P. aeruginosa* Csy4 (Cas6f) to generate this selectivity while maintaining an extremely high-affinity interaction with its ligand. The potential benefits of these attributes for molecular biology applications will be exciting to explore further.

# Chapter 5

# Utilizing Csy4 to engineer modular and predictable gene expression

*A portion of the work presented in this chapter has been submitted for publiciation as part of the following manuscript: Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A., Arkin, A.P. (2012). RNA Processing Enables Predictable Programming of Prokaryotic Gene Expression. *Nat Biotechnol* submitted.

*Lei Qi and Rachel Haurwitz designed the constructs. Lei Qi and Wenjun Shao cloned the constructs. Lei Qi performed the quantitative, *in vivo* assays. Rachel Haurwitz performed the Northern blotting.

## 5.1 Introduction

The field of synthetic biology strives to create standard genetic parts that function modularly, predictably, and reliably in a context-free, plug-and-play manner. Large consortia, such as the Registry of Standard Biological Parts, founded in 2003 at the Massachusetts Institute of Technology, offers a large catalog of genetic parts, such as promoters, ribosome binding sites (RBSs), and transcriptional terminators. However, there remain a number of stumbling blocks to researchers attempting to engineer synthetic constructs. A recent review defined five broad areas in which synthetic biology fails: many of the genetic parts are poorly characterized, constructed pathways are frequently unpredictable, the complexity of larger networks requires a tremendous amount of human hours to test and assemble, many parts are incompatible with each other and/or the host organism, and biological noise or mutation can destroy synthetic circuits (Kwok, 2010).

Genetic systems often behave unpredictably due to structural interactions between DNA, RNA, and protein components as well as functional interactions with host factors and metabolites (Ellis et al., 2009). Due to these complexities, the ability to program gene expression quantitatively based on the characteristics of individual components is very limited. In nature, transcription, translation, and degradation of an RNA transcript are crucial to any gene expression event (Culler et al., 2010), and all three events are controlled by a combination of promoters, RBSs, and *cis*-regulatory signals encoded in untranslated regions (UTRs) (Endy, 2005). These elements can unpredictably interact with each other through formation of RNA structures and recruitment of factors that affect global transcript accessibility and stability. We postulate that physically separating genetic elements at the transcript level will allow modular programming of predictable genetic systems.

## 5.2 Methods

### 5.2.1 Strains and media

The *E. coli* strain Top10 (Invitrogen) was used in all experiments. EZ rich defined media (EZ-RDM, Teknoka) was used as the growth media for *in vivo* fluorescence assays. The antibiotics used were 100 μg $ml^{-1}$ carbenicillin (Fisher) and 34 μg $ml^{-1}$ chloramphenicol (Acros). Culturing, genetic transformation, and verification of transformation were done as previously described (Lucks et al., 2011), using either ampicillin resistance or chloramphenicol resistance as selectable markers.

### 5.2.2 Plasmids construction

The *csy4* gene was cloned from the previously described vector pHMGWA-Csy4 using primers 5'-TTCAAAAGATCTAAAGAGGAGAAAGGATCTATGGACCACTACCTC GACATTCGCTTGCGA-3' and 5'-TCCTTACTCGAGTTATCAGAACCAGGGAACGAAA CCTCC-3', and inserted into a vector containing a tetracycline-inducible promoter $P_L$tetO-1 (Lutz and Bujard, 1997), an ampicillin-selectable marker, and a ColE1 replication origin (pCsy4). A second vector containing a chloramphenicol-selectable marker and a pSC101 replication origin (low copy) was used for cloning reporter constructs following standard cloning techniques (pControl library and pCRISPR library). In the case of complex cis-regulatory circuits, the vector containing a

chloramphenicol-selectable marker and a p15A replication origin was used. Detailed cloning procedures are described below.

### 5.2.3 Time course measurements

The *E. coli* strain expressing Csy4, Top10-Csy4, was derived by transforming *E. coli* Top10 cells with pCsy4. Reporter plasmids were then transformed into the Top10-Csy4 cells and plated on Difco LB+Agar plates containing 100 μg $ml^{-1}$ carbenicillin and 34 μg $ml^{-1}$ chloramphenicol, followed by incubation at 37 °C overnight. For the strains with random 30-nt UTRs, one single colony with a unique 30-nt UTR insertion (sequencing verified) was picked. In all other experiments, three single colonies of each construct were picked. The picked colonies were grown in 300 μL of EZ-RDM containing 100 μg $ml^{-1}$ carbenicillin and 34 μg $ml^{-1}$ chloramphenicol in 2 mL 96-well deep well plates (Costar 3960) overnight at 37 °C and 1000 r.p.m. in a high-speed Multitron shaker (ATR Inc.). One μL of this overnight culture was then added to 149 μL of fresh EZ-RDM with the same antibiotic concentrations with or without 2 μM anhydrotetracycline (Fluka) supplemented. The temporal fluorescence expression from the lag growth phase to the stationary phase was monitored using a high-throughput fluorescence plate reader (Tecan M1000) for 24 hours. The excitation and emission wavelengths used for sfGFP were 485 nm and 510 nm. The excitation and emissions wavelengths used for mRFP were 587 nm and 610 nm. Optical densities were measured at 600 nm ($OD_{600}$); the shaking period between consecutive measurements is 900 seconds with a shaking diameter of 2.5 mm.

### 5.2.4 Flow cytometry and analysis

Flow cytometry measurements were carried out as previously described (Lucks et al., 2011). The 300 μl cell cultures were grown for 8 hrs to an $OD_{600}$ of 0.3 in a Multitron shaker before measurement. Cells were sampled with a medium flow rate until 80,000 cells had been collected. Data were analyzed using FCS Express (De Novo Software) by gating on a polygonal region containing 75% cell population in the forward scatter-side scatter plot.

### 5.2.5 Northern blotting

Single colonies containing the indicated genomic UTR-GFP reporter construct with or without Csy4 co-expression were grown in 5 ml EZ-RDM containing 100 μg $ml^{-1}$ carbenicillin, 34 μg $ml^{-1}$ chloramphenicol, and 2 μM anhydrotetracycline. Samples were pelleted by centrifugation at $OD_{600}$ of 0.6~0.7. Total RNA was extracted using the mirVana miRNA Isolation Kit (Ambion) as per the manufacturer's instructions. 5 μg of each total RNA sample were separated by electrophoresis on a 9% urea polyacrylamide gel. Samples were subsequently transferred to a nylon membrane (Hybond-N+, GE Healthcare) using a semi-dry transfer cell (Bio-Rad). The membrane was pretreated with ULTRAHyb-Oligo Hybridization Buffer (Ambion) and probed overnight with a 5'-[$^{32}$P]-radiolabeled DNA oligonucleotide complementary to the GFP open reading frame (5'-CTTCAGCACGCGTCTTGTAGGTCCCGTCATC-3'). The membrane was washed twice with 2X saline-sodium citrate (SSC) buffer containing 0.5% SDS and visualized by phosphorimaging. The probe was stripped from the membrane by rocking the membrane in 200 ml pre-boiled 0.1% SDS at 66 ºC for 20 min. The membrane was

pretreated with hybridization buffer again and then probed with a 5'-[$^{32}$P]-radiolabeled DNA oligonucleotide complementary to 16S ribosomal RNA (5'-CGTCAATGAG CAAAGGTATTAACTTTACTCCCTTCCTCCCCGC-3'). After washing with 2X SSC as before, the membrane was visualized by phosphorimaging.

### 5.2.6 Construction of random 30-nucleotide UTR libraries

The last six nucleotides in the $\sigma^{70}$ promoter J23119 (http://partsregistry.org/Part:BBa_J23119) were modified to be a BglII restriction enzyme site to enable BioBrick cloning (Shetty et al., 2008). We performed inverse polymerase chain reaction (iPCR) to insert various RBS sequences and the 28-nt cleavage elements (5'-GTTCACTGCCGTATAGGCAGCTAAGAAA-3') upstream of the sfGFP or mRFP genes (Huang, 1994). The RBS sequences used are (with SD sequences in bold):

| **RBS** | **Sequences** |
|---|---|
| Bujard RBS (Lutz and Bujard, 1997) | GAATTCATTAAAG**AGGAGA**AAGGTACC |
| B0030 RBS<br>(http://partsregistry.org/Part:BBa_B0030) | TTTAAGA**AGGAGA**TATACAT |
| Weiss RBS (Weiss et al., 2001) | ATTAAAG**AGGAGA**AATTAAGC |
| Anderson RBS<br>( http://partsregistry.org/Part:BBa_J61100) | TCTAGAGAAAG**AGGGGA**CAAACTAGT |

We then performed iPCR-based saturation mutagenesis to insert random 30-nt sequences either between the promoter and the RBS sequences or between the promoter and the cleavage element. A common reverse primer 5'-NNNNNNNNNN AGATCTATTATACCTAGGACTGAGCTAGCTG-3' with different forward primers 5'-NNNNNNNNNNNNNNNNNNNN-overlap sequences-3' were used together for iPCRs, where N represents a random nucleotide. All primers were PAGE purified. We picked random colonies and used sequencing to verify that each colony contained a circuit with a unique 30-nt UTR insertion. All UTR insertions were computed to not contain explicit SD sequences using the RBS calculator with a low efficiency of translation initiation (Salis et al., 2009).

### 5.2.7 Cloning genomic UTR sequences into reporter plasmids

The 5' UTR sequences of 12 genes from the *E. coli* MG1655 strain were obtained from EcoCyc (http://ecocyc.org). The genes were randomly chosen from a list of genes with known transcriptional start sites which were annotated as "promoter experimentally verified." For all 12 UTRs except for *lldP*, we performed iPCR to insert the sequences into the plasmid containing the Bujard RBS and sfGFP gene with and without the cleavage element. For the *lldP* UTR, we applied the CPEC cloning method

to insert its sequence into the circuits (Quan and Tian, 2011). Details for the 12 UTRs are summarized below:

| Gene | Length (nt) | Strand | Description |
|---|---|---|---|
| *lacZp1* | 18 | -1 | *β*-D-galactosidase; The first gene in *lacZYA* operon; The first one of four promoters |
| *serB* | 19 | +1 | Phosphoserine phosphatase; *SerB-RadA* operon |
| *chiA* | 23 | -1 | Endochitinase |
| *lacY* | 31 | -1 | Lactose MFS transporter; The second gene in *lacZYA* operon |
| *sodA* | 31 | +1 | Superoxide dismutases |
| *ompRp3* | 35 | -1 | OmpR response regulator; *ompR-envZ* operon; The third of multiple promoters |
| *trpR* | *36* | +1 | Tryptophan transcriptional repressor |
| *glpA* | 44 | +1 | The large unit of the glycerol-3-phosphate dehydrogenase; *glpABC* operon |
| *rhoL* | 45 | +1 | *Rho* operon leader peptide |
| *CRISPRI* | 53 | +1 | The leader sequence between the promoter $P_{CRISPRI}$ and the first repeat hairpin |
| *fixA* | 58 | +1 | Hypothetical flavoprotein subunit required for anaerobic carnitine metabolism; *fixABCX* operon |
| *lldP* | 90 | +1 | Lactate transporter; *lldPRD* operon |

### 5.2.8 Construction of twenty-eight combinatory circuits

We constructed these circuits by modifying the promoter sequences using either iPCR (for J23101, J23105, and J23110, http://partsregistry.org/Part:BBa_J23119) or CPEC cloning (for $P_{T7A1}$ (Higuchi et al., 1988), $P_L$lacO-1, and $P_{A1}$lacO-1 (Lutz and Bujard, 1997)). The promoter sequences are summarized below:

| Promoter | Constitutive or Inducible | Sequence (**<u>bold underline</u>** - transcriptional start site; ***bold italic*** – operator sites) |
|---|---|---|
| J23119 | Constitutive | TTGACAGCTAGCTCAGTCCTAGGTATAATAGATC**<u>T</u>** |
| J23101 | Constitutive | TTTACAGCTAGCTCAGTCCTAGGTATTATAGATC**<u>T</u>** |
| J23105 | Constitutive | TTTACGGCTAGCTCAGTCCTAGGTACTATAGATC**<u>T</u>** |
| J23110 | Constitutive | TTTACGGCTAGCTCAGTCCTAGGTACAATAGATC**<u>T</u>** |

| | | |
|---|---|---|
| $P_{T7A1}$ | Constitutive | CGAGGCCAACTTAAAGAGACTTAAAAGATTAATTTAAAATT TATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATA CTTACAGCC**A**TCGAGAGGGA |
| $P_L$lacO-1 | Inducible* | ***ATAAATGTGAGCGGATAACA***TTGACA***TTGTGAGCGGATA ACAA***GATACTGAGCAC**A**TCAGCAGGACGCACTGACC |
| $P_{A1}$lacO-1 | Inducible* | AAAATTTATCAAAAAGAGTGTTGACT***TGTGAGCGGATAACA A***TGATACTTAGATTC***AATTGTGAGCGGATAACAATTTCAC ACA*** |

*Induced with 500 μM IPTG.

### 5.2.9 Construction of synthetic operons

Multiple-step BioBrick (Shetty et al., 2008) cloning was used to clone all operon circuits. First we introduced BioBrick cloning sites (BglII/BamHI with AatII) into the monocistronic plasmids using iPCR. Second we performed double digestions of the monocistronic plasmids using either the combination of AatII and BglII or the combination of AatII and BamHI, and ligated the digestion products following standard protocols. Third we performed double digestions again to assemble multiple cistrons onto a single plasmid. The same procedure was used to clone the bicistronic circuits with orthogonal IS10 *cis*-regulatory elements. The orthogonal IS10wt and IS10-9 sequences and their antisense RNAs are described below.

| **Description** | **Sequence** (**bold** - Shine-Dalgarno sequences; ***bold italic*** – specificity sites) |
|---|---|
| IS10wt UTR | ***GCGAAAAA***TCAATA**AGGAGA**CAACAAG |
| IS10-9 UTR | ***GGCTTAAA***TCAATA**AGGAGA**CAACAAG |
| IS10wt antisense RNA | TCGCACATCTTGTTGTCTGATTATTGA***TTTTTCGC***GAAACCATTTGATCAT ATGACAAGATGTGTATCCACCTTAACTTAATGATTTTTACCAAAATCATTA GGGGATTCATCAG |
| IS10-9 antisense RNA | TCGCACATCTTGTTGTCTGATTATTGA***TTTAAGCC***GAAACCATTTGATCAT ATGACAAGATGTGTATCCACCTTAACTTAATGATTTTTACCAAAATCATTA GGGGATTCATCAG |

### 5.2.10 Construction of synthetic circuits with composite UTR functions

We used BioBrick cloning to assemble the PT181 *cis*-regulatory element with the IS10wt *cis*-regulatory elements onto a single sfGFP reporter plasmid. The antisense plasmids were constructed by inserting the cassette that expressed the PT181 antisense RNA under the promoter J23119 into pCsy4. This cassette was also inserted into the ColE1 vector harboring both the *csy4* gene and the IS10wt antisense RNA expression cassette to obtain a plasmid that expresses both antisense RNAs. The sequences of the PT181wt system are given below.

| Description | Sequence |
|---|---|
| PT181 UTR | AACAAAATAAAAAGGAGTCGCTCACGCCCTGACCAAAGTTTGTGAACGA<br>CATCATTCAAAGAAAAAAACACTGAGTTGTTTTTATAATCTTGTATATTTA<br>GATATTAAACGATATTTAAATATACATAAAGATATATATTTGGGTGAGCGA<br>TTCCTTAAACGAAATTGAGATTAAGGAGTCGCTCTTTTTTATGTATAAAAA<br>CAATCATGCAAATCATTCAAATCATTTGGAAAATCACGATTTAGACAATTT<br>TTCTAAAACCGGCTACTCTAATAGCCGGTTGTAA |
| PT181 antisense RNA | ATACAAGATTATAAAAACAACTCAGTGTTTTTTTCTTTGAATGATGTCGTT<br>CACAAACTTTGGTCAGGGCGTGAGCGACTCCTTTTTATTT |

### 5.2.11 Calculation of protein production rates (PPRs)

The PPRs are calculated following the mathematical equation (Leveau and Lindow, 2001): $\left.\frac{\partial P}{\partial t}\right|_{production,ss} = \left.\frac{\partial F}{\partial OD}\right|_{ss} \cdot \mu \cdot (1+\frac{\mu}{\nu})$, where P is the PPR, F is fluorescence, v is protein maturation rate (per hour), and μ is growth rate (per hour). For each construct, we measured the temporal curves of F and $OD_{600}$. The plot of the measured fluorescence F as a function of $OD_{600}$ showed a linear curve for all circuits tested during the exponential growth phase, which allowed us to calculate the slopes of this linear curve ($\left.\frac{\partial F}{\partial OD}\right|_{ss}$). We calculated the growth rate (μ) during the exponential phase for each construct by plotting Log($OD_{600}$) over time. The maturation constants for sfGFP and mRFP are $m_{sfGFP} = 7h^{-1}$ (Pedelacq et al., 2006) and $m_{mRFP} = 3.5h^{-1}$ (Campbell et al., 2002). The data were then analyzed using Mathematica 7 (Wolfram). The statistics were performed using Prism 5 (GraphPad Software), and the Spearman's rank correlation coefficients were calculated using Microsoft Excel.

### 5.2.12 Measurement of RNA polymerase dropoff rate

We calculated the average dropoff rate of RNA polymerase with following the mathematical equation $Log(RFP) = Log(1-\lambda) \cdot L$, where λ (per nucleotide, 0 < λ < 1) is the average dropoff rate, and L is the length of the upstream sequence. We shortened the length of L to 690, 594, 495, 393, 294, 198, 150, and 0 (bp), and measured the RFP expression from the second cistron. The data were used for linear regression between RFP and L. We calculated that the 95% confidence interval (CI) of the average RNA polymerase dropoff rate was $\lambda = 8.3\times10^{-4} : 1.04\times10^{-3}$ (per nucleotide).

## 5.3 Results

### 5.3.1 The CRISPR RNA processing system eliminates interactions between UTRs and RBSs

In order to physically separate genetic elements at the transcript level, we exploited a synthetic RNA processing platform derived from the pre-crRNA processing system in *P. aeruginosa* UCBPP-PA14. Our synthetic cleavage platform is modularly

comprised of the *csy4* gene and its 28-nucleotide cleavage element (the crRNA repeat sequence). The cleavage element is inserted into reporter transcripts between other components, and the Csy4 protein is induced to cleave at designed loci and generate well-defined RNA segments. Csy4 is highly specific for only its cognate RNA target (Sternberg et al., 2012), and is therefore well-suited for use in a synthetic platform.

We first constructed two reporter libraries in *E. coli* to test whether RNA cleavage could eliminate interactions between UTRs and translational elements such as RBSs. In the control library, random 30-nt UTR sequences were placed between a constitutive promoter and a characterized RBS fused to a green fluorescent protein-coding gene. In the CRISPR library, we inserted the 28-nt CRISPR hairpin between the random UTRs and the RBS. The protein production rates (PPRs) of clones randomly sampled from each library were calculated during exponential cell growth (Fig. 5.1A). The control library demonstrated a wider distribution of PPRs (Fig. 5.1B) than the CRISPR library (Fig. 5.1C). The relative standard deviation (RSD) of the distribution with RNA processing was three-fold less than that without. Moreover, the difference in mean expression between the control library and a baseline construct (the baseline bias, BB) lacking the 30-nt UTR insertion was larger than that of the CRISPR library. We observed that both RSD and BB increased in the absence of Csy4 expression (Fig. 5.1D). Co-expression of Csy4 with the cleavage element led to a reduction in size of the transcript as detected by Northern blotting, consistent with cleavage of the mRNA at the CRISPR hairpin (Fig. 5.2). We conducted a series of similar experiments by inserting random 30-nt UTRs between different RBSs and genes with the Csy4 cleavage element, and demonstrated a consistent reduction of expression variability and BB compared to those control constructs lacking Csy4 cleavage (Fig. 5.3). Since genomic 5' UTRs exist in varying lengths and encode a variety of structures that might not be captured in the above random library, we tested 12 UTR sequences found in the *E. coli* MG1665 genome that range from 18- to 90-nts in length in the reporter system. As expected, a much lower variance of the PPR was observed with RNA cleavage compared to that without (Fig. 5.1E).

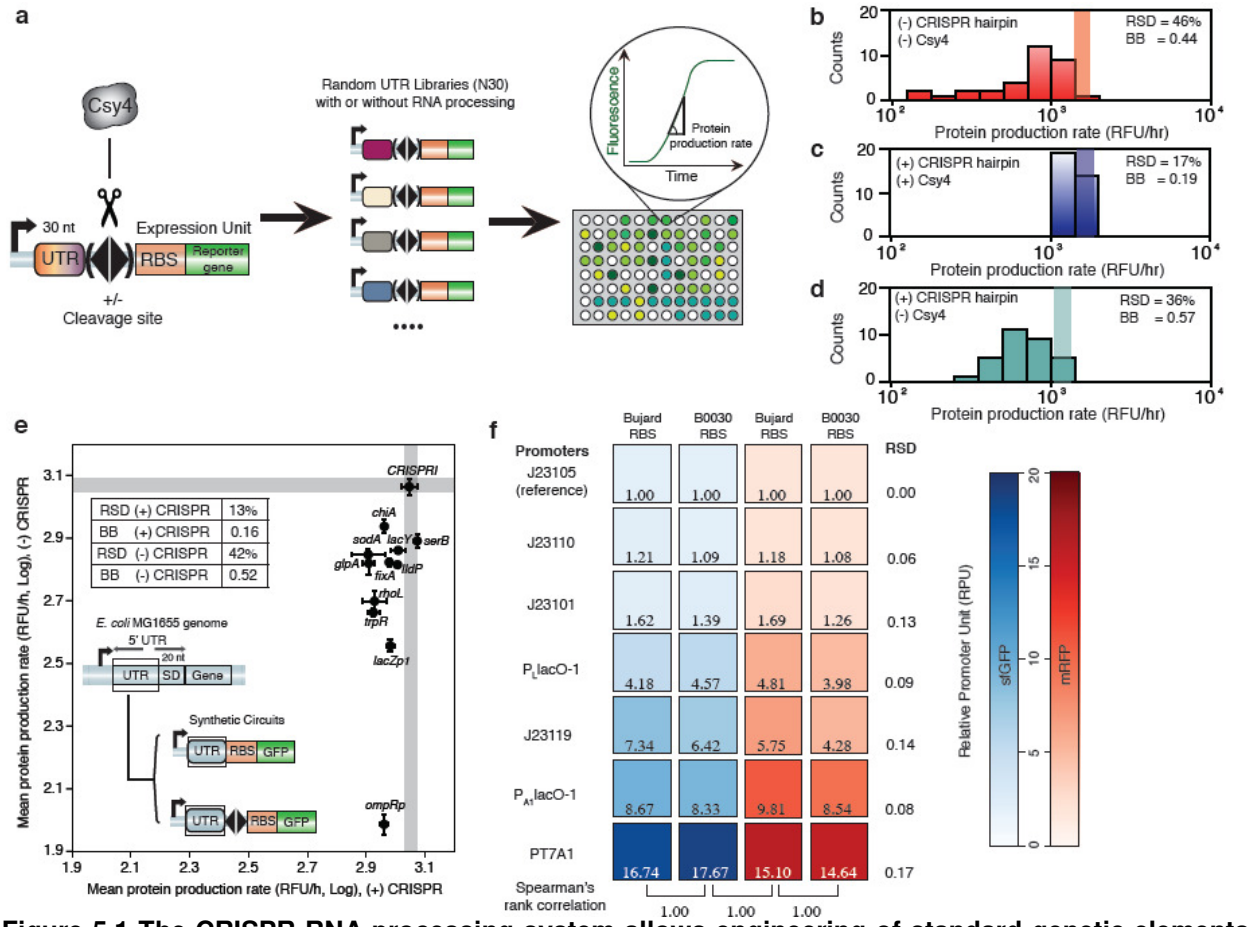



**Figure 5.1 The CRISPR RNA processing system allows engineering of standard genetic elements in various contexts.** (**A**) Experimental procedure for measuring the effects of random UTRs on gene expression with and without the CRISPR cleavage element (diamond). (**B** – **D**) Statistical analysis of protein production rates (PPRs). The expression of baseline constructs without 30-nt UTR insertions are plotted as shaded lines, with the width representing the standard deviation of biological triplicates. (**E**) The 2D plot of the mean PPRs for 12 genomic UTR insertions with and without RNA processing. The gray lines show the PPR values of the baseline constructs. (**F**) Experimental data for twenty-eight combinatory circuits composed of seven promoters, two RBSs, and two reporter genes with the cleavage element inserted between the promoters and RBSs. The heat maps show the relative promoter unit (RPU) values, with each column normalized to a reference promoter J23105. The RSD values for each row and Spearman's rank correlation coefficients between adjacent columns are labeled.

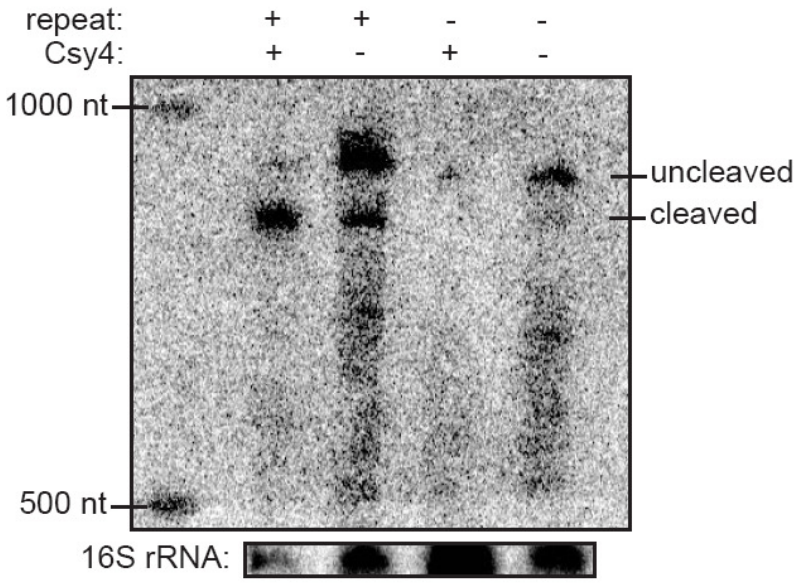
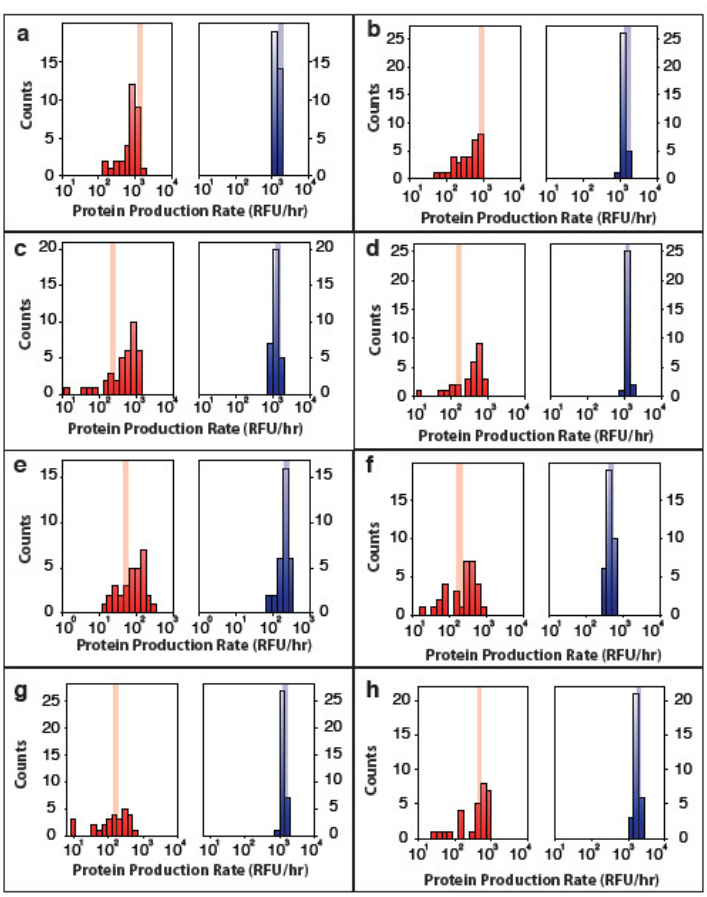



**Figure 5.2 Northern analysis of total RNA from *E. coli* cells to verify *in vivo* Csy4 cleavage.** The cells in all columns contained an expression cassette that transcribed a segment of the genomic *lldP* UTR (90-nt) and the GFP-coding gene (744-nt). The first two columns further contained a 28-nt CRISPR cleavage hairpin inserted between the *lldP* UTR and GFP. When Csy4 is co-expressed, cleavage should reduce the transcript size to 752-nts, which was observed in the first column (more than 70% cleaved). Uncleaved transcript is 864-nts, as shown in the second column (more than 65% remained uncleaved). In the absence of a CRISPR repeat hairpin, no cleavage was observed as shown in columns 3 and 4. The band migrating approximately the same as the cleaved samples seen in column 2 is likely due to an intrinsic transcriptional start site inside the *lldP* UTR sequence, which will be further investigated.




i

| # | RBS | Reporter | (±) CRISPR | BB | RSD (%) |
|---|---|---|---|---|---|
| a | Bujard | sfGFP | (-) | 0.56 | 46% |
| | | | (+) | 0.74 | 19% |
| b | B0030 | sfGFP | (-) | 2.21 | 60% |
| | | | (+) | 0.73 | 13% |
| c | Anderson | sfGFP | (-) | 1.88 | 68% |
| | | | (+) | 0.81 | 27% |
| d | Weiss | sfGFP | (-) | 1.77 | 67% |
| | | | (+) | 0.97 | 22% |
| e | Bujard | mRFP | (-) | 0.55 | 32% |
| | | | (+) | 0.70 | 15% |
| f | B0030 | mRFP | (-) | 2.35 | 61% |
| | | | (+) | 0.88 | 11% |
| g | Anderson | mRFP | (-) | 1.39 | 80% |
| | | | (+) | 0.79 | 13% |
| h | Weiss | mRFP | (-) | 0.94 | 62% |
| | | | (+) | 0.80 | 12% |

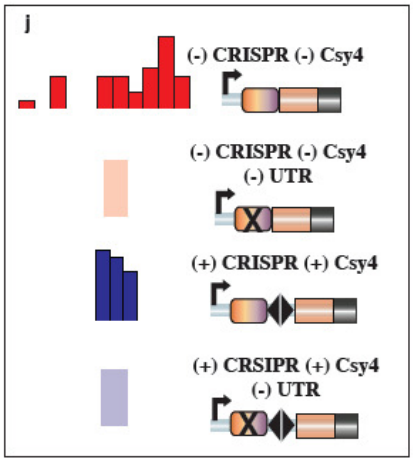


**Figure 5.3 The synthetic RNA processing system improves the predictability of different RBSs and genes.** (**A** – **H**) Eight translation units composed of four RBS sequences and two reporter genes are tested with randomized 30-nt UTR sequences with or without the Csy4 hairpin. The histograms in red show the constructs without RNA processing, and the histograms in blue show those with processing. The mean values of the control construct without a 30-nt UTR sequence are shown as lines, and the width of the lines represent the standard deviation of biological triplicates. (**I**) The statistical summary of (A – H). (**J)** The legend for (A – H).

### 5.3.2 CRISPR processing standardizes promoter strength

Most promoters used in genetic engineering are derived from natural DNA sequences, usually with poorly annotated operator sequences and transcriptional start sites (Mutalik et al., 2012a). As a consequence, direct assembly of promoters with other elements such as UTRs and RBSs often changes translational efficiency and transcript stability and alters the apparent activity of the promoter in different contexts. To determine if RNA processing could remove this interference, we compared two expression libraries with and without RNA processing. Both libraries consisted of twenty-eight parallel constructs generated via combinatorial assembly of seven promoters with two RBSs and two reporter genes (see Methods). We inserted the cleavage element between the promoters and RBSs in the CRISPR library and assayed the relative promoter unit (RPU) of each promoter by normalizing its production rate to that of a reference promoter. The measured RPU values in the control library varied widely across different RBSs and genes (Fig. 5.4). In contrast, the RPU of each promoter in the CRISPR library was nearly constant and the rank orders were completely conserved across different contexts (Fig. 5.1F), implying that cleavage of the mRNA between promoters and downstream elements allowed precise characterization of the standard RPUs of promoters.

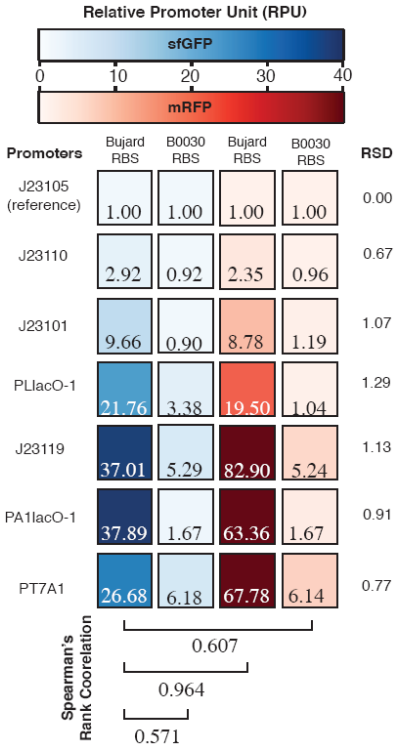



**Figure 5.4 Measured relative promoter units (RPUs) of the promoters without RNA processing.** The heat map shows the RPU value of each promoter by normalizing to a reference promoter J23105. The numbers after each row show the RSD values of RPUs (variance) across the contexts. Spearman's rank correlation coefficients between columns are labeled on the bottom.

### 5.3.3 RNA processing enables design of operonic systems

We applied RNA processing to the rational design of multi-cistronic operons. Operons are ubiquitous in prokaryotes and present engineering advantages due to their compactness and coordinated regulation (Pfleger et al., 2006). However, multiple cistrons encoded on the same transcript may interfere with one another through RNA structures and translational coupling which affect translation initiation and transcript stability. Here we compared expression of a set of genes in monocistronic format to different arrangements in bicistronic operons, when these cistrons were or were not bounded by the cleavage elements. Six monocistrons were first constructed by combining three RBSs with two fluorescent proteins, and the bicistrons were constructed by pairing every RFP monocistron with all GFP monocistrons such that both serve as the first and second gene in the operon. This provided two parallel sets of eighteen bicistrons in permutation: one without RNA processing and one with RNA processing (Fig. 5.5A). In the control library, expression of the first cistron was linearly correlated to the monocistronic format, but expression of the second cistron was highly variable (Fig. 5.5B). The results from the CRISPR library were strikingly different. If a gene appeared as the first cistron in a bicistronic operon, its production was almost the same as the corresponding monocistron; if a gene appeared as the second cistron, there was a strong linear correlation between its expression and the corresponding monocistron (Fig. 5.5C and Fig. 5.6). The consistent difference in the second cistron expression reflects transcriptional polarity (Wek et al., 1987), an effect that results in lower expression levels of operonic genes distal to the promoter compared to proximal genes. Notably, our synthetic cleavage system allows precise measurement of the RNA polymerase dropoff rate during elongation, which was estimated as $8\times10^{-4}$ per nucleotide in our constructs (Fig. 5.7).

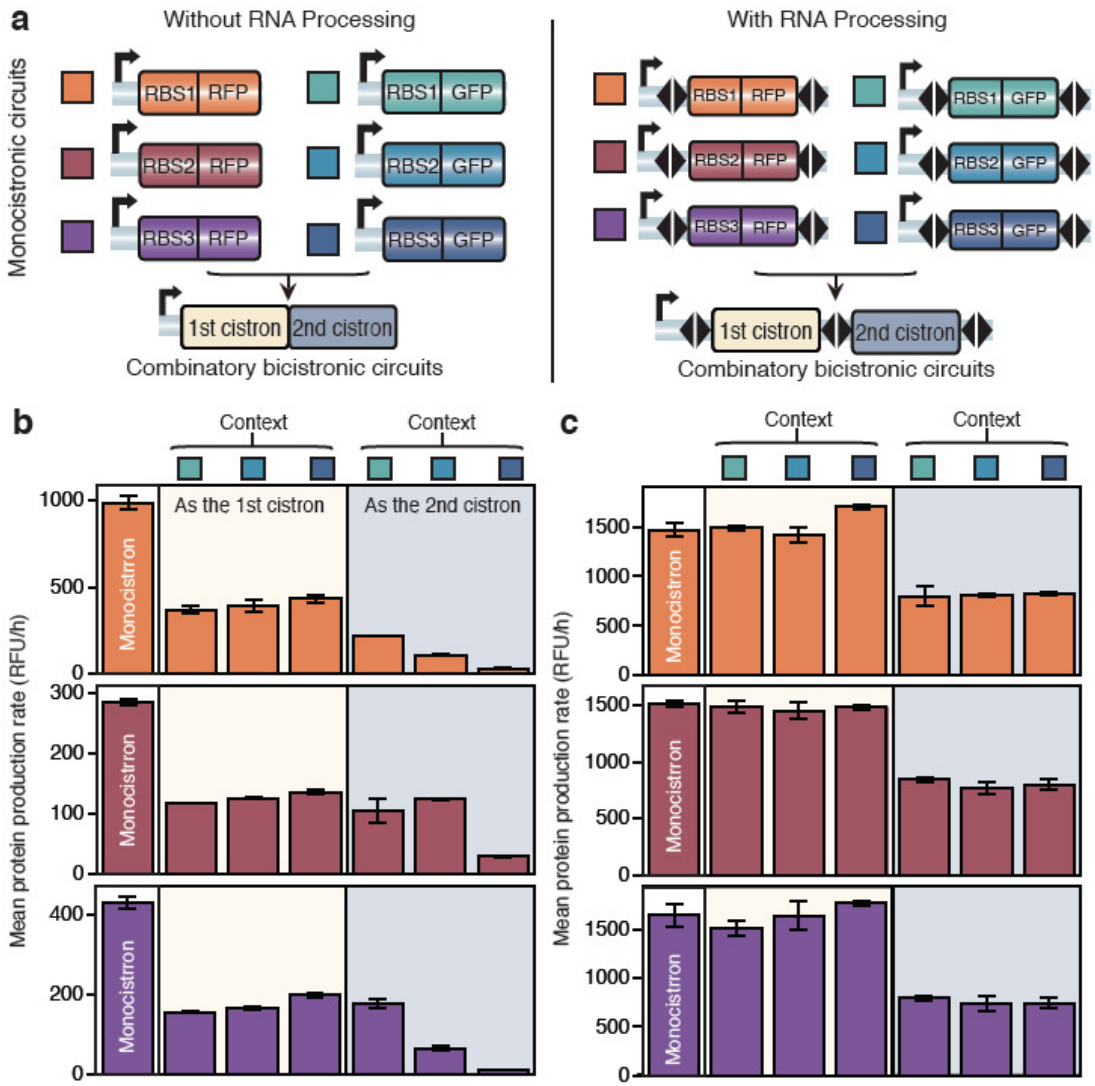



**Figure 5.5 RNA processing enables design of operonic systems.** (**A**) Six monocistrons are combined in pairs to generate eighteen bicistrons without (left) and with (right) cleavage elements (see Methods). (**B**) Measured PPRs for the monocistrons and bicistrons without RNA processing. The bicistrons with RFP in the first and second cistrons are shaded in yellow and gray. The context is labeled on the top. (**C**) PPRs with RNA processing. Coloring as in (B).

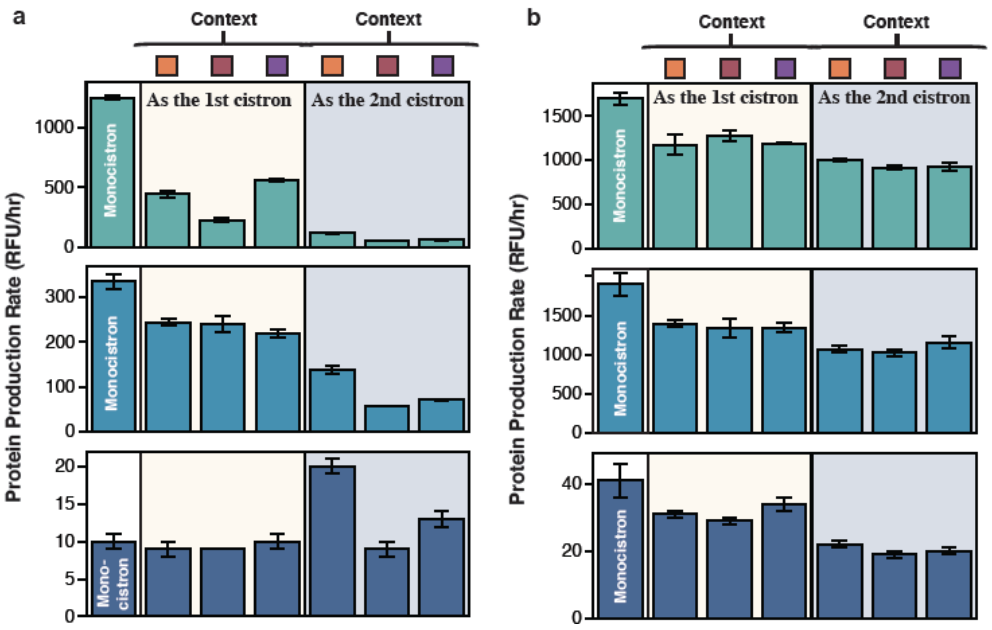



**Figure 5.6 Measurement of GFP expression as the first or second cistron in the operon with and without RNA processing.** (**A**) The bar graphs show the calculated mean GFP PPRs from biological triplicates for each monocistronic or bicistronic circuit without RNA processing. Their contexts are shown at the top. (**B**) The measured GFP PPRs for the constructs with RNA processing.

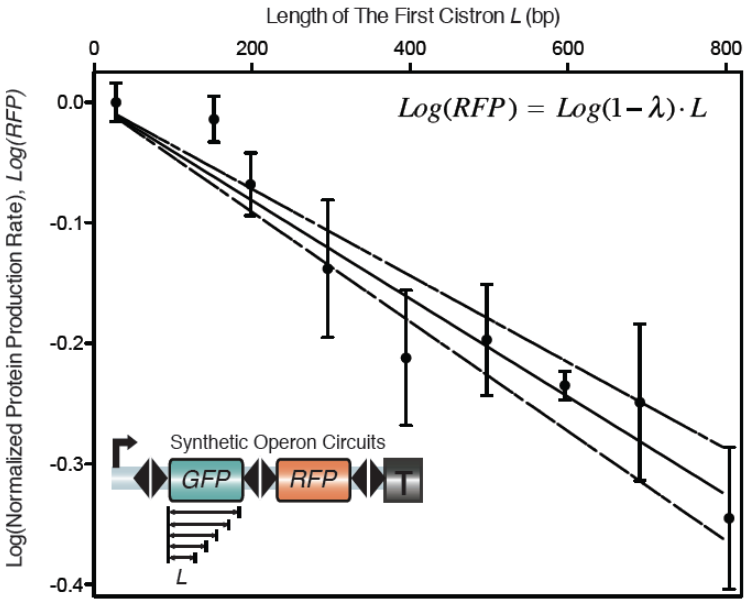



**Figure 5.7 Measurement of transcriptional polarity using synthetic operons.** We systematically shortened the length of the first cistron from the 3' end (bottom left inset), and measured the expression of RFP in the second cistron. The plot of the normalized RFP expression data and the length of the first cistron shows a linear correlation, whose slope is used to calculated the average dropoff rate of RNA polymerase. The dotted lines show the 95% confidence interval of the fitness.

### 5.3.4 RNA processing permits the predictable engineering of complex *cis* regulatory systems

We next applied the RNA processing system to design complex *cis*-regulatory systems. Two families of antisense RNA-mediated *cis* elements were used for multi-input regulation: two orthogonal pairs of translational repressors (IS10wt and IS10-9) (Mutalik et al., 2012b) and one transcriptional attenuator (PT181wt) (Lucks et al., 2011). Attempts to use orthogonal IS10 elements to differentially control translation of individual genes inside an operon failed (Fig. 5.8A, left panel). This lack of functionality is likely due to coupled transcript stability and structural interactions between the two cistrons. However, when the precursor transcript was cleaved at designed loci to free the 5' and 3' ends, each antisense RNA individually repressed the cognate gene without affecting the other (Fig. 5.8A, right panel and Fig. 5.9).

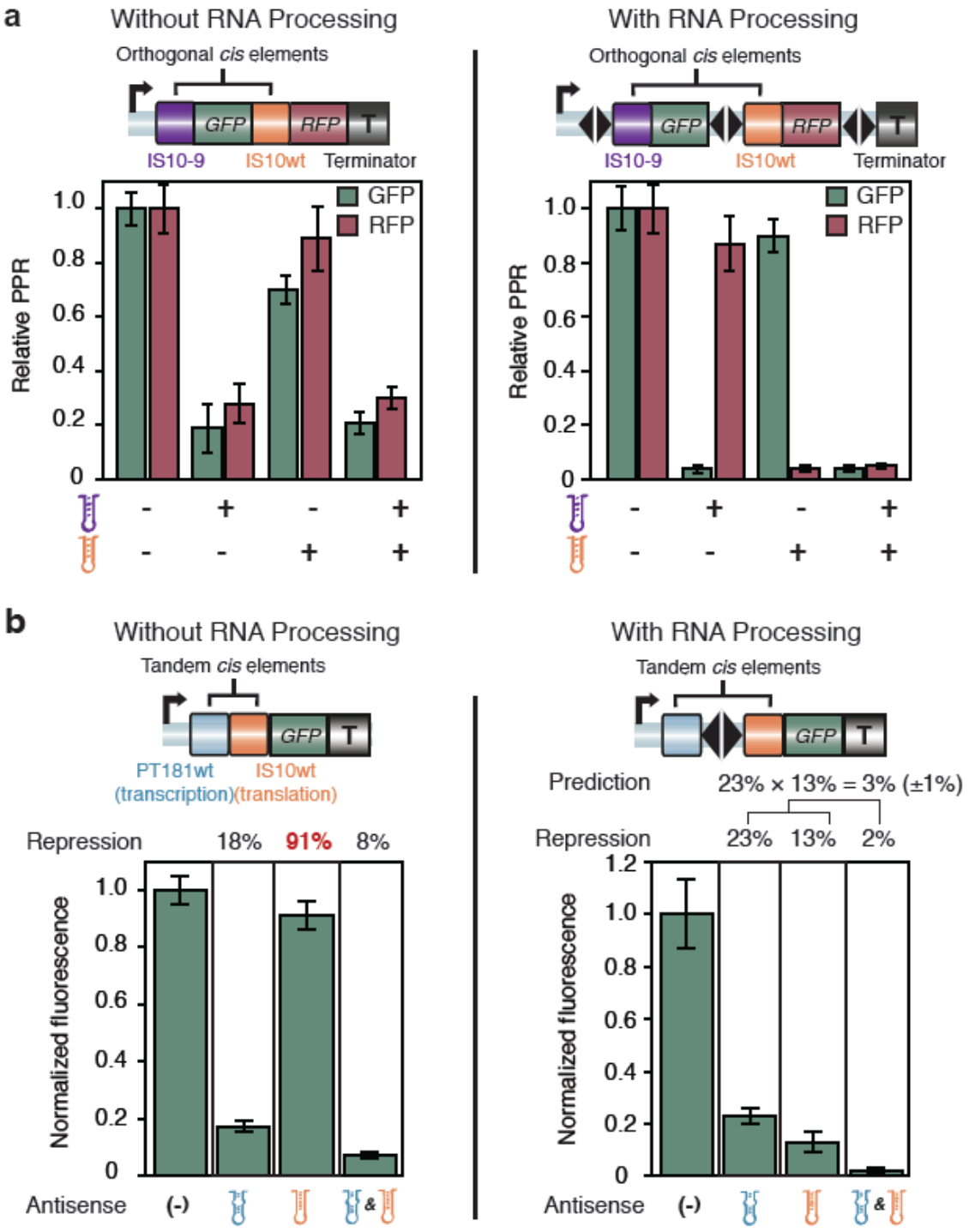



**Figure 5.8 Application of RNA processing to the predictable engineering of complex *cis* regulatory systems.** (**A**) Orthogonally-acting IS10 antisense RNA-mediated *cis* elements are used to control translation of individual genes in an operon without (left) or with (right) RNA processing. (**B**) Design of a complex *cis*-regulatory system without (left) or with (right) RNA processing. All data are normalized to gene expression without any antisense RNA.

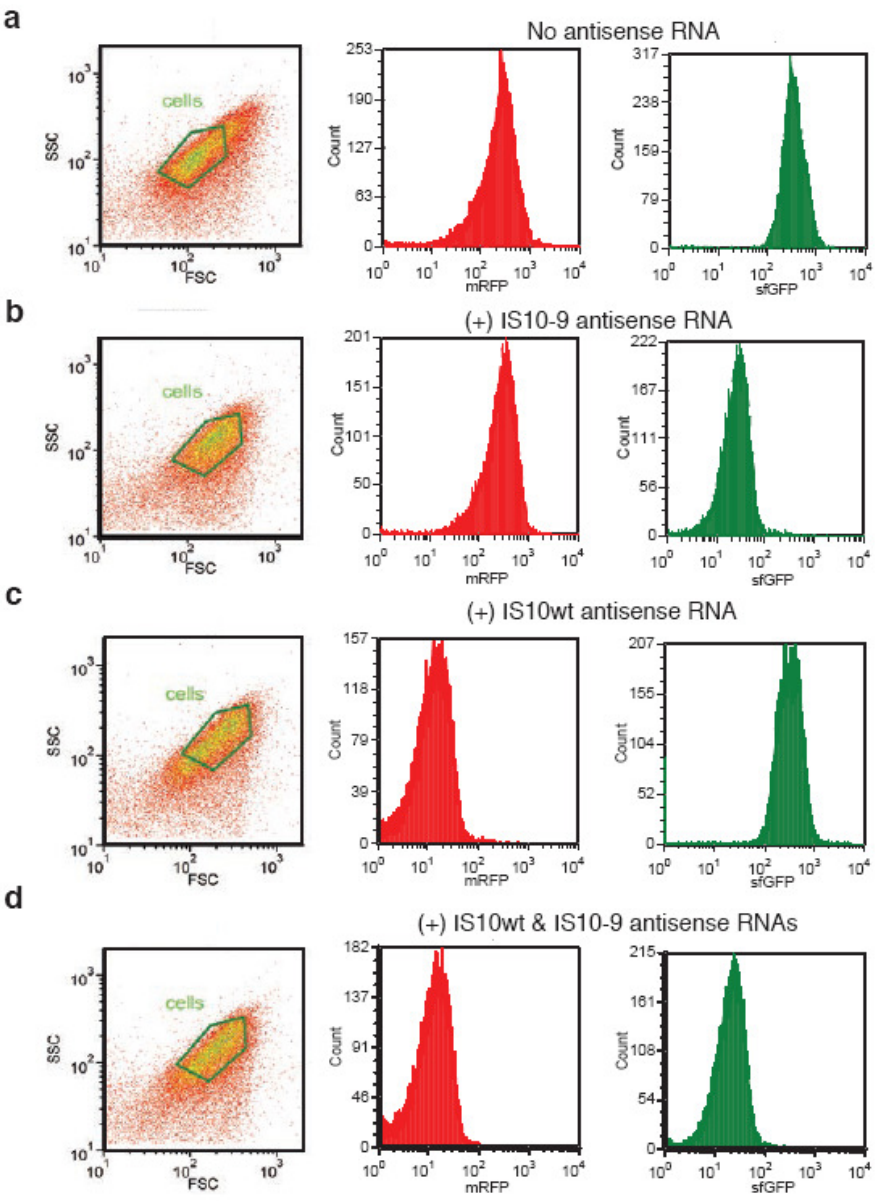



**Figure 5.9 Flow cytometry analysis of the synthetic operon controlled by orthogonal IS10 cis-regulatory systems.** (**A**) No antisense; (**B**) Only IS10-9 antisense; (**C**) Only IS10wt antisense; (**D**) Both antisense RNAs. The first column shows the forward scatter-side scatter 2D plots with the polygon gating 75% of cell populations. The second and third columns show histograms of RFP and GFP expression.

We designed an additional version of complex RNA-level regulations by combining two *cis*-regulatory systems, PT181wt and IS10wt, in tandem to control a monocistron. Ideally, such tandem multi-input control would result in a multiplicative function of the two elements (Lucks et al., 2011). Without RNA cleavage, this function was not obtained (Fig. 5.8B, left panel). In contrast, cleavage between the two tandem elements allowed ideal multiplicative regulation (Fig. 5.8B, right panel), indicating that complex *cis* regulations could only be achieved using RNA processing.

We then compared the efficacy of CRISPR-based cleavage to other RNA cleavage elements by inserting different cleavage elements into the complex *cis*-regulatory system (Fig. 5.10). None of these elements behaved as effectively as Csy4-

based system, suggesting that the CRISPR system may be more robust to different genetic contexts, which could be a unique feature of CRISPR, since Csy4 naturally cleaves its cognate target in a wide variety of unique genetic contexts derived from phage and plasmid genomes (Wiedenheft et al., 2012).

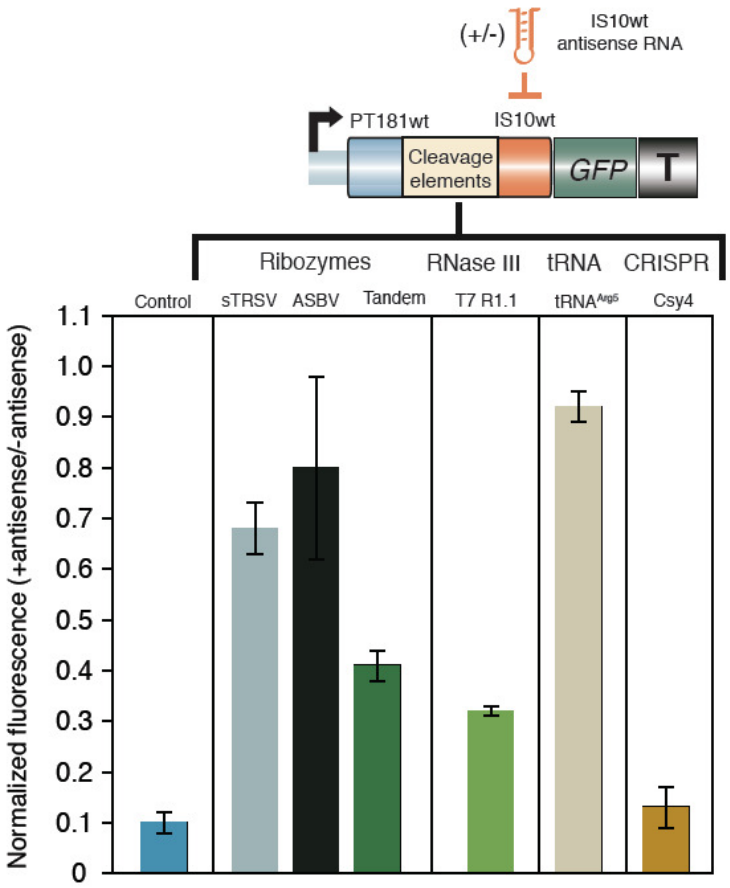



**Figure 5.10 Comparison of the efficacy of different RNA cleavage elements using the complex *cis*-regulatory system.** Different RNA cleavage elements were inserted between the tandem *cis* regulators (PT181wt and IS10wt). Fluorescence of each construct in the presence of IS10wt antisense RNA was measured using flow cytometry, averaged over biological triplicates, and normalized to a construct without IS10wt antisense expression. The control shows the repressive activity of IS10wt antisense RNA on an IS10wt UTR that is not fused in tandem to the PT181wt UTR. sTRSV is short for small Tobacco Ring Spot Virus hammerhead ribozyme (Hampel and Tritz, 1989). ASBV is short for Avocado Sun Blotch Virus ribozyme (Daros et al., 1994). T7 R1.1 is the RNase III recognition site derived from the T7 phage (Dunn and Studier, 1973). $tRNA^{Arg5}$ served as a highly structured element that is not cleaved and was obtained from (Espeli et al., 2001). Among all constructs, Csy4 allowed maximal repression and exhibited the highest efficacy of restoring riboregulator-UTR function.

## 5.4 Discussion

Our results demonstrate that RNA processing enforces high levels of modularity between physically linked and functionally coupled elements within a precursor transcript, leading to predictable gene expression programming. CRISPR-based controllable RNA processing allows creation of standard genetic parts such as promoters and RBSs that behave consistently across diverse genetic contexts. More broadly, together with technologies for genome-scale modification (Wang et al., 2009), our work establishes a foundation for the efficient and predictable engineering of genetic systems to meet demands for new therapies, manufacturing, and the environment.


**March 2012 *RNA* cover featuring one of the six Csy4-crRNA co-crystal structures presented in this thesis.**

# Bibliography


Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213-221.

Al-Attar, S., Westra, E.R., van der Oost, J., and Brouns, S.J. (2011). Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol Chem *392*, 277-289.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Anand, K., Schulte, A., Vogel-Bachmayr, K., Scheffzek, K., and Geyer, M. (2008). Structural insights into the cyclin T1-Tat-TAR RNA transcription activation complex from EIAV. Nat Struct Mol Biol *15*, 1287-1292.

Andersson, A.F., and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. Science *320*, 1047-1050.

Auweter, S.D., Oberstrass, F.C., and Allain, F.H. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res *34*, 4943-4959.

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

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

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

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

Brussow, H., and Hendrix, R.W. (2002). Phage genomics: small is beautiful. Cell *108*, 13-16.

Busso, D., Delagoutte-Busso, B., and Moras, D. (2005). Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. Anal Biochem *343*, 313-321.

Cady, K.C., and O'Toole, G.A. (2011). Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. J Bacteriol *193*, 3433-3445.

Cai, Z., Gorin, A., Frederick, R., Ye, X., Hu, W., Majumdar, A., Kettani, A., and Patel, D.J. (1998). Solution structure of P22 transcriptional antitermination N peptide-boxB RNA complex. Nat Struct Biol *5*, 203-212.

Calnan, B., Tidor, B., Biancalana, S., Hudson, D., and Frankel, A. (1991). Arginine-mediated RNA recognition: the arginine fork. Science *252*, 1167-1171.

Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A., and Tsien, R.Y. (2002). A monomeric red fluorescent protein. Proc Natl Acad Sci U S A *99*, 7877-7882.

Cantor, C.R., and Schimmel, P.R. (1980). The Conformation of Biological Macromolecules (San Francisco, W.H. Freeman and Co.).

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

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

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *66*, 12-21.

Chen, V.B., Davis, I.W., and Richardson, D.C. (2009). KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. Protein Sci *18*, 2403-2409.

Cilley, C.D., and Williamson, J.R. (1997). Analysis of bacteriophage N protein and peptide binding to boxB RNA using polyacrylamide gel coelectrophoresis (PACE). RNA *3*, 57-67.

Cochrane, J.C., and Strobel, S.A. (2008). Catalytic strategies of self-cleaving ribozymes. Acc Chem Res *41*, 1027-1035.

Collaborative Computational Project, N. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr *50*, 760-763.

Culler, S.J., Hoff, K.G., and Smolke, C.D. (2010). Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. Science *330*, 1251-1255.

Daros, J.A., Marcos, J.F., Hernandez, C., and Flores, R. (1994). Replication of avocado sunblotch viroid: evidence for a symmetric pathway with two rolling circles and hammerhead ribozyme processing. Proc Natl Acad Sci U S A *91*, 12813-12817.

DeLano, W.L. (2002). http://www.pymol.org.

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

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

Draper, D.E. (1995). Protein-RNA recognition. Annu Rev Biochem *64*, 593-620.

Dunn, J.J., and Studier, F.W. (1973). T7 early RNAs and Escherichia coli ribosomal RNAs are cut from large precursor RNAs in vivo by ribonuclease 3. Proc Natl Acad Sci U S A *70*, 3296-3300.

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

Ellis, T., Wang, X., and Collins, J.J. (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat Biotechnol *27*, 465-471.

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

Endy, D. (2005). Foundations for engineering biology. Nature *438*, 449-453.

Espeli, O., Moulin, L., and Boccard, F. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. J Mol Biol *314*, 375-386.

Fersht, A.R. (1987). The hydrogen bond in molecular recognition. Trends Biochem Sci *12*, 301-304.

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

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., and Macmillan, A.M. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat Struct Mol Biol *18*, 688-692.

Gherghe, C.M., Mortimer, S.A., Krahn, J.M., Thompson, N.L., and Weeks, K.M. (2008). Slow conformational dynamics at C2'-endo nucleotides in RNA. J Am Chem Soc *130*, 8884-8885.

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

Grosse-Kunstleve, R.W., and Adams, P.D. (2003). Substructure search procedures for macromolecular structures. Acta Crystallogr D Biol Crystallogr *59*, 1966-1973.

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

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

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M.*, 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*, 292-302.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell *139*, 945-956.

Hampel, A., and Tritz, R. (1989). RNA catalytic properties of the minimum (-)sTRSV sequence. Biochemistry *28*, 4929-4933.

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

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

Haurwitz, R.E., Sternberg, S.H., and Doudna, J.A. (2012). Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. EMBO J.

Higuchi, R., Krummel, B., and Saiki, R.K. (1988). A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Res *16*, 7351-7367.

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

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

Huang, S.H. (1994). Inverse polymerase chain reaction. An efficient approach to cloning cDNA ends. Mol Biotechnol *2*, 15-22.

Huppler, A., Nikstad, L.J., Allmann, A.M., Brow, D.A., and Butcher, S.E. (2002). Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. Nat Struct Biol *9*, 431-435.

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

Johnson, J.E., Jr., and Hoogstraten, C.G. (2008). Extensive backbone dynamics in the GCAA RNA tetraloop analyzed using 13C NMR spin relaxation and specific isotope labeling. J Am Chem Soc *130*, 16757-16769.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R.*, et al.* (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol *18*, 529-536.

Kabsch, W. (2010). XDS. Acta Crystallogr D Biol Crystallogr *66*, 125-132.

Katoh, H., Yoshinaga, M., Yanagita, T., Ohgi, K., Irie, M., Beintema, J.J., and Meinsma, D. (1986). Kinetic studies on turtle pancreatic ribonuclease: a comparative study of the base specificities of the B2 and P0 sites of bovine pancreatic ribonuclease A and turtle pancreatic ribonuclease. Biochim Biophys Acta *873*, 367-371.

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

Kwok, R. (2010). Five hard truths for synthetic biology. Nature *463*, 288-290.
Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat Rev Microbiol *8*, 317-327.

Ladner, J.E., Wladkowski, B.D., Svensson, L.A., Sjolin, L., and Gilliland, G.L. (1997). X-ray structure of a ribonuclease A-uridine vanadate complex at 1.3 A resolution. Acta Crystallogr D Biol Crystallogr *53*, 290-301.

LeCuyer, K.A., Behlen, L.S., and Uhlenbeck, O.C. (1995). Mutants of the bacteriophage MS2 coat protein that alter its cooperative binding to RNA. Biochemistry *34*, 10600-10606.

Lee, D.G., Urbach, J.M., Wu, G., Liberati, N.T., Feinbaum, R.L., Miyata, S., Diggins, L.T., He, J., Saucier, M., Deziel, E.*, et al.* (2006). Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial. Genome Biol *7*, R90.

Legault, P., Li, J., Mogridge, J., Kay, L.E., and Greenblatt, J. (1998). NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. Cell *93*, 289-299.

Leveau, J.H., and Lindow, S.E. (2001). Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. J Bacteriol *183*, 6752-6762.

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

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

Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copie, V.*, et al.* (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem *286*, 21643-21656.

Lucks, J.B., Qi, L., Mutalik, V.K., Wang, D., and Arkin, A.P. (2011). Versatile RNA-sensing transcriptional regulators for engineering genetic networks. Proc Natl Acad Sci U S A *108*, 8617-8622.

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res *25*, 1203-1210.

Maag, D., and Lorsch, J.R. (2003). Communication between eukaryotic translation initiation factors 1 and 1A on the yeast small ribosomal subunit. J Mol Biol *330*, 917-924.

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

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F.*, et al.* (2011). Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol *9*, 467-477.

Manica, A., Zebec, Z., Teichmann, D., and Schleper, C. (2011). In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. Mol Microbiol *80*, 481-491.

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

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

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658-674.

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

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

Mortimer, S.A., and Weeks, K.M. (2009). C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. Proc Natl Acad Sci U S A *106*, 15622-15627.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Mai, Q., Christoffersen, M.J., Martin, L., Yu, A., Lam, C., Rodriguez, C., Bennett, G.*, et al.* (2012a). Composition and quality of irregular transcription and translation genetic elements. Submitted.

Mutalik, V.K., Qi, L., Guimaraes, J.C., Lucks, J.B., and Arkin, A.P. (2012b). Rationally designed families of orthogonal RNA regulators of translation. Nat Chem Biol.

Nolan, S., Shiels, J., Tuite, J., Cecere, K., and Baranger, A. (1999). Recognition of an essential adenine at a protein-RNA interface: Comparison of the contributions of hydrogen bonds and a stacking interaction. J Am Chem Soc *121*, 8951-8952.

Ogawa, T., Inoue, S., Yajima, S., Hidaka, M., and Masaki, H. (2006). Sequence-specific recognition of colicin E5, a tRNA-targeting ribonuclease. Nucleic Acids Res *34*, 6065-6073.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321-329.

Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol *24*, 79-88.

Pfleger, B.F., Pitera, D.J., Smolke, C.D., and Keasling, J.D. (2006). Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. Nat Biotechnol *24*, 1027-1032.

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

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

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

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

Quan, J., and Tian, J. (2011). Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. Nat Protoc *6*, 242-251.

Raines, R.T. (1998). Ribonuclease A. Chem Rev *98*, 1045-1066.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248-1253.

Rousseau, C., Gonnet, M., Le Romancer, M., and Nicolas, J. (2009). CRISPI: a CRISPR interactive database. Bioinformatics *25*, 3317-3318.

Rupert, P.B., and Ferre-D'Amare, A.R. (2001). Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature *410*, 780-786.

Rupert, P.B., Massey, A.P., Sigurdsson, S.T., and Ferre-D'Amare, A.R. (2002). Transition state stabilization by a catalytic RNA. Science *298*, 1421-1424.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol *27*, 946-950.

Sashital, D.G., Jinek, M., and Doudna, J.A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat Struct Mol Biol *18*, 680-687.

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Molecular Cell.

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

Shetty, R.P., Endy, D., and Knight, T.F., Jr. (2008). Engineering BioBrick vectors from BioBrick parts. J Biol Eng *2*, 5.

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

Snoussi, K., and Leroy, J.L. (2001). Imino proton exchange and base-pair kinetics in RNA duplexes. Biochemistry *40*, 8898-8904.

Snyder, J.C., Bateson, M.M., Lavin, M., and Young, M.J. (2010). Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol *76*, 7251-7258.

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

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

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci U S A *99*, 7536-7541.

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

Terns, M.P., and Terns, R.M. (2011). CRISPR-based adaptive immune systems. Curr Opin Microbiol *14*, 321-327.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, P.H., Hung, L.W., Read, R.J., and Adams, P.D. (2008). Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr *64*, 61-69.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

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

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci *34*, 401-407.

van Gelder, C.W., Gunderson, S.I., Jansen, E.J., Boelens, W.C., Polycarpou-Schwarz, M., Mattaj, I.W., and van Venrooij, W.J. (1993). A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation. EMBO J *12*, 5191-5200.

Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2007). Automated structure solution with autoSHARP. Methods Mol Biol *364*, 215-230.

Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*, 894-898.

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

Weeks, K.M., and Crothers, D.M. (1993). Major groove accessibility of RNA. Science *261*, 1574-1577.

Weiss, R., Knight, T.F., Jr., and Sussman, G. (2001). Genetic Process Engineering. In Cellular Computing, M. Amos, ed. (Oxford, Oxford University Press).

Wek, R.C., Sameshima, J.H., and Hatfield, G.W. (1987). Rho-dependent transcriptional polarity in the ilvGMEDA operon of wild-type Escherichia coli K12. J Biol Chem *262*, 15256-15261.

Westra, E.R., Pul, U., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L.*, et al.* (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol *77*, 1380-1393.

Westra, E.R., van Erp, P.B., Kunne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K.*, et al.* (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Molecular Cell.

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

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

Wiedenheft, B., van Duijn, E., Bultema, J.B., Waghmare, S.P., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J., Boekema, E.J., Dickman, M.J.*, et al.* (2011b). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A *108*, 10092-10097.

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

Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry *37*, 14719-14735.

Yang, W. (2011). Nucleases: diversity of structure, function and mechanism. Q Rev Biophys *44*, 1-93.

Ye, X., Gorin, A., Ellington, A.D., and Patel, D.J. (1996). Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. Nat Struct Biol *3*, 1026-1033.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res.

Zamel, R., Poon, A., Jaikaran, D., Andersen, A., Olive, J., De Abreu, D., and Collins, R.A. (2004). Exceptionally fast self-cleavage by a Neurospora Varkud satellite ribozyme. Proc Natl Acad Sci U S A *101*, 1467-1472.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H., and O'Toole, G.A. (2009). Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. J Bacteriol *191*, 210-219.

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