IDENTIFICATION AND CHARACTERIZATION OF RNA-PROTEIN COMPLEXES INVOLVED IN A NOVEL RNA-BASED PROKARYOTIC IMMUNE SYSTEM

by

CARYN RATCLIFF HALE

(Under the Direction of Michael and Rebecca Terns)

ABSTRACT

Recent evidence supports the presence of a small RNA-based genome defense system in prokaryotic organisms. This system is comprised of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and a large family of CRISPR-associated proteins. The presence of invader-derived sequences within the CRISPR loci has been shown to confer immunity against that invader to the CRISPR-containing organism. This system is predicted to depend on small RNAs that are transcribed from the CRISPR loci. This work focuses on characterization of these prokaryotic silencing (psi)RNAs and their associated proteins in the hyperthermophilic archaeon *Pyrococcus furiosus*. Characterization of RNA products from the CRISPR loci revealed unidirectional transcription of all of the CRISPRs in *Pyrococcus furiosus* originating from the conserved leader sequences. Small RNAs derived from all parts of a given CRISPR locus were found, with a greater number found of the more recently incorporated leader-proximal invader targeting sequences. Furthermore, CRISPR transcripts were shown to be processed into small RNAs that contain mostly invader-derived sequence, with an 8-nucleotide portion of the 5' repeat sequence called the psi-tag. Mature psiRNAs were

found to exist in several chromatographically distinct RNA-protein complexes, one of which contained the Cas module RAMP (Cmr) proteins. The psiRNA-Cmr protein complex was found to recognize and cleave single-stranded RNAs that are complementary to the incorporated psiRNAs. Cleavage was shown to occur at a site that is 14 nucleotides from the 3' end of the psiRNA. This works presents the first evidence of small RNA-guided RNA destruction by the CRISPR/Cas system, and may represent a powerful system for sequence-specific degradation of RNAs *in vivo*.

INDEX WORDS: CRISPR, Cas, *Pyrococcus furiosus*, RNA protein complexes, psiRNA, RAMP

IDENTIFICATION AND CHARACTERIZATION OF RNA-PROTEIN COMPLEXES INVOLVED IN A NOVEL RNA-BASED PROKARYOTIC IMMUNE SYSTEM

by

CARYN RATCLIFF HALE

B.A., Louisiana Scholars' College at Northwestern State University, 2004

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

DOCTOR OF PHILOSOPHY

ATHENS, GA

2010

© 2010

Caryn Ratcliff Hale

All Rights Reserved

IDENTIFICATION AND CHARACTERIZATION OF RNA-PROTEIN COMPLEXES INVOLVED IN A NOVEL RNA-BASED PROKARYOTIC IMMUNE SYSTEM

by

CARYN RATCLIFF HALE

Major Professors:	Michael Terns Rebecca Terns
Committee:	Robert Scott Lance Wells

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

DEDICATION

To all of the strong women in my life who have taught me what it means to be independent, intelligent and fiercely loyal to my family.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentors, Drs. Michael and Rebecca Terns. You took a chance on both me and this project, and it's been a wild ride, but I sincerely believe that I've had the best graduate school experience possible. You guys have taught me more than you know not only about how to be a scientist and love the work, but how to sacrifice all for family when necessary. I couldn't have asked for better teachers, or better examples, both professionally and personally.

I would also like to thank my committee members, Dr. Lance Wells and Dr. Robert Scott, as well as Dr. Claiborne Glover, for help and support throughout this project.

I would also like to thank the past and present members of the Terns' Lab. I would especially like to thank the rest of the "catalytic triad:" Jason Carte and Sonali Majumdar. What fun it has been obsessing over CRISPR with you both. To the rest of the CRISPR group, past and present: Joshua Elmore, Agniva Saha, Neil Pfister, Lindsay Jones, Megan Tibbetts, Duane Jurma, and Ross Christopher: thanks for your help and support throughout this project. And finally, to two undergraduates, Kyle Kleppe and Rachel Applebaum, two of the most talented undergraduates I have ever encountered. It was an honor working with both of you; I could not have completed this work without your help. I also thank some past members of the Terns' Lab: Osama Youssef, Dan Baker, and Kevin Polach. You guys were all instrumental in teaching me how to be a great, clean scientist, and I'm thankful for the effort you put into my training.

I would also like to thank my family. To Dad, who taught me all that I'll ever need to know about ambition and the importance of education, and to my Mom, my biggest fan, the kindest, most thoughtful person I've ever met. You guys have been behind me every step of the way, and I would never have made it without your support. To the rest of my family, both in blood and in marriage, I thank you all for making me the person that I am.

Last and most importantly, I thank my husband, Conor. No one, including myself, believes in me like you do. You constantly refuse to let me settle, and I owe my success to you and your stubbornness. Our life so far has been an incredible ride, and I can't wait to see where it goes from here. And finally, to my children: I don't expect you ever to understand this monster of a dissertation, but I pray that one day you find a passion that allows you to understand the sacrifices I've made for my passion for science.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

RNA-protein complexes

Since the discovery of catalytic RNAs, many biologists have proposed the theory of an ancient "RNA world" that predated the evolution of both DNA and proteins. In this proposed world, both the genomes and cellular machinery of organisms consisted entirely of RNA components (Talini et al., 2009). True or untrue, this hypothesis is a testament to the biological power of RNAs. RNAs are not only able to confer specificity by base-pairing with other nucleic acids, but they have been shown to play catalytic roles in several critical cellular processes, including the ribosome, vital cellular machinery found in all organisms (Eddy, 2001). Although life has obviously evolved well past this ancient world, RNAs still play a major role in many cellular processes. In fact, in many cases collaboration between the specificity that RNAs can provide and the chemically powerful nature of proteins allows for a molecular machine with power that far outweighs the sum of its parts (Brosius, 2003). The ribonucleoproteins (RNPs) involved in these sorts of processes contain non-coding RNAs, or those that are not translated by the ribosome into proteins. RNPs are involved in a variety of cellular functions such as protein translation, insertion of membrane proteins, and regulation of gene expression (Eddy, 2001; Huttenhofer et al., 2005).

One example of a non-coding RNP contains RNAs called small nucleolar (sno)RNAs in eukaryotes and small RNAs in archaea (which lack a nucleolus). These

RNAs act as sequence-specific guides to lead the associated proteins to modification sites in ribosomal RNAs (Matera et al., 2007). Although some modification can occur in the absence of the snoRNAs (see Appendix 1), the addition of a variety of snoRNAs allow for a single set of proteins to guide numerous different modifications on a given rRNA molecule. In bacteria, where these complexes are not present, an entire family of proteins is necessary for these modification events (Ferre-D'Amare, 2003). The ability to incorporate a variety of specificity-conferring small RNAs into a single complex allows for enormous evolutionary benefit.

RNA-protein complexes involved in genome defense

Another recently discovered function for cellular RNPs in all three kingdoms of life is in genome defense. All organisms must participate in the war between its genome and nucleic acid invaders, such as viruses, phages and transposons. Viruses are one of the most widespread parasites on earth, and are thought to far outnumber the organisms that they seek to infect (Bergh et al., 1989). These invaders must reproduce by entering a host and hijacking its cellular machinery to replicate, transcribe and translate its genome, sometimes by incorporating into the host cell's genome itself, potentially causing deleterious effects to the host (Weinbauer, 2004). A system of defense must therefore exist to protect an organism from unchecked insertions by these foreign genetic elements. Indeed, complex defense systems have been detected in both prokaryotic and eukaryotic organisms that work to silence these mobile genetic elements. However, viruses have shown an incredible aptitude for acquired mutation and recombination events (Deveau et al., 2008; Heidelberg et al., 2009). Therefore, many viruses have evolved means to evade host immune systems, forcing the creation

of new modes of defense. This process has led to the biological arms race between viruses and the organisms they seek to infect, out of which many elegant genome defense systems have evolved.

RNAi and piRNAs in eukaryotes

In eukaryotes, the RNA interference (RNAi) system is implicated in genome defense in two ways. Many of the viral invaders of eukaryotic cells utilize double-stranded (ds)RNA within some portion of their life cells (van Rij and Berezikov, 2009). In eukaryotic cells, dsRNA is recognized by a protein called Dicer and cleaved into small dsRNAs between 21-28 nucleotides long (Figure 1.1). One strand of this dsRNA is incorporated into an RNA-Induced Silencing Complex (RISC) that is made up of a family of proteins called Argonaute (AGO) proteins. Once loaded with the small RNA referred to as a small interfering (si)RNA, the complex recognizes and silences complementary elements (Figure 1.1) (Marques and Carthew, 2007; Martinez et al., 2002). This silencing has been shown to occur through both cleavage of an mRNA, and through direct binding with the gene (or nascent RNA) on the chromosome and subsequent hetereochromatin formation (Figure 1.1) (Buhler and Moazed, 2007).

Another class of small RNAs has recently been found to be involved in protecting the genome of germ cells from transposons. These small RNAs are found associated with a group of Ago proteins called Piwi proteins, and are therefore called piwi-interacting (pi)RNAs (Aravin et al., 2007; Hartig et al., 2007; Houwing et al., 2007; Kim, 2006). These piRNAs have been shown to originate from genomic clusters that contain many transposable elements. These small RNAs are typically 24-30 nucleotides long, and their biogenesis is believed to be Dicer-independent. Rather, they are transcribed

as a long RNA spanning an entire cluster, and processed into small RNAs by an unknown mechanism. Mature piRNAs associate with the Piwi proteins (Piwi, Aubergine (Aub) and Ago3), and form a silencing complex that can cleave transposon mRNAs (Aravin et al., 2007; Brennecke et al., 2007; Gunawardane et al., 2007; Lau et al., 2009). The piRNA pathway also possesses a curious amplification pathway referred to as the ping-pong mechanism (See Figure 1.2) (Brennecke et al., 2007; Gunawardane et al., 2007; Saito et al., 2006). Sequencing efforts found a small fraction of piRNAs that were identical, and not complementary to transposable elements (Aravin et al., 2007). These RNAs were found exclusively associated with the Ago3 Piwi protein . Additionally, while all other piRNAs had a strong bias for a 5' uridine, no such preference was seen in these RNAs. Instead, an adenosine residue was invariably found in the 10^{th} position (Lau et al., 2009). This led to the proposal of a mechanism where active transposons are recognized and cleaved by Piwi- and Aub- containing complexes. Once cleavage occurs, Ago3 associates with the cleaved transposon RNA. The 3' end is processed, and the resulting complex can recognize and cleave precursor piRNA in a manner that will release piRNAs that target active transposons (Figure 1.2). The result of the ping-pong pathway is an amplification of the number of cellular piRNAs that target the transposable elements that are currently active within the cell, and thus an amplification of transposon silencing (Figure 1.2) (Brennecke et al., 2007; Gunawardane et al., 2007; Saito et al., 2006).

Genome defense in prokaryotic organisms

Prokaryotic organisms have not been shown to contain a pathway that directly parallels the eukaryotic RNAi system. Very few homologues of the proteins involved in

the RNAi pathway have been found in archaeal or bacterial genomes. An Argonaute-like protein has been identified in many prokaryotic organisms, but the physiological function of these proteins remains to be elucidated (Makarova et al., 2009). In fact, the crystal structure of an Argonaute-like protein in *Pyrococcus furiosus* led to the hypothesis that Argonaute is responsible for the "Slicer" activity that cleaves target RNAs in the eukaryotic RNAi pathway (Song et al., 2004). However, no functional data is available for this protein, and no evidence of dsRNA-mediated RNA degradation has been seen in prokaryotes. However, bacteria and archaea survive in environments that are rich in phages and other mobile elements, and yet seem to be able to control insertion into their small genomes in a manner more efficiently than that of eukaryotes, when comparing the relative amount of transposable elements found in the genomes. Indeed, several mechanisms of genome defense have been discovered in prokaryotes.

A very well studied system of genome defense in prokaryotes is the restriction-modification (R/M) system. The hallmark of this system is two closely related proteins: a restriction endonuclease, and a methyltransferase. Both of these enzymes recognize a short DNA sequence motif. The endonuclease specifically cleaves this sequence. The methyltransferase protects host DNA by methylating the same sequence within host DNA, which is recognized by DNA that is methylated on only one strand. This cooperative activity means that the only non-methylated DNA within the cell is of foreign origin, providing an important means of specifically targeting foreign DNA elements (Tock and Dryden, 2005).

Another means of protection against phages in bacteria is the process of phage abortive infection. In organisms containing these abortive infection (Abi) mechanisms, a

phage is able to normally infect and enter a host cell, but the process of replication and propagation is aborted, and the cell undergoes programmed cell death, preventing the phage to propagate. This process, while detrimental to the cell, is overall beneficial to the bacterial community. One example of an Abi system is the AbiD1 system in *Lactococcus lactis*. In this system, a viral-encoded protein induces the expression of the AbiD1 protein, which inhibits a protein that is required for resolving branch structures that occur in DNA replication. Activation of this protein inhibits viral replication, and is also toxic to the host cell. Many similar systems have been discovered in *Escherichia coli* and other prokaryotic organisms (Chopin et al., 2005).

CRISPRs: Clustered Regularly Interspaced Short Palindromic Repeats

In the past few years, a novel system for RNA-mediated genome defense in prokaryotes has been discovered. This system, called the CRISPR/Cas system is both inheritable and remarkably adaptable. This system is comprised of two main features: Clustered, Regularly Interspaced Short Pallindromic Repeats (CRISPRs) and a large family of CRISPR-associated (Cas) proteins. CRISPRs were first identified in 1987 in the genome of *Escherichia coli* (Ishino et al., 1987). They have since been found in nearly all sequenced archaeal and ~40% of sequenced bacterial genomes (Bolotin et al., 2005; Godde and Bickerton, 2006; Mojica et al., 2005). They are characterized by short, direct repeats that range from 24 to 48 nucleotides, which are separated by similarly-sized unique sequences (Figure 1.3) (Kunin et al., 2007). These unique spacer sequences are unique to both the genome and among even closely related organisms. CRISPR repeat sequences are often weakly palindromic and are highly variable between organisms. A single CRISPR locus can have from two to several

hundred repeats. Within a locus, the repeat sequences are typically invariant. A given CRISPR-containing organism can have from one to 18 CRISPR loci, and the direct repeat sequences are typically well conserved, but may be highly variable (Figure 1.3) (Bolotin et al., 2005; Godde and Bickerton, 2006; Grissa et al., 2007).

Most CRISPR loci have a conserved sequence directly upstream of the first repeat. This AT-rich sequence, which can be up to 550 base pairs, is termed the leader and has been predicted to be both a promoter for transcription of the loci, and the site of addition of new spacers. The leaders are not generally conserved among organisms, but are often conserved within a single CRISPR-containing genome (Lillestol et al., 2006).

CRISPR-associated (Cas) proteins

A large family of proteins is invariably found associated with the CRISPR loci. These genes, dubbed, CRISPR-associated (*cas*) genes are never found in organisms that do not contain CRISPRs, and are often directly linked to one end of the repeat clusters. Both of these observations point to not only physical, but also functional coupling throughout evolution. The *cas* genes were first identified in 2002 by Makarova et al. as a predicted thermophile-specific DNA repair system (Makarova et al., 2002). Later that year, the genes were shown to be invariably linked to the CRISPR loci, and the term CRISPR-associated, or *cas* genes was coined (Jansen et al., 2002). Four highly conserved *cas* genes were identified and named *cas1 – cas4*. All CRISPR-containing organisms contain *cas1* along with one or more of the other three core cas genes. Its invariable presence in CRISPR-containing organisms has led to *cas1* being considered the hallmark of the CRISPR/Cas system.

Although extensive sequence analysis was performed, no predictions could be originally made for the Cas1 or Cas2 proteins. Cas3 proteins showed similarity to the superfamily 2 DNA helicases. The Cas3 genes were often found either fused or adjacent to a predicted HD nuclease. The Cas4 proteins were predicted to be a RecB family exonuclease due to the presence of conserved active site residues (Jansen et al., 2002; Makarova et al., 2002). Recent biochemical studies have shown that cas1 from *Pseudomonas aeruginosa* cleaves DNA in a metal dependent manner into fragments ~80 nts in length. The crystal structure of the protein revealed a novel fold and a conserved divalent metal binding site (Wiedenheft et al., 2009). Another biochemical study of Cas1 from *Sulfolobus solfataricus* did not observed such cleavage activity. Instead, this Cas1 protein was observed to bind both RNA and DNA with high affinity, and was also shown to promote annealing of complementary nucleic acids (Han et al., 2009).

Biochemical characterization has also been performed on Cas2 and Cas3 in *Solfolobus solfataricus*. Cas2 was found to be a single-stranded RNA specific endonuclease. This protein cleaved in U-rich regions, and required both Mg^{+2} and K^{+} ions for optimal activity (Beloglazova et al., 2008). Cas3 from *Solfolobus solfataricus* was found to possess double-stranded nucleic acid specific endonuclease activity. This protein cleaved either dsRNA or dsDNA, with no clear preference, but showed very little activity with single-stranded nucleic acids. This activity was also seen to be Mg^{+2}-dependent (Han and Krauss, 2009).

The biochemical studies of the core Cas proteins from various organisms have provided glimpses into potential functions of these proteins in the CRISPR/Cas system.

However, it is currently unknown what the true role of any of these proteins is in the mechanism of genome defense.

Another, more extensive bioinformatic study of the cas genes, classed CRISPR-associated proteins into 45 families of proteins (Haft et al., 2005). This work identified two additional core *cas* genes, *cas5* and *cas6* as highly conserved families of *cas* genes. In addition to the core *cas* genes, the rest of the genes were grouped into 8 subtypes and one module of proteins. The subtypes are defined as groups of genes that are often found together in a CRISPR-containing organism. They are named according to an archetypal organism that contains only core genes and the subtype-specific genes. For example, *Escherichia coli* (strain K12) has only the core *cas* genes and the subtype-specific genes that are so called the Ecoli subtype. Other subtypes are Ypest, Nmeni, Dvulg, Tneap, Hmari, Apern, and Mtube. Members of these subtypes are named accordingly. For example, the five proteins in the Mtube subtype are named Csm1 (for Cas subtype Mtube) through Csm5, etc. (Haft et al., 2005). This nomenclature will be used for the remainder of this work. Table 1.1 describes the various Cas proteins and their predicted or observed functions.

In addition to the Cas protein subtypes, one module of proteins was identified that is not present in a particular set of organisms, but is found both archaea and bacteria that contain a variety of the subtypes of Cas proteins. Most of the families of proteins within this module are proteins that were identified previously (Jansen et al., 2002), and were originally termed Repair Associated Mysterious Proteins (RAMPs). Once their association with CRISPRs was detected, the name was changed to Repeat Associated Mysterious proteins (Haft et al., 2005). The RAMP module consists of four

families of RAMP genes (annotated as cas module RAMP, or *cmr1*, *3*, *4* and *6*). Cmr proteins are predicted to contain a ferredoxin-like fold, which may be responsible for RNA-binding. In addition, Cmr proteins contain a conserved C-terminal G-rich loop, which is often considered the "signature" of the RAMP proteins. Two other families of genes are also found associated with the RAMPs (*cmr2* and *5*). Cmr2 is predicted to be a novel polymerase/nuclease and Cmr5 has no functional prediction. A crystal structure was determined for Cmr5. The protein was crystallized as a homotrimer which contains large basic patches that may be involved in RNA-binding (Sakamoto et al., 2009). This RAMP module of proteins is always found as a cluster of genes, but is not necessarily found closely associated with a repeat locus, although it often is (Haft et al., 2005; Jansen et al., 2002). The proteins from this module are of great importance to this work, and greater detail about the structure and function of the proteins will be discussed in greater detail in later chapters.

Until 2005, extensive bioinformatics had been performed on the components of the CRISPR/Cas system, but very little was known about the true biological function of the CRISPRs and their associated proteins. One study found that when a plasmid containing CRISPR repeats was transformed into *Haloferax volcanii*, the resulting cells had reduced cell viability, and the daughter cells were found to have uneven distribution of parental extrachromosomal plasmids (Mojica et al., 1995). All of the daughter cells contained the necessary genome, but the distribution of plasmids within the cells was variable. This led to the prediction that the CRISPRs were involved in partitioning of parental DNA in prokaryotes (Mojica et al., 1995). However, no other data was available to provide a more firm hypothesis.

The function of the CRISPR/Cas system

New insight into the potential function of CRISPRs came with the identification of the remarkable origin of the spacer sequences. Three independent works published in 2005 determined that although the vast majority of spacers had no determinable origin, those that did were derived from extrachromosomal elements, mainly from phages, plasmids and transposons (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). It was further noted that among the spacers derived from viral or plasmid ORFs, few homologs of these ORFs were found in the CRISPR-containing genome. In the few cases where homologous ORFs were found in the chromosome, the spacer-derived sequence was either absent or highly degenerate (Mojica et al., 2005). In *Yersenia pestis*, it was observed that new spacers were integrated in a polarized manner, with the most recent addition being closest to the leader sequence (Pourcel et al., 2005). Finally, in *Streptococcus thermophilus*, the number of phage-derived spacers was shown to be correlated with previously-established degrees of phage resistance (Bolotin et al., 2005). These findings led to the theory that CRISPRs were somehow involved in phage defense in prokaryotic organisms.

Unrelated RNomics studies identified the presence of small RNAs derived from the CRISPR loci in two archaeal organisms: *Sulfolobus solfotaricus* (Tang et al., 2005) and *Archaeoglobus fulgidus* (Tang et al., 2002). In both of these studies, RNAs were cloned and sequenced that contained portions of both repeat and invader-targeting sequences. Subsequent Northern analysis in both organisms revealed a ladder of RNA products that corresponded to multiple repeat-spacer units, with the smallest product being a size consistent with one repeat and one invader targeting sequence. All of the

cloned RNAs were of the same orientation relative to the repeat sequence, which predicted that the leader sequence (described above) is indeed a transcriptional start site for the CRISPR loci (Tang et al., 2002; Tang et al., 2005). This was the first evidence that the CRISPR loci are transcribed into non-coding RNAs that may play an important role in the biological function of the CRISPR/Cas system.

Extensive bioinformatic analysis of the repeat sequences led to the hypothesis that the dyad symmetry found in many of the repeats allowed for potential secondary structure in these CRISPR-encoded RNAs. The repeat sequences were clustered according to sequence. Of the 12 significant clusters that were found, six were predicted to form stable secondary structure. It was also seen that the predicted clusters of repeat sequences corresponded with the subtype-specific proteins that were predicted by Haft et al. (Kunin et al., 2007). This provided support for the hypothesis that classes of cas genes accompany particular repeat sequences.

A hypothesis emerges

Increased understand of the predicted function of the Cas proteins, combined with the knowledge that the CRISPR loci are transcribed into small RNAs led to a groundbreaking new hypothesis: the CRISPR/Cas system is an RNA-based genome defense system that is functionally analogous to the eukaryotic RNAi system (Makarova et al., 2006). This work even proposed functions for the Cas proteins in the predicted system. Cas3, a predicted helicase fused to a nuclease, was proposed to be a dicer-like enzyme that cleaves CRISPR RNAs into monomers of repeat and invader-targeting units. Cas4 (of the RecB exonuclease family) was tentatively predicted to be the prokaryotic or p-Slicer, and the Cmr proteins were predicted to be RNA-binding proteins

that bind the RNAs of various sizes and facilitate the binding to target RNAs and recruitment of p-Slicer (Makarova et al., 2006). Although none of these hypotheses could be tested at the time, the exciting nature of it spurred the beginning of a massive amount of research that has led to the delineation of much of this exciting system.

Experimental evidence for the role of CRISPR in genome defense

In 2007 came the first conclusive evidence that CRISPRs are involved in viral defense in prokaryotic organisms (Barrangou et al., 2007). In this study, the CRISPR loci of *Streptococcus thermophilus* were sequenced before and after phage infection. Indeed, phage-resistant strains were found to have taken up one or more sequences from the invading phage. It was further shown that deletion of these spacers led to phage sensitivity, and insertion of artificial viral sequences in sensitive strains led to phage resistance. One cas gene, *csn1*, was shown to be necessary for CRISPR-mediated phage resistance. This pioneering work conclusively showed a major role of the CRISPR/Cas system in phage defense (Barrangou et al., 2007).

Sequence analysis of natural microbial communities provided another line of evidence for the role of the CRISPR/Cas defense system in phage response (Tyson and Banfield, 2008). In this work, sequencing was performed on bacteria from two biofilms found in an iron mine in California. Analysis of the CRISPR loci of the Leptospirillum species found in these biofilms showed that the CRISPR loci were highly variable, even to the point where it can be assumed that no two cells had the same CRISPR spacer sequences. This finding implies that in nature, changes in the content of the CRISPR loci are happening extremely rapidly, and likely in response to a community-wide phage infection (Tyson and Banfield, 2008).

Another line of evidence in support of the CRISPR/Cas system playing a role in genome defense is the response of phages to interference by the CRISPR/Cas system. Viruses and other mobile elements have long been known to be extremely active in recombination and mutations, due to their constant need to avoid host defense systems (Weinbauer, 2004). It has been shown in phages that infect several species that viruses mutate both the sequence that is homologous to the spacer sequence, referred to as the proto-spacer, and the protoco-spacer adjacent motif (PAM) sequences, a short sequence adjacent to the proto-spacer that has been shown to be a recognition site for the Cas protein machinery (discussed below) (Mojica et al., 2009). Mutations in either of these sequences allows the organisms to evade CRISPR-mediated silencing (Andersson and Banfield, 2008; Barrangou et al., 2007; Heidelberg et al., 2009; Semenova et al., 2009).

From this point, the amount of attention paid to the study of the mechanism of CRISPR-mediated genome defense increased exponentially, and the amount of available publications naturally increased as well. From all of these works, a tentative model has emerged, although many aspects of the system are still not understood. Therefore, the following section will outline the predicted mechanism for the CRISPR/Cas system, describe evidence in support or opposition to this model, and identify areas that are unknown or still being disputed.

Cas protein diversity

Before moving too far into the mechanics of the CRISPR/Cas system, it should be noted that the CRISPR/Cas systems of various organisms are highly variable and divergent. That is, other than Cas1, there are no proteins that are found in all CRISPR-

containing organisms. Therefore, it must be assumed that Cas proteins in general represent a biological variations on a theme. That is, many organisms have evolved many different ways to accomplish the same task. This makes it unreasonable to assume that any Cas protein functions that can be assigned in one organism is true for all organisms.

The mechanism of CRISPR-mediated genome defense

The biological function of the CRISPR/Cas system has been broken into three stages (Figure 1.4). The first stage of function is referred to as adaptation. In the adaptation phase, phage RNA or DNA is recognized by the Cas protein machinery and DNA is inserted into the CRISPR locus. In the expression phase, the CRISPR locus is transcribed into RNA, and processed into a form that is loaded into Cas-CRISPR RNPs. Finally, in the interference stage, these RNPs utilize the CRISPR RNAs to recognize target RNA or DNA, and the element is silenced, either through cleavage, binding or some other mechanism.

Stage 1: Adaptation

The first step in the adaptation stage of CRISPR/Cas function is recognition of foreign nucleic acids (see Figure 1.5). There are several arguments that this nucleic acid is double stranded DNA: 1) phage-derived spacers have been seen that originate from both sense and antisense strands of phage DNA, and no preference has been seen for open reading frame-derived spacers 2) most viruses that invade prokaryotes are DNA, and 3) RNA-based uptake of spacers would require RNA-dependent DNA polymerase, or reverse transcriptase (RT) activity, and few CRISPR containing organisms contain Cas proteins with predicted RT activity (Mojica et al., 2009). It

therefore seems reasonable to assume that spacers are taken up from double-stranded DNA molecules.

Short sequences have been found near the spacer-targeting, proto-spacers that seem to act as recognition sequences. These proto-spacer-adjacent motifs, or PAMs, have been seen in numerous phage-spacer matches. PAMs are typically 2-5 nucleotides long, and are found directly adjacent of the protospacer (Deveau et al., 2008; Mojica et al., 2009). When analyzed in the context of the repeat sequence clusters described above, it was found that the clusters that were predicted to produce folded RNAs possessed unique PAM sequences, while the unfolded clusters all possessed a generic "NGG" PAM sequence (Mojica et al., 2009). The PAMs were shown to be in a uniform orientation with respect to their orientation in the CRISPR locus. That is, the end of the protospacer that was adjacent to the PAM is oriented towards the leader. This suggests that the PAMs are not only used as recognition sites for the Cas proteins, but may also be involved in orienting the protospacers for insertion into the CRISPR locus, which is the second step in the adaptation phase.

Little is known about the mechanism for insertion of novel spacers into the CRISPR loci. Spacers are inserted in a polarized fashion from the leader. In other words, the newest additions to a given CRISPR locus are found closer to the leader, and the further away you move from the leader, the older the insertions (Figure 1.3). It has been seen by sequencing that both insertion and deletion of spacers occurs rapidly during phage infection (Andersson and Banfield, 2008; Deveau et al., 2008; Tyson and Banfield, 2008). No Cas proteins have been unequivocally linked to this adaptation phase of the CRISPR/Cas system. Barrangou et al. observed that they were unable to

make phage-resistant mutants in a *Streptococcus thermophilus* strain devoid of a protein that they call Cas7, but no additional evidence has been observed (Barrangou et al., 2007). There is speculation that the highly conserved core proteins are involved in the adaptation phase, as integration of spacers into CRISPR loci is a step that must exist in all CRISPR-containing organisms, but much more experimentation is needed in this area (Makarova et al., 2006).

Stage 2: Expression

The second stage of the CRISPR/Cas mechanisms is termed expression. This phase of the system involves the biogenesis of CRISPR RNA-Cas protein complexes, and will be the primary focus of Chapter 2 of this work. Early studies showed that RNAs derived from the CRISPR loci are transcribed as long RNAs, presumably starting from a promoter-like sequence within the leader and encompassing the entire locus, although transcripts have been seen from the opposite strand in *Sulfolobus solfataricus* (Brouns et al., 2008; Hale et al., 2008; Lillestol et al., 2006; Lillestol et al., 2009). These long RNAs are rapidly processed into smaller RNAs that possess a single spacer-repeat unit (Tang et al., 2002; Tang et al., 2005). Further analysis presented in this work and others has shown that these RNAs are even further processed into RNAs containing one invader-targeting sequence and only a small portion of the repeat sequence (Brouns et al., 2008; Hale et al., 2008; Lillestol et al., 2006). These small RNAs are thought to be the mature species that are involved in phage defense in CRISPR-containing organisms, and are called either CRISPR (cr)RNAs or prokaryotic silencing (psi)RNAs. This work will use the term psiRNAs. In *Escherichia coli* and *Pyrococcus furiosus*, it was clearly shown that the 5' end of the small CRISPR RNAs is formed with

precisely eight nucleotides of the 5' repeat. This small sequence is referred to as the psi-tag (this work, (Hale et al., 2009)) or the handle (Brouns et al., 2008), and has also been observed in *Staphylococcus epidermis* (Marraffini and Sontheimer, 2008). It is thought that this small sequence may serve as a recognition signal for the Cas proteins. Further, recent work has shown this sequence to be involved in discrimination between self and non-self DNA (Marraffini and Sontheimer, 2010). This will be discussed further below.

The discovery of the psi-tag led to the hypothesis that there exists a Cas protein that is responsible for cleaving the repeat sequence of the precursor CRISPR RNAs in a manner that leaves a precise 5' tag sequence. Indeed, this activity has been seen in both *Pyrococcus furiosus* and *Escherichia coli*. In *Pyrococcus furiosus*, one of the core cas proteins, Cas6, was found to specifically bind and cleave the *Pyrococcus furiosus* repeat sequence exactly 8 nucleotide from the 3' end of the repeat, forming the 5' AUUGAAAG psi-tag and leaving 22 nucleotides of repeat on the 3' end (Carte et al., 2008). Cas6 is a member of the RAMP family of Cas proteins, and was identified as a core Cas protein by Haft et al. (Haft et al., 2005). This activity was found to be metal-independent, and could occur on both a single repeat and a substrate that contained multiple repeat sequences (Carte et al., 2008). This implies that this protein alone could be responsible for cleaving the long CRISPR transcript down from multiple to single units of one invader-derived sequence and portions of the repeat on each end (see Figure 1.6). An *Escherichia coli* Cas protein, CasE or Cse3, has been shown to perform a highly similar function (Brouns et al., 2008). Like Cas6, cse3 was found to generate a 5' 8-nucleotide tag sequence, AUAAACCG. This protein was purified as a

part of a Cas protein complex termed Cascade, for CRISPR-associated complex for antiviral defense. This complex contains the *Escherichia coli* subtype-specific proteins Cse1, Cse2, Cse3, Cse4 and Cas5e. However, like in *Pyrococcus furiosus*, activity of the cse3 protein was not dependent on any other Cas proteins (Brouns et al., 2008).

Crystal structures were obtained of both *Pyrococcus furiosus* Cas6 and a Cse3 homolog in *Thermus thermophilus* (Brouns et al., 2008; Carte et al., 2008). Both proteins have been classified as potential members of the RAMP superfamily, which was originally proposed to be simply a large group of RNA-binding proteins. The structures of these proteins revealed a duplicated ferredoxin-like fold, which is common in many RNA-binding domains, such has the RNA Recognition Motif, (RRM) (Brouns et al., 2008; Carte et al., 2008).

Although it appears that the 5' end of psiRNAs in both *Escherichia coli* and *Pyrococcus furiosus* is well-defined, both organisms have shown heterogeneity in the 3' end. In *Pyrococcus furiosus*, it is shown (in this work) that most mature-sized psiRNAs contain no 3' repeat sequence, and end at or near the spacer-repeat boundary (Hale et al., 2008; Hale et al., 2009). In *Escherichia coli*, the major species of psiRNAs often contain portions of the 3' repeat, which was referred to as the 3' handle (Brouns et al., 2008). This heterogeneity likely implies that the generation of the 3' end of the psiRNAs is performed by a less precise fashion, perhaps by an uncharacterized exonuclease.

Processed psiRNAs have been shown to be incorporated into Cas protein-containing RNPs. In *Escherichia coli*, the processed psiRNAs remain incorporated in the Cascade complex, which contained the Cse3 protein responsible for biogenesis (Brouns et al., 2008). In this work, we show that mature-sized *Pyrococcus furiosus*

psiRNAs are loaded into at least two chromatographically distinct complexes, one of which contains the Cmr proteins, one which is yet uncharacterized, and neither of which has been shown to contain Cas6 (Hale et al., 2009). Therefore, while sorting may not be necessary for a simplified system such as *Escherichia coli*, in *Pyrococcus furiosus* which contains a more complete set of cas proteins, it seems that mature psiRNAs are sorted into several Cas-protein containing complexes by a currently unknown mechanism.

Stage 3: Interference

Once the psiRNAs have been processed and sorted into the appropriate Cas protein complexes, we move into the third and final stage of the CRISPR/Cas mechanism: interference. In this stage, target RNAs or DNAs are recognized, presumably through basepairing with the psiRNAs, and the invader is silenced, either through cleavage, binding, or another unknown mechanism. The mechanism of interference in *Pyrococcus furiosus* is the main focus of Chapter 3 of this work.

A major argument exists in the field about the nature of the targeted element: does the CRISPR/Cas system target DNA or RNA? Initial comparisons between the CRISPR/Cas system and the RNAi pathway in eukaryotes led to the hypothesis that RNA is targeted. Indeed, in this work, we show that a complex of Cmr proteins and psiRNAs in *Pyrococcus furiosus* can target and cleave complementary RNA sequences. However, there is evidence in other organisms that DNA may also be the target of the interference mechanism. For example, bioinformatic analysis of the spacers with known origins has not shown any preference for the coding strand of extrachromosomal elements. Indeed, in *Streptococcus thermophilus*, the spacers that were shown to be

taken up in response to viral infection were done so from both strands of the phage genome (Barrangou et al., 2007). Also, no preference is shown for spacers that derive from the open reading frames of viral genomes (Barrangou et al., 2007). In *Escherichia coli*, interference was shown to depend on both the Cascade complex and the core protein cas3. This interference was shown to function on artificial spacers in both orientations, indicating that Cascade-mediated interference likely occurs on the DNA level (Brouns et al., 2008). In *Staphylococcus epidermis*, a spacer exists against the *nickase* gene, which is present on all known Straphylococcus conjugative plasmids. This spacer is identical, not complementary to the *nickase* mRNA, and has been shown to be functional. A self-splicing intron was engineered into the *nickase* gene of the plasmid's proto-spacer, causing the proto-spacer to be destroyed in the DNA, but not the mRNA of the nickase gene. Conjugation was not inhibited, which indicated a lack of CRISPR response, chalking up another argument for interference on the DNA, not RNA of the targeted sequenced in this organism (Marraffini and Sontheimer, 2008).

Although much energy has been spent on the debate between DNA and RNA, it seems likely that, as mentioned above, different classes of Cas proteins are likely to function on different targets. Indeed, as mentioned above, two distinct complexes of Cas proteins have been shown in two organisms to target two different substrates. In *Escherichia coli*, CRISPR-mediated interference was assayed by engineering artificial psiRNAs against the λ phage, and sensitivity to λ infection was assayed by plaque assays. Introduction of neither the Cascade complex or Cas3 alone was sufficient to cause phage resistance, but in a strain with both Cascade and Cas3, efficient resistance was shown (Brouns et al., 2008)

In this work, we show that the Cmr proteins act to sequence-specifically recognize and cleave target RNAs. A complex containing Cmr1-6 was found to be associated with two size classes of psiRNAs, a 45-nt species and a 39-nt species. This complex was purified from *Pyrococcus furiosus* cell extract, and was also reconstituted *in vitro*. Both recombinant and native complexes were shown to cleave any RNA that has a complementary sequence. This complex does not cleave DNA, or double stranded RNA. *In vitro* assays showed that only Cmr5 was not required for activity. A mechanism is proposed in which the cleavage site is determined by counting from the 3' end of the incorporated psiRNA (Chapter 3, (Hale et al., 2009).

An important factor in the interference stage of the CRISPR/Cas system, especially on DNA-targeting mechanisms is self vs. non-self recognition. If the CRISPR system targets DNA that contains the same sequence that is in the CRISPR loci, it is extremely important that CRISPR loci within the genome are not cleaved. Therefore, some mechanism to recognize self vs. non-self DNA must exist. Indeed, recent work in *Staphylococcus epidermis* has established an elegant mechanism for such CRISPR locus protection (Marraffini and Sontheimer, 2010). This work found that non-complementarity between the protospacer and the psi-tag region of the psiRNAs allowed for silencing to occur. When the target sequence was mutated to allow for base-pairing between the tag region of the psiRNA and the target, the target was able to evade silencing (Marraffini and Sontheimer, 2010). This work not only proposes a role for the conserved psi-tag sequence, but also proposes an efficient mechanism for avoiding targeting and destruction of self DNA, a vital part of the interference pathway.

Although this work reveals a mechanism for RNA-targeted CRISPR interference, no mechanism has been shown for DNA-targeted interference. This silencing could occur through cleavage, or may involve other mechanisms, such as irreversible binding, or inhibition of transcription through other mechanisms, such as inducing repressive chromatin structures. Indeed, it seems likely that there are diverse mechanisms of silencing in organisms with various subsets of cas proteins, and work must be performed in a variety of organisms in order to truly understand the role of Cas proteins in CRISPR-mediated interference.

The CRISPR/Cas system of *Pyrococcus furiosus*

The work that is presented here will focus on the CRISPR/Cas system of *Pyrococcus furiosus*. *Pyrococcus furiosus* is a hyperthermophilic archaeon that was first isolated in geothermal vents off the coast of Italy. This organism is somewhat unique in that it has an optimal growth temperature of 100°C, the boiling point of water. The naturally heat-stable nature of its proteins in both the native and recombinant form makes it an ideal model system for studying RNA-protein complexes (Adams, 1994).

Another advantage of studying the CRISPR/Cas system in *Pyrococcus furiosus* is the complete nature of its components (Figure 1.8). There are 7 CRISPR loci in the genome of *Pyrococcus furiosus* which encode for a total of about 200 psiRNAs. Each of the 7 loci contains a leader sequence. The sequence of the repeats is highly conserved among the loci, with only a few nucleotides difference between any two of the repeat sequences. All of the repeats fall in the unstructured cluster #6 (Kunin et al., 2007).

Pyrococcus furiosus contains two clusters of cas genes. The largest cluster is located immediately adjacent to CRISPR locus 7 (as annotated by UCSC, (Schneider et al., 2006)) and contains 15 genes. Included in this cluster are the core cas genes 1-6, the RAMP module (cmr1-6), and a cluster of three Tneap subtype-specific genes. A second cluster of cas genes is found between loci 5 and 6. This smaller group of genes contains a group of 5 Apern subtype-specific genes, two genes that together form the core protein Cas3, and one unidentified cas gene (see Figure 1.8). Also found throughout the *Pyrococcus furiosus* genome are six genes that have also been predicted to be CRISPR-associated genes. All in all, *Pyrococcus furiosus* contains 200 psiRNAs, 29 cas genes of which 9 are core cas genes, 15 are members of two different subtypes and the RAMP module, and 5 have unidentified classifications (Haft et al., 2005). This wide variety of proteins makes *Pyrococcus furiosus* an excellent system for the study of the functions of the proteins, both by *in vitro* studies involving recombinant proteins, and by isolation and characterization of native complexes from *Pyrococcus furiosus* cells.

Organization of this work

This work will focus on the identification and characterization of psiRNAs and psiRNPs in *Pyrococcus furiosus*. In Chapter 2, the expression and processing of RNAs derived from the CRISPR loci is examined. In addition, mature sized psiRNAs were shown to exist in several chromatographically distinct protein-containing complexes. In Chapter 3, a one of these Cas protein-psiRNA containing complexes was isolated and characterized. This complex was shown to contain the Cmr proteins, and had the ability to sequence-specifically recognize and cleave target RNAs. In the appendix, a

separate work has been included which focuses on tRNA modification in *Pyrococcus furiosus* either by a single protein, PsuX or by the protein components of the H/ACA sRNP that guides pseudouridylation of rRNAs in the presence of guide H/ACA sRNAs.

REFERENCES

Adams, M.W. (1994). Biochemical diversity among sulfur-dependent, hyperthermophilic microorganisms. FEMS Microbiol Rev *15*, 261-277.

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

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

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.

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

Bergh, O., Borsheim, K.Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. Nature *340*, 467-468.

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.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089-1103.

Brosius, J. (2003). Gene duplication and other evolutionary strategies: from the RNA world to the future. J Struct Funct Genomics *3*, 1-17.

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.

Buhler, M., and Moazed, D. (2007). Transcription and RNAi in heterochromatic gene silencing. Nat Struct Mol Biol *14*, 1041-1048.

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.

Chopin, M.C., Chopin, A., and Bidnenko, E. (2005). Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol *8*, 473-479.

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.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat Rev Genet *2*, 919-929.

Ferre-D'Amare, A.R. (2003). RNA-modifying enzymes. Curr Opin Struct Biol *13*, 49-55.

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

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

Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science *315*, 1587-1590.

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

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

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

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.

Hartig, J.V., Tomari, Y., and Forstemann, K. (2007). piRNAs--the ancient hunters of genome invaders. Genes Dev *21*, 1707-1713.

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

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B.*, et al.* (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell *129*, 69-82.

Huttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet *21*, 289-297.

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

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

Kim, V.N. (2006). Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. Genes Dev *20*, 1993-1997.

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

Lau, N.C., Robine, N., Martin, R., Chung, W.J., Niki, Y., Berezikov, E., and Lai, E.C. (2009). Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res *19*, 1776-1785.

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.

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

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

Marques, J.T., and Carthew, R.W. (2007). A call to arms: coevolution of animal viruses and host innate immune responses. Trends Genet *23*, 359-364.

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). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568-571.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563-574.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol *8*, 209-220.

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.

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

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.

Saito, K., Nishida, K.M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and Siomi, M.C. (2006). Specific association of Piwi with rasiRNAs derived from

retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev *20*, 2214-2222.

Sakamoto, K., Agari, Y., Agari, K., Yokoyama, S., Kuramitsu, S., and Shinkai, A. (2009). X-ray crystal structure of a CRISPR-associated RAMP superfamily protein, Cmr5, from Thermus thermophilus HB8. Proteins *75*, 528-532.

Schneider, K.L., Pollard, K.S., Baertsch, R., Pohl, A., and Lowe, T.M. (2006). The UCSC Archaeal Genome Browser. Nucleic Acids Res *34*, D407-410.

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

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434-1437.

Talini, G., Gallori, E., and Maurel, M.C. (2009). Natural and unnatural ribozymes: back to the primordial RNA world. Res Microbiol *160*, 457-465.

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.

Tock, M.R., and Dryden, D.T. (2005). The biology of restriction and anti-restriction. Curr Opin Microbiol *8*, 466-472.

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 Rij, R.P., and Berezikov, E. (2009). Small RNAs and the control of transposons and viruses in Drosophila. Trends Microbiol *17*, 163-171.

Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. FEMS Microbiol Rev *28*, 127-181.

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.

Figure 1.1. **The eukaryotic RNAi pathway.** In eukaryotic cells, double-stranded RNA (such as is formed during most viral life cycles) is recognized by a RNaseIII-like protein Dicer, which cleaves it into short (21-28), double-stranded RNAs with 3' overhangs. One strand of these small interferening, or siRNAs are loaded into the Argonaute protein containing RNA-Induced Silencing Complex, or RISC. The RNA-loaded RISC complex recognizes complementary RNA sequences, and leads to cleavage of the viral mRNA.

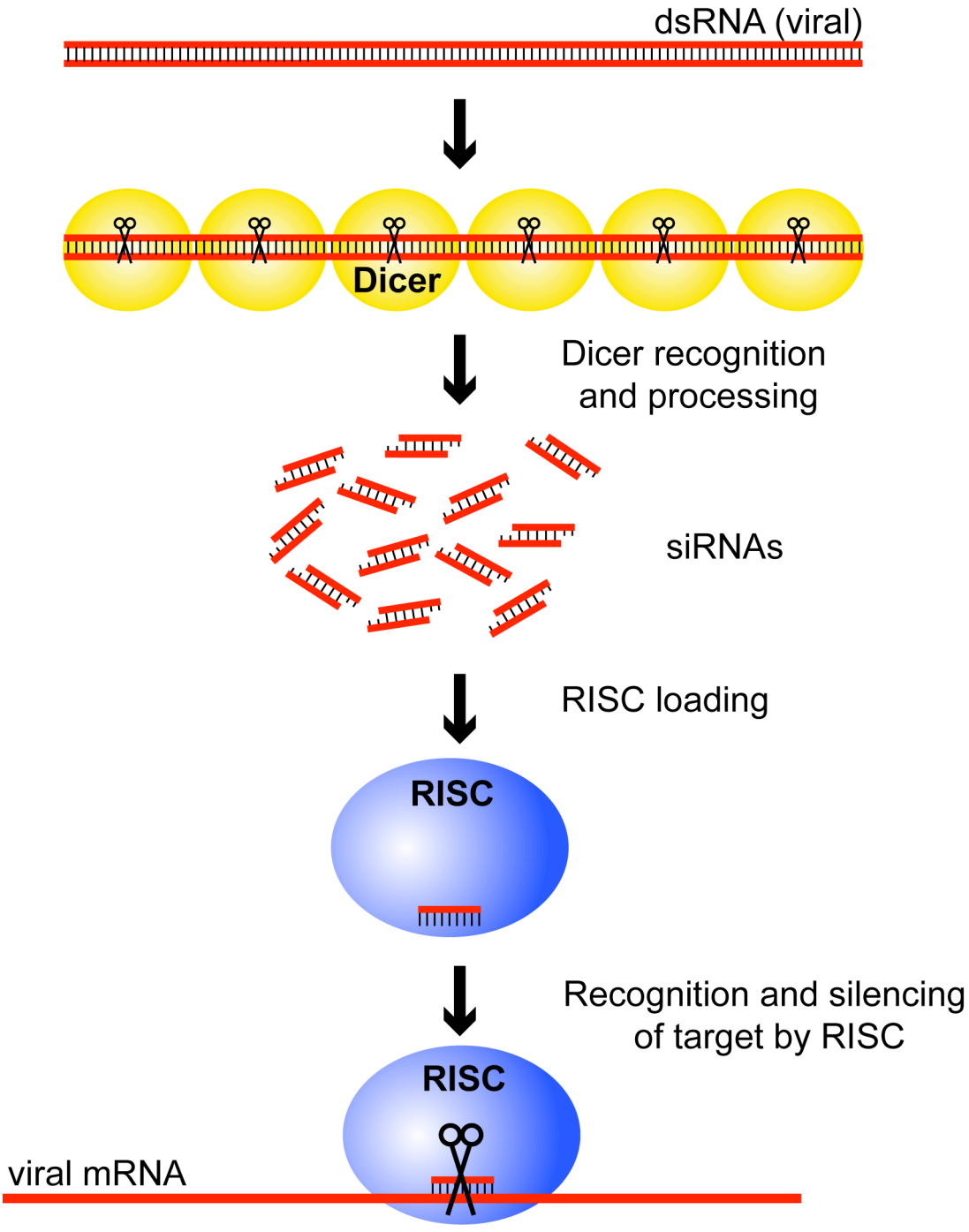

dsRNA (viral)
Dicer
Dicer recognition
and processing
siRNAs
RISC loading
RISC
Recognition and silencing
of target by RISC
RISC
viral mRNA

Figure 1.2. **The ping-pong cycle for amplification of piRNAs.** Primary piRNAs, which are transcribed from genomic piRNA clusters and contain a 5' adenosine, are bound by the Argonaute family proteins Piwi and Aubergene (red). These RNA-protein complexes recognize and cleave transposon mRNAs. The cleavage products are recognized by Ago3, another Argonaute family protein (blue). 3' end processing occurs by an unknown mechanism, generating a secondary piRNA that contains a U in the 10^{th} position. This secondary piRNA-Ago3 complex recognizes and cleaves piRNA cluster transcripts, which generates a pool of primary piRNAs that are antisense to currently expressed transposon mRNAs.

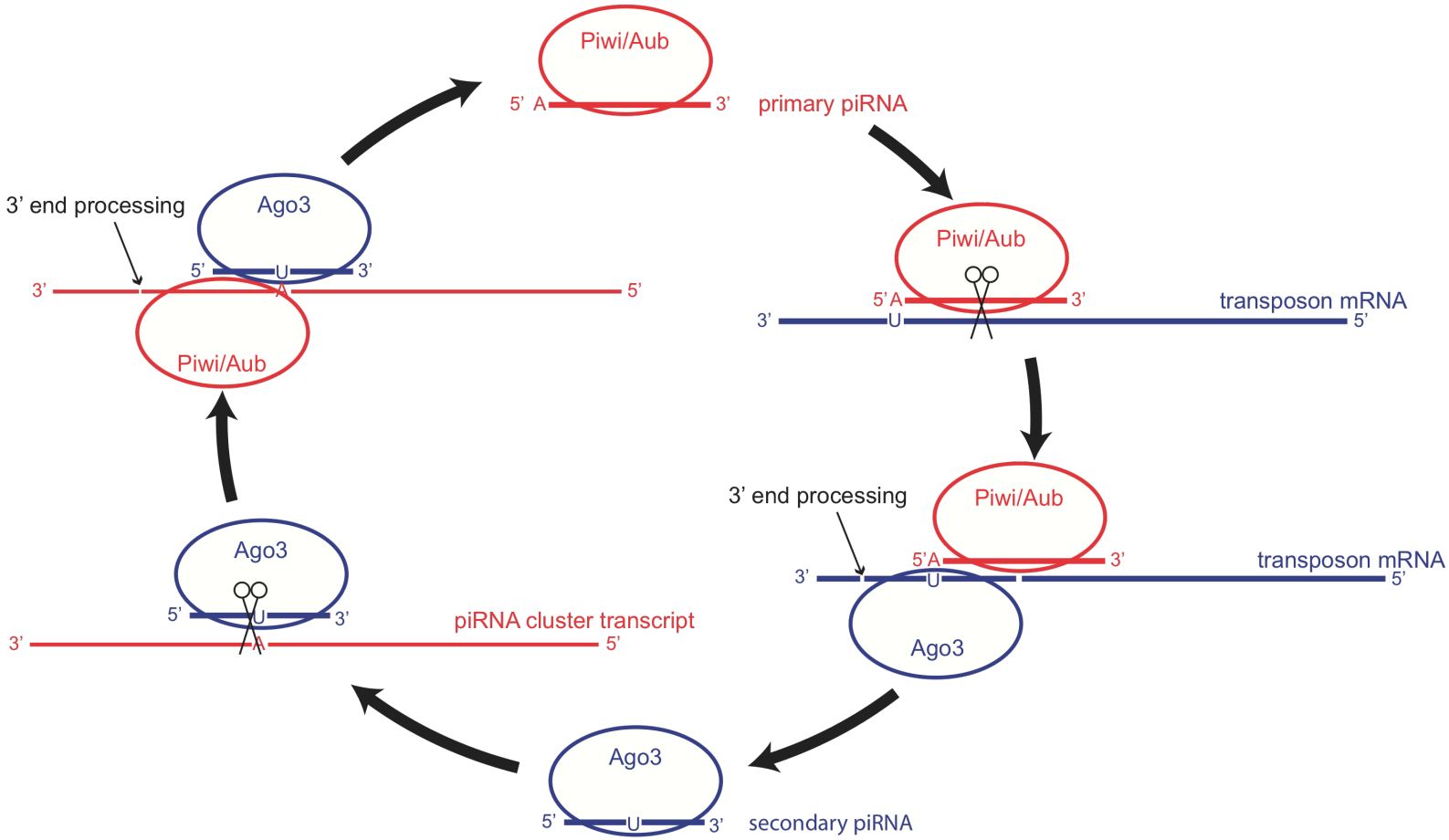

Piwi/Aub
5' A 3'
primary piRNA
Piwi/Aub
5' A 3'
3' U 5'
transposon mRNA
3' end processing
Piwi/Aub
5' A 3'
3' U 5'
transposon mRNA
Ago3
Ago3
5' U 3'
3' A 5'
piRNA cluster transcript
3' end processing
Ago3
5' U 3'
3' A 5'
Piwi/Aub
Ago3
5' U 3'
secondary piRNA

Figure 1.3. **A typical CRISPR locus.** A typical CRISPR locus consists of short, direct repeats (shown in black), separated by unique spacer sequences (shown in colors). Upstream of the repeats is a leader sequence (white) that is conserved among the CRISPR loci of a given organism, but not among different organisms. Often, an operon of cas genes is found adjacent to one end of the CRISPR locus (shown here in gray, adjacent to the leader). It has been shown that the spacers that are nearest to the leader are the most recently acquired spacers, while those more distal to the leader are derived from older infections.


newly acquired spacers
older spacers
cas gene operon
leader
R S R S R S R S R S R S R
• • •

Table 1.1. **Classificiation of the CRISPR-associated proteins, and their predicted and observed functions.** References are provided as appropriate.

Gene Group	Gene Name	Associated TIGRs[1]	Predicted function	Observed function
Core	cas1	TIGR00287	Nuclease[1]	DNA endonuclease (P. aerigonosas)[3] RNA/DNA binding, promotes annealing (S. solfataricus)[4]
	cas2	TIGR01573 TIGR01873	VapD-like nuclease[2]	ssRNA endonuclease (S. solfataricus)[5]
	cas3	TIGR01587 TIGR01596 TIGR02562	Helicase/HD-Nuclease[1,2]	double-strand specific endonuclease (S. solfataricus)[6] Involved in interference (*Escherichia coli*)[7]
	cas4	TIGR00372	RecB-family exonuclease[1,2]	--
	cas5	TIGR02593	RAMP[2]	--
	cas6	TIGR01877	RAMP[1]	Endonuclease, cleavage of repeat sequences (*Pyrococcus furiosus*)[8]
Apern	csa1	TIGR01896	--	--
	csa2	TIGR02583	Regulator[1]	--
	csa3	TIGR01884	--	--
	csa4	TIGR01914	--	--
	csa5	TIGR01878	--	--
	cas5a	TIGR01874	--	--
Dvulg	csd1	TIGR01863	--	--
	csd2	TIGR02589	--	--
	cas5d	TIGR01876	--	--
Ecoli	cse1	TIGR02547	--	Form the Cascade complex, which is involved in both psiRNA biogenesis and interference in *Escherichia coli*. Cse3 cleaves repeat sequences[7]
	cse2	TIGR02548	--	
	cse3	TIGR01907	--	
	cse4	TIGR01869	--	
	cas5e	TIGR01868	--	
Hmari	csh1	TIGR02591	--	--
	csh2	TIGR02590	Regulator[1]	--
	cas5h	TIGR02592	--	--
Mtube	csm1	TIGR02578	Polymerase[1]	--
	csm2	TIGR01870	--	--
	csm3	TIGR02582	RAMP[1]	--
	csm4	TIGR01903	RAMP[1]	--
	csm5	TIGR01899	RAMP[1]	--
Nmeni	csn1	TIGR01865	Endonuclease[1]	Interference? (S. thermophilus)[9]
	csn2	TIGR01866	--	--

Tneap	cst1	TIGR01908	--	--
	cst2	TIGR02585	Regulator[1]	--
	cas5t	TIGR01895	--	--
Ypest	csy1	TIGR02564	--	--
	csy2	TIGR02565	--	--
	csy3	TIGR02566	--	--
	csy4	TIGR02563	--	--
RAMP	cmr1	TIGR01864	RAMP[1]	Form an RNA-protein complex with mature psiRNAs, cleaves complementary ssRNA in *Pyrococcus furiosus* (Cmr5 is not required)[10]
	cmr2	TIGR02577	Polymerase[1]	
	cmr3	TIGR01888	RAMP[1]	
	cmr4	TIGR02580	RAMP[1]	
	cmr5	TIGR01881	--	
	cmr6	TIGR01898	RAMP[1]	

[1] Haft, D. H., J. Selengut, et al. (2005). "A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes." PLoS Comput Biol 1(6): e60.

[2] Makarova, K. S., N. V. Grishin, et al. (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.

[3] Wiedenheft, B., K. Zhou, et al. (2009). "Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense." Structure 17(6): 904-12.

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

[5] Beloglazova, N., G. Brown, et al. (2008). "A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats." J Biol Chem 283(29): 20361-71.

[6] Han, D. and G. Krauss (2009). "Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2." FEBS Lett 583(4): 771-6.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

Figure 1.4. **Overview of a proposed 3-stage mechanism of the CRISPR/Cas system for genome defense.** The first stage of the CRISPR/Cas mechanism is the adaptation phase. In this phase, foreign DNA (shown here as phage DNA, blue and red) is recognized by the Cas proteins, and the protospacer (red) is inserted into the CRISPR locus near the leader. In stage 2, expression, the CRISPR locus is transcribed and processed into small RNAs that contain a single spacer sequence and small amounts of the 5' repeat sequence. The resulting small RNAs are loaded into Cas-protein containing complexes (gray). In the third stage, interference, the resulting psiRNA-Cas protein complexes recognize and silence either foreign single-stranded RNA transcribed from the phage DNA through cleavage, or directly silence the phage DNA by an unknown mechanism.

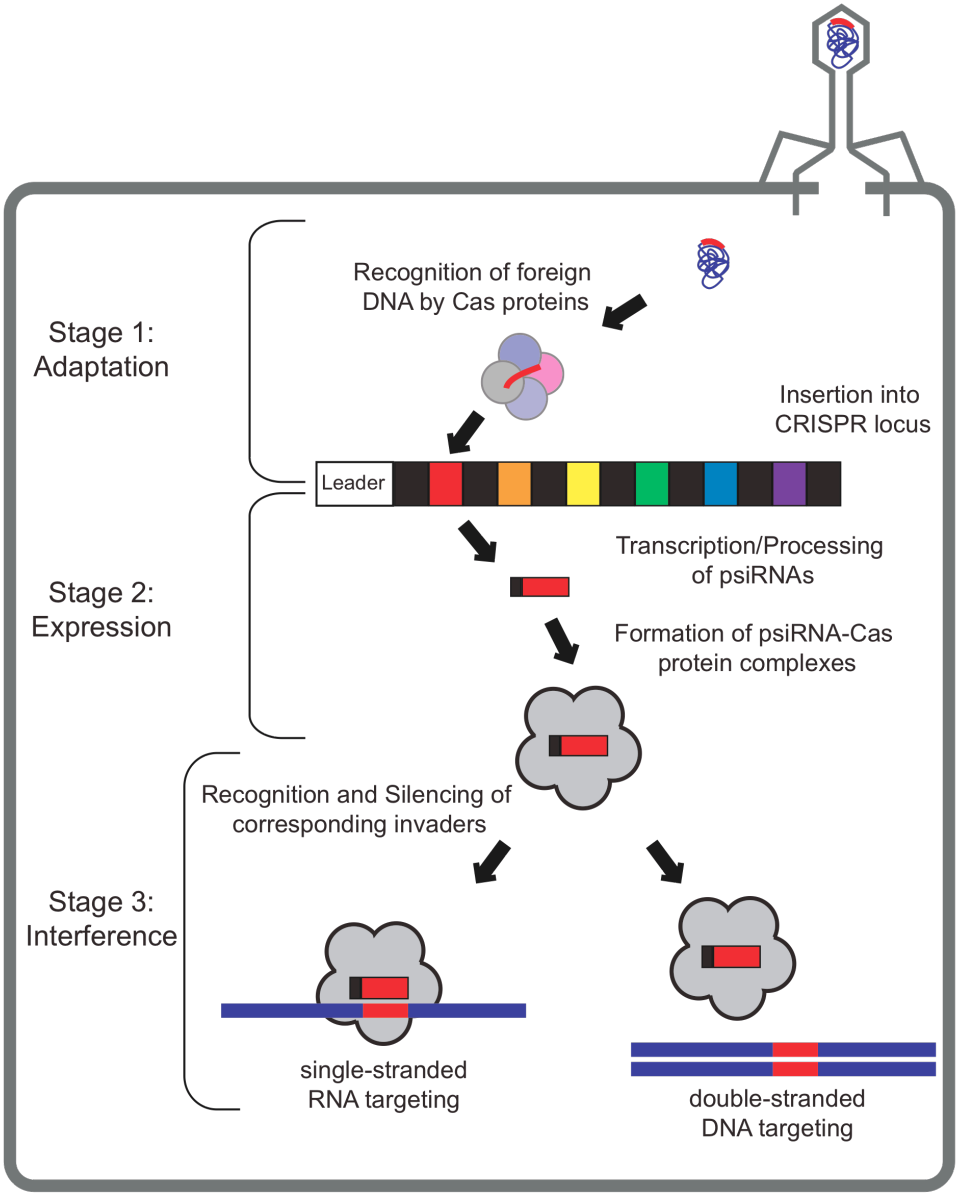

Stage 1:
Adaptation
Recognition of foreign
DNA by Cas proteins
Insertion into
CRISPR locus
Leader
Stage 2:
Expression
Transcription/Processing
of psiRNAs
Formation of psiRNA-Cas
protein complexes
Recognition and Silencing of
corresponding invaders
Stage 3:
Interference
single-stranded
RNA targeting
double-stranded
DNA targeting

Figure 1.5. The adaptation phase. In the adaptation phase of the CRISPR/Cas mechanism, viral DNA (red) is proposed to be recognized by the Cas proteins (gray, pink and purple), and proto-spacers are determined by the presence of a short protospacer-adjancent motif, or PAM (green). The viral DNA is cleaved and inserted into the CRISPR locus at the leader end. The PAM is used to orient the spacer in the CRISPR locus at the leader. The result of this process is the addition of a novel repeat/spacer unit into the leader-proximal end of the CRISPR.

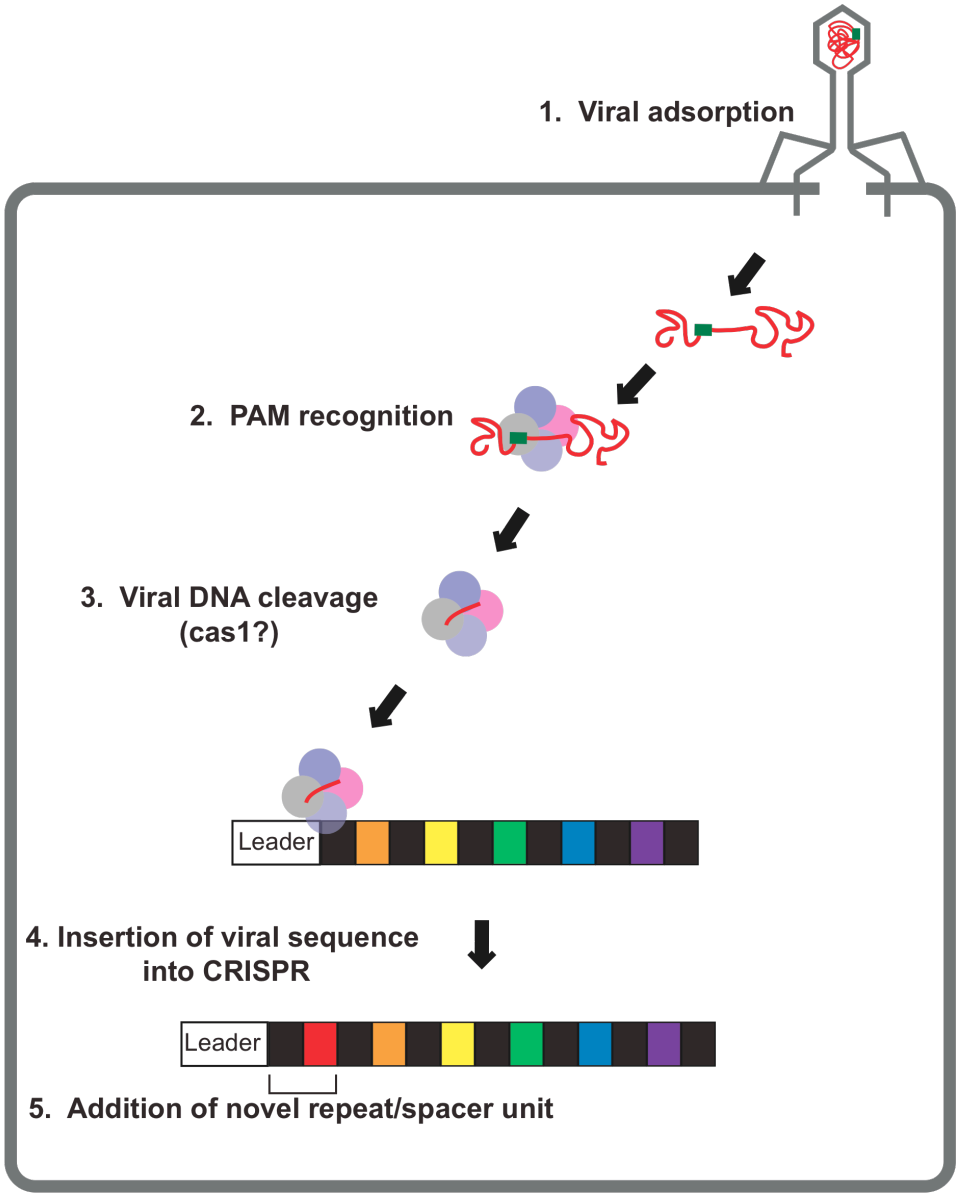

1. Viral adsorption
2. PAM recognition
3. Viral DNA cleavage
(cas1?)
Leader
4. Insertion of viral sequence
into CRISPR
Leader
5. Addition of novel repeat/spacer unit

Figure 1.6. **Expression of the CRISPR loci.** A CRISPR locus is transcribed as a long transcript that encompasses the entire locus. This transcript is processed by recognition and cleavage of the repeat sequences by the Cas6/Cse3 homologs. These enzymes cleave the repeats 22 nts from the 5' end of the repeat. This cleavage results in a pool of 1X intermediates, which contain a single spacer, 8 nts of repeat on the 5' end, and 22 nts of repeat on the 3' end. These intermediates are undergo 3' end processing by an uncharacterized mechanism. This results in a pool of mature psiRNAs which contain a full or most of a spacer, and 8 nucleotides of repeat on the 5' end. We refer to this 8-nucleotide sequence as a psi-tag. These mature psiRNAs are sorted by an unknown mechanism into various Cas protein-containing complexes.

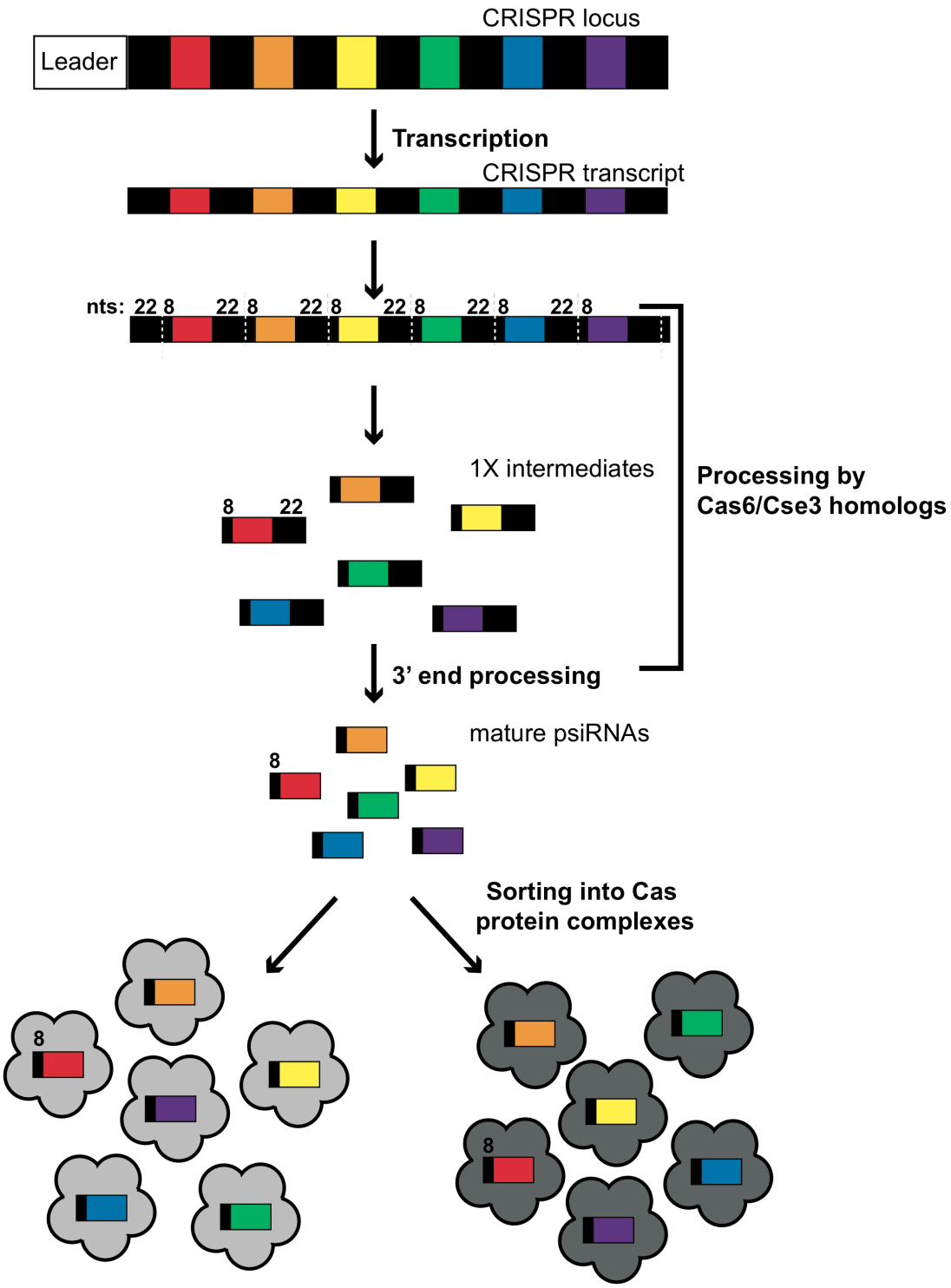

CRISPR locus
Leader
Transcription
CRISPR transcript
nts: 22 8 22 8 22 8 22 8 22 8 22 8
1X intermediates
8 22
Processing by
Cas6/Cse3 homologs
3' end processing
mature psiRNAs
8
Sorting into Cas
protein complexes
8
8

Figure 1.7. **The interference stage: silencing of foreign nucleic acids.** Once the psiRNA-Cas protein complexes are formed (gray and red), they are able to recognize foreign nucleic acids (blue and red). This targeted nucleic acid may be DNA or RNA. In an RNA-targeting pathway (top), the viral genome is transcribed into mRNAs, which, if complementary to incorporated psiRNAs, is cleaved, preventing translation. In *Pyrococcus furiosus*, this activity is performed by the members of the cas module RAMP, or Cmr proteins. In a DNA-targeting system, the psiRNA-Cas protein complex (light gray and green) directly recognizes the viral double-stranded DNA (blue and green) and prevents infection, either through cleavage, irreversible binding, or another uncharacterized mechanism. The complexes involved in DNA targeting may be made of Cas proteins that are grouped into the subtype specific complexes, or may consist of a subset of the core Cas proteins.

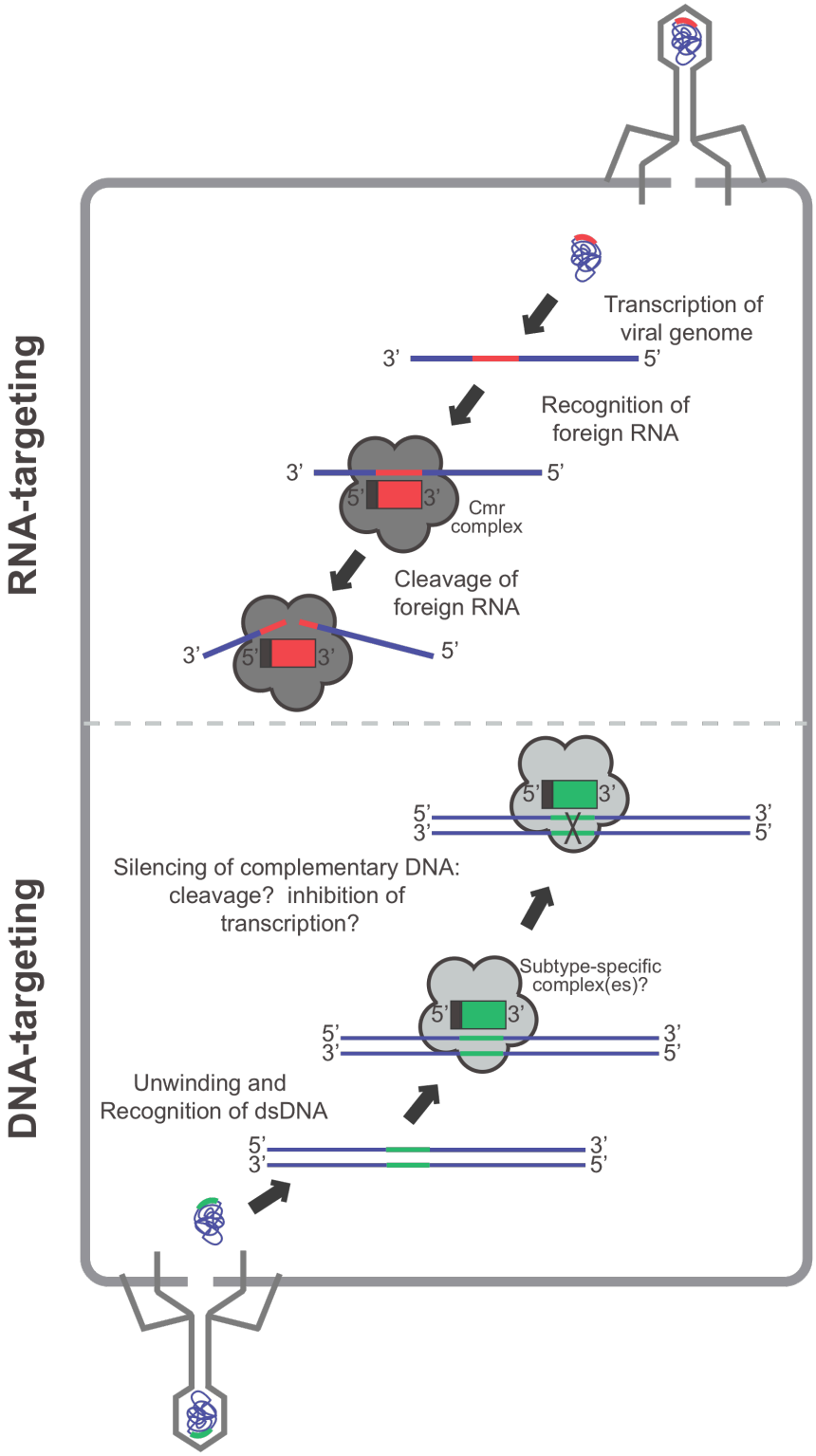
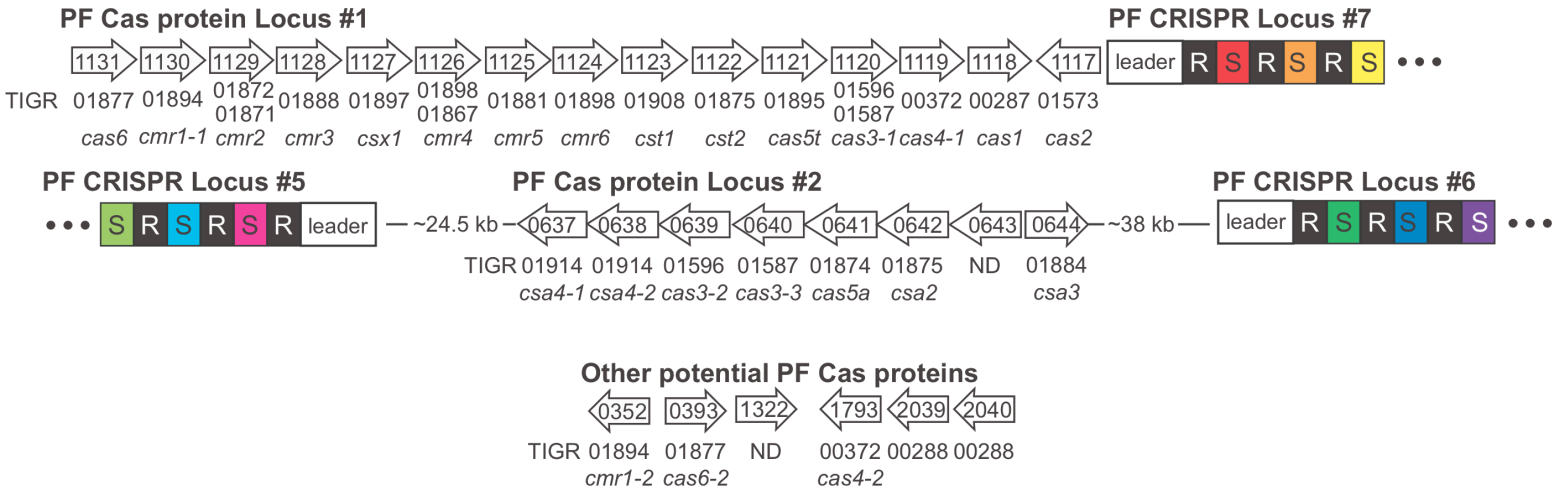

RNA-targeting
DNA-targeting
Transcription of viral genome
3'
5'
Recognition of foreign RNA
Cmr complex
Cleavage of foreign RNA
Silencing of complementary DNA: cleavage? inhibition of transcription?
Subtype-specific complex(es)?
Unwinding and Recognition of dsDNA

Figure 1.8. **The CRISPR/Cas system of *Pyrococcus furiosus*.** *Pyrococcus furiosus* contains two large operons of cas genes. The first, larger operon (PF Cas protein Locus #1) is directly adjancent to PF CRISPR Locus #7. White arrows represent the cas genes. The numbers within the arrow represent the gene name. Under the arrows, the TIGR number assigned to each gene is given (Haft et al., 2005), and the gene name is also given (Haft et al., 2005). The second cas gene operon is located between CRISPR locus #5 and #6. There are also a few other genes that have TIGR number assignments that are associated with the CRISPR/Cas system.


PF Cas protein Locus #1
PF CRISPR Locus #7
1131 1130 1129 1128 1127 1126 1125 1124 1123 1122 1121 1120 1119 1118 1117
leader R S R S R S
TIGR 01877 01894 01872 01871 01888 01897 01898 01867 01881 01898 01908 01875 01895 01596 01587 00372 00287 01573
cas6 cmr1-1 cmr2 cmr3 csx1 cmr4 cmr5 cmr6 cst1 cst2 cas5t cas3-1 cas4-1 cas1 cas2
PF CRISPR Locus #5
S R S R S R leader
~24.5 kb
PF Cas protein Locus #2
0637 0638 0639 0640 0641 0642 0643 0644
~38 kb
PF CRISPR Locus #6
leader R S R S R S
TIGR 01914 01914 01596 01587 01874 01875 ND 01884
csa4-1 csa4-2 cas3-2 cas3-3 cas5a csa2 csa3
Other potential PF Cas proteins
0352 0393 1322 1793 2039 2040
TIGR 01894 01877 ND 00372 00288 00288
cmr1-2 cas6-2 cas4-2

CHAPTER 2

PROKARYOTIC SILENCING (PSI)RNAS IN *PYROCOCCUS FURIOSUS*[1]

[1] Hale, C., K. Kleppe, et al. (2008). "Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus." RNA **14**(12): 2572-9. Reprinted here with permission of publisher.

Abstract

In many prokaryotes, non-coding RNAs that arise from the CRISPR loci are now thought to mediate defense against viruses and other molecular invaders by an RNAi-like pathway. CRISPR (clustered regularly interspaced short palindromic repeat) loci contain multiple short regions of similarity to invader sequences separated by short repeat sequences, and are associated with resistance to infection by corresponding viruses. It is hypothesized that RNAs derived from these regions, termed prokaryotic silencing (psi)RNAs, guide Slicer-like complexes of partner proteins to destroy invader nucleic acids. Here we have investigated CRISPR-derived RNAs in the archaeon *Pyrococcus furiosus*. Northern analysis revealed multiple RNA species consistent with a proposed biogenesis pathway that includes full-length CRISPR locus transcripts and intermediates generated by endonucleolytic cleavages within the repeat sequences. However, our results identify the principal products of the CRISPR loci as small psiRNAs comprised primarily of invader-targeting sequence with perhaps only 5-10 nucleotides of CRISPR repeat sequence. These RNAs are the most abundant CRISPR RNA species in *Pyrococcus furiosus* and are likely the guides for the effector complexes of the proposed prokaryotic RNAi (pRNAi) system. We analyzed cell-free extracts fractionated under non-denaturing conditions and found that the various CRISPR RNA species are components of distinct RNA-protein complexes, including at least two complexes that contain mature-length psiRNAs. Finally, RNAs are produced from all 7 CRISPR loci present in the *Pyrococcus furiosus* genome and interestingly, the most recently acquired psiRNAs encoded proximal to the leader sequence of a CRISPR locus appear to be the most abundant.

Introduction

Small, non-coding (nc)RNAs are found in all domains of life and function in a wide array of essential cellular processes. In eukaryotes, small ncRNAs including siRNAs and microRNAs have been shown to function in post-transcriptional gene silencing by targeting exogenous or endogenous RNAs, respectively, in a process called RNA interference, or RNAi (Hannon, 2002). Another class of small RNAs referred to as piRNAs (piwi-associated) or rasiRNAs (repeat associated small interfering) regulate spreading of selfish genetic elements such as transposons or repeat elements in organisms including mammals, plants and flies (Aravin et al., 2007; Hartig et al., 2007; Kim, 2006; Lin, 2007; Nishida and Siomi, 2006).

An RNAi-like system that functions in genome defense has recently been proposed to exist in prokaryotes (Deveau et al., 2008; Makarova et al., 2006; Sorek et al., 2008; Tyson and Banfield, 2008). The hallmark of the proposed prokaryotic RNAi (or pRNAi) system is the CRISPR locus, a cluster of short direct repeats that separate short variable sequences (i.e. clustered regularly interspaced short palindromic repeat). A number of the variable sequences (also sometimes called “spacers”) found in CRISPR loci display complementarity (or identity) to known prokaryotic viruses, plasmids and transposons (Bolotin et al., 2005; Lillestol et al., 2006; Makarova et al., 2006; Mojica et al., 2005; Pourcel et al., 2005). The other signature component of the hypothesized pRNAi system is a set of protein-coding genes referred to as CRISPR-associated or Cas genes that are found in CRISPR-containing genomes (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2002; Makarova et al., 2006). The Cas genes are predicted to encode nucleases, helicases, RNA-binding proteins and a

polymerase (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2002; Makarova et al., 2006). These bioinformatically-predicted properties of the CRISPR and Cas gene products led to the hypothesis that they comprise an RNAi-like system of genome defense in prokaryotes, in which RNAs derived from the variable regions of CRISPR loci (prokaryotic silencing or psiRNAs) guide the degradation of genome invaders by Cas proteins (Bolotin et al., 2005; Lillestol et al., 2006; Makarova et al., 2006). The Cas proteins are also expected to function in the processing of the psiRNAs and in the integration of new psiRNA genes (directed against newly encountered pathogens) into the genome.

Recent studies have provided strong evidence for a role of CRISPR loci in viral resistance in prokaryotes. Several groups have observed that virus exposure leads to the appearance of new virus-derived sequence elements within the CRISPR loci of surviving (resistant) isolates (Barrangou et al., 2007; Deveau et al., 2008; Horvath et al., 2008). In addition, Barrangou et al. showed that an alteration of an organism's CRISPR sequences that generates or destroys correspondence with a viral sequence results in viral resistance and viral sensitivity, respectively (Barrangou et al., 2007). However, the pathway by which CRISPR loci confer viral resistance remains hypothetical and undefined.

CRISPR loci are present in about half of bacterial genomes and nearly all archaeal genomes (Godde and Bickerton, 2006; Makarova et al., 2006). A given locus can contain as few as 2, and as many as several hundred repeat-psiRNA units (Grissa et al., 2007; Sorek et al., 2008). The repeat sequences are generally 25 to 45 nucleotides long and often weakly palindromic at the 5' and 3' termini (Jansen et al.,

2002). Interspersed between the repeats are the variable, putative psiRNA-encoding sequences, which are usually similar in length to the repeats. RNAs arising from CRISPR loci have been detected by RNA cloning and/or Northern blotting in 3 archaeal species: *Archaeoglobus fulgidus*, *Sulfolobus solfataricus* and *Sulfolobus acidocaldarius* (Lillestol et al., 2006; Tang et al., 2002; Tang et al., 2005). These studies provided convincing evidence of transcription of entire CRISPR loci from the predicted transcriptional leader sequences that are found at one end of the loci, and of a discrete series of smaller RNAs that correspond in length to multiples of repeat-psiRNA units (e.g. ~70, 140, 210, 280 nts, etc. (Tang et al., 2002; Tang et al., 2005)). These findings along with RNA sequence analysis led to a hypothesized biogenesis pathway in which primary CRISPR transcripts are endonucleolytically cleaved within repeat sequences to produce psiRNAs flanked by repeat sequence at both the 5' and 3' ends (Tang et al., 2002; Tang et al., 2005). Intriguingly, very diffuse Northern blot signals suggesting smaller RNAs were also detected in *S. acidocaldarius* (Lillestol et al., 2006).

Here we have investigated the RNAs expressed from the seven CRISPR loci in *Pyrococcus furiosus*. Our results support the previously proposed steps in CRISPR RNA biogenesis, but extend the pathway to include, as the apparent end products of biogenesis, small psiRNAs that likely contain less than 10 nucleotides of repeat sequence. Additionally, fractionation of *Pyrococcus furiosus* extract reveals the presence of at least four potential psiRNA-protein complexes (psiRNPs), each containing distinct classes of CRISPR RNA species. The mature psiRNAs identified here and their associated complexes are leading candidates for the primary agents in the proposed prokaryotic RNAi viral defense system.

Results

psiRNAs cloned from the seven CRISPR loci in Pyrococcus furiosus

The *Pyrococcus furiosus* genome contains seven CRISPR loci, each encoding between 11 and 51, and together encoding 200 potential psiRNAs (Grissa et al., 2007) (Figure 2.1). To investigate whether psiRNAs are produced from the 7 CRISPR loci, we isolated and cloned small RNAs (less than 50 nucleotides) from total *Pyrococcus furiosus* RNA preparations. Among 872 small RNA clones sequenced, 144 (17.3%) were derived from CRISPR loci. In addition, 42.2% corresponded to rRNA, 23.9% were derived from ORFs, and 12.4% were from sRNAs (snoRNA homologs). The remaining 4.2% of sequences were derived from tRNAs, transposons, Hhc RNAs (Klein et al., 2002) and intergenic sequences.

Most of the CRISPR clones consisted primarily of psiRNA (variable) sequence and included some flanking repeat sequence. The clones included 64 of the 200 potential *Pyrococcus furiosus* psiRNAs and represented all seven CRISPR loci (Figure 2.1). We have adopted a simple system of nomenclature for psiRNAs, in which the psiRNA is designated by a 3-digit number. The first digit indicates the locus number (1, 2 and 4 - 8 in *Pyrococcus furiosus*) and the second two digits, separated from the first by a decimal point, indicate the position of the psiRNA within that locus (relative to the leader). For example, the first psiRNA in CRISPR locus 1 is Pf psiRNA 1.01, the first psiRNA in CRISPR locus 2 is psiRNA 2.01, and the last psiRNA in CRISPR locus 1 is psiRNA 1.51 (see Figure 2.1). Figure 2.1 shows the number of times each individual psiRNA was cloned. The psiRNA clones ranged between 17 and 50 nucleotides in length (see Table 2.1 for psiRNA clone sequences). The clones included variable

amounts of the psiRNA sequence (12 to 40 nucleotides) and of repeat sequence at the 5' (0 to 8 nucleotides) and/or 3' (0 to 22 nucleotides) end. In addition to the psiRNA clones, we isolated a few CRISPR-derived clones that lacked psiRNA sequence and consisted of a portion of leader sequence upstream of a repeat (See "Leader" section of Table 2.1), indicating that the 3' end of the CRISPR leader is also transcribed.

Our sampling is not apparently at saturation, however, we cloned psiRNAs from the beginning, middle and end of CRISPR loci (Figure 2.1), indicating that RNAs are produced from across the length of the loci. Interestingly, however, the likelihood of cloning was significantly higher for psiRNAs encoded within the first part of a CRISPR locus, suggesting a greater abundance in the organism of psiRNAs from these regions. Two-thirds of the psiRNAs that we cloned were from the first third of their CRISPR locus and 45% were one of the first four psiRNAs in a given locus (Figure 2.1). With the exception of locus 2, this trend was observed within each individual CRISPR locus.

Comparison of the percentage of psiRNAs cloned from a given locus (Table 2.2, % of clones) to the percentage of the total psiRNAs encoded by that locus (Table 2.2, % of psiRNAs) revealed that most of the loci are represented proportionately within the clones. However, locus 6 seems to be significantly underrepresented in the cDNA library. Locus 6 encodes ~22% of the psiRNAs in *Pyrococcus furiosus*, however only ~12% of the cloned RNAs were derived from this locus. This suggests that the psiRNAs encoded within locus 6 are less abundant in *Pyrococcus furiosus* than those encoded by the six other CRISPR loci.

The results of the RNA cloning suggest the presence of novel, small psiRNAs in *Pyrococcus furiosus*. However, the cloned RNAs were not of a uniform size or

composition. To determine whether discrete psiRNA species are present in *Pyrococcus furiosus*, we undertook additional analysis.

Northern analysis of RNAs derived from the CRISPR loci in Pyrococcus furiosus

In order to further investigate the RNAs that arise from the CRISPR loci in *Pyrococcus furiosus*, we undertook Northern analysis with probes against both repeat and psiRNA sequences. Probes were designed for detection of both sense (transcription from the leader sequence) and antisense RNAs. Figure 2.2A shows results obtained with a probe that recognizes the repeat sequence (sense orientation) that is common to *Pyrococcus furiosus* CRISPR loci 1, 5 and 6. This probe detected a prominent band at ~65 nucleotides, a less prominent band at ~130 nucleotides, and an unresolved set of bands of greater than 150 nucleotides near the top of the gel. For this and all other probes tested, no significant differences in the patterns were observed from total RNA samples prepared with and without DNase treatment indicating that the bands represent RNAs. Consistent with the observations and CRISPR RNA processing pathway proposed by others (Lillestol et al., 2006; Makarova et al., 2006; Sorek et al., 2008; Tang et al., 2002; Tang et al., 2005), the set of bands above 150 nucleotides likely represents a mixture of primary transcripts from the 3 loci as well as larger intermediates generated by cleavages within repeat regions. The most prominent band detected with the repeat probe in *Pyrococcus furiosus* (~65 nucleotides) corresponds well to the primary product of the CRISPR loci reported previously in other organisms (Lillestol et al., 2006; Tang et al., 2002; Tang et al., 2005), however in this work this RNA is identified as the “1X intermediate” (Figure 2.2). This band corresponds in length to a psiRNA (~35-40 nts) and repeat (~30 nts), and likely represents psiRNAs with

flanking repeat sequences generated by cleavages within the adjacent repeats (see Figure 2.2C). The detection of this RNA by the repeat probe suggests that cleavage may be asymmetric within the repeat sequence, leaving a substantial contiguous region of the 30 nucleotide repeat on one side (e.g. the 3' end as modeled in Figure 2.2C) for efficient detection by Northern probes. A less abundant band of ~130 nucleotides corresponds in length to two psiRNAs with flanking repeat sequences and likely represents the immediately upstream 2X intermediates (Lillestol et al., 2006; Tang et al., 2002; Tang et al., 2005) (Figure 2.2C).

Northern analysis with a probe against one of the variable psiRNA sequences revealed novel CRISPR RNA species. Using a probe against psiRNA 4.02 (sense orientation), we detected a band at ~60 nucleotides and a very faint signal near the top of the gel, but the most prominent band is ~46 nucleotides (Figure 2.2B) and corresponds in size to that of the psiRNA (35 nts in the case of psiRNA 4.02) and ~10 nucleotides of repeat sequence. A significant secondary band was detected at ~39 nucleotides (Figure 2.2B).

Importantly, similar results were observed in Northern analysis of RNAs from other CRISPR loci. Results for all RNAs analyzed, both sense and antisense, are compiled in Figure 2.3. First, like the repeat 1, 5, 6 probe, a probe for the repeat sequence common to loci 2, 4 and 7 (sense orientation) detected prominent diffuse bands of ~65 and ~130 nucleotides (theoretical 1X and 2X processing intermediates, see Figure 2.2C). In addition, we probed for psiRNAs from the first part of each CRISPR locus (1.01, 2.01, 4.02, 5.02, 6.01, 7.01 and 8.01) as well as for psiRNA sequences from the middle and end of locus 7 (7.11 and 7.21). Strikingly, probes for

each of the psiRNA sequences (sense orientation) detected a single predominant RNA species (indicated with dots in Figure 2.3). Most of these predominant RNAs were ~43 to ~46 nucleotides. The observed size of the major RNA species was generally 5 to 10 nucleotides longer than the encoded psiRNA sequence. The psiRNA with the longest observed primary product (psiRNA 1.01) has an unusually long psiRNA sequence. These findings, together with the observation that these RNAs are recognized by psiRNA but not repeat sequence probes, suggest that the primary mature psiRNA species in *Pyrococcus furiosus* consists of a psiRNA with ~5-10 nucleotides of repeat.

In addition to the primary psiRNA species, each of the psiRNA probes detected other RNAs. These often included an RNA close to the size of the ~65 nucleotide 1X intermediates that were detected by repeat probes, and in some cases (e.g. psiRNAs 101 and 402) an RNA was detected that was the size of the theoretical 2X intermediate. Many of the psiRNA probes detected other faint bands. However, in many cases the most prevalent secondary species was a slightly smaller RNA of ~38 to ~45 nucleotides.

We did not detect antisense RNAs with most of the CRISPR probes (Figure 2.3B). Prominent bands were detected with probes from psiRNAs 2.01 and 7.11, however the absence of corresponding bands with the repeat probes suggests that these are not CRISPR locus-derived RNAs.

Northern analysis of CRISPR RNA distribution in fractionated Pyrococcus furiosus extract

The CRISPR RNAs are hypothesized to function in complex with proteins in various aspects of RNA-guided genome defense in prokaryotes (Makarova et al., 2006).

To assess whether the CRISPR-derived RNA species that we identified may be components of complexes, we analyzed the distribution of the RNAs across fractions from anion exchange chromatography of *Pyrococcus furiosus* S100 cell extract performed under anaerobic conditions. Fractions were evaluated by Northern analysis using probes against the repeat sequence common to loci 1, 5 and 6, and psiRNA 4.02 and 7.01 sequences (Figure 2.4). For reference, the profile of RNAs detected in unfractionated extract is shown in the first lane.

The distribution of the various RNAs across the fractions suggests the presence of several distinct CRISPR RNA-containing complexes. The novel primary and secondary psiRNAs (~45 and ~39 nts) from both loci co-fractionated in a distinct set of fractions denoted as peak A (Figure 2.4, fractions 10-14). Other larger CRISPR RNA species were not observed in peak A. The primary mature psiRNA is also present in peak B (fractions 20-23) along with a fraction of the 1X intermediate RNA and some of the variable-size psiRNA species. However, the highest concentration of the 1X intermediate is found in a distinct set of fractions that lack small psiRNAs, termed peak C (fractions 23-26). The mixture of RNAs that likely includes full CRISPR locus transcripts and larger intermediates, and the 2X intermediate are found primarily in peak D (fractions 31-34). These results suggest the presence of multiple complexes each containing distinct subsets of CRISPR-derived RNAs in *Pyrococcus furiosus*.

To verify the presence of CRISPR RNA-protein complexes in the fractionated *Pyrococcus furiosus* extract, we examined the complexes on native gels. Using the probe for psiRNA 7.01, we compared the mobility of the RNAs, both in the presence of the co-fractionating proteins in peaks A, B and C and following protein extraction, by

non-denaturing PAGE and Northern analysis. In each case, a significant shift in the mobility of the RNAs was observed in the presence of the proteins. Peak D was not examined on native gels. Together our results indicate that the novel primary mature psiRNA is a component of at least 2 chromatographically distinct RNA-protein complexes (peaks A and B), the 1X intermediate is found primarily in a third complex (peak C), and a fourth complex includes larger psiRNA precursors. These complexes are likely candidates for the mediators of psiRNA production, invader destruction and CRISPR element integration in the proposed prokaryotic RNAi pathway.

Discussion

Novel CRISPR RNAs.

The CRISPR loci found in many prokaryotes encode alternating repeat and "spacer" or psiRNA sequences, and have been shown to give rise to a series of RNAs that decrease in increments from the full-length locus transcript to a single psiRNA and repeat (i.e. 1X intermediate, see Figure 2.2C) (Lillestol et al., 2006; Tang et al., 2002; Tang et al., 2005). Current evidence indicates that processing occurs by endonucleolytic cleavages within the repeat sequences(Tang et al., 2002; Tang et al., 2005). The RNA products of the CRISPR loci are hypothesized to guide silencing of viruses and other genome invaders. In this work, we have identified a novel class of smaller discrete CRISPR-derived RNAs that we have termed psiRNAs, which appear to be the ultimate gene products of the CRISPR loci (Figure 2.2C).

The mature psiRNA species that we have identified is the most abundant CRISPR-derived RNA detected in *Pyrococcus furiosus*. The primary mature psiRNAs are approximately 5 to 10 nucleotides longer than the corresponding psiRNA sequence

(i.e. approximately 45 nucleotides long). These RNAs are shorter than the smallest discrete CRISPR RNA products previously reported (i.e. the ~60-65 nt 1X intermediate species) (Lillestol et al., 2006; Tang et al., 2002; Tang et al., 2005), and are presumably generated by exonucleolytic processing of the 1X intermediate. Our Northern and sequence analysis indicates that these RNAs are comprised primarily of psiRNA sequence and do not contain substantial repeat sequence. A secondary psiRNA of about 39 nucleotides was also consistently observed among psiRNA profiles.

CRISPR RNA-protein complexes.

The common primary and secondary psiRNA species are likely candidates for the guide RNA component of the effector complex in the proposed pRNAi genome defense pathway. Both of these psiRNA species, but not larger intermediate CRISPR RNAs, are found in RNA-protein complexes in anion exchange chromatography peak A (Figure 2.4 A and B), thus peak A could contain the effector complex. The primary mature psiRNA is even more abundant in peak B, which contains relatively less of the secondary psiRNA but also contains some 1X intermediate RNA (Figure 2.4). Peak B may also contain the effector complex and/or a complex involved in the exonucleolytic processing of the 1X intermediate to psiRNAs. The 1X intermediate RNA is most abundant in peak C, which is adjacent to peak B and processing may occur across peak B and C fractions. Longer CRISPR RNAs are found in peak D. Our results indicate that the various RNA species are components of distinct RNA-protein complexes in *Pyrococcus furiosus*. Extensive purification and analysis will determine whether these hypothesized activities and the Cas proteins predicted to function in CRISPR RNA

biogenesis and invader silencing (e.g. RNA binding proteins and nucleases) are present in these complexes.

psiRNA expression.

The psiRNAs are hypothesized to act in a manner similar to the antibodies of the human immune system and expression would be expected even in the absence of active infection to patrol for returning invaders. Our results indicate that psiRNAs are actively produced from all 7 CRISPR loci in *Pyrococcus furiosus*. Moreover, expression levels appear to be equivalent between the loci under the growth conditions examined with the possible exception of one locus that yielded 50% fewer psiRNA clones than expected (see locus 6, Table 2.2). Our results confirm that CRISPR RNAs are transcribed from the leader sequence in *Pyrococcus furiosus* and indicate that a portion of the leader sequence is also transcribed.

Interestingly, we found evidence of significantly higher levels of expression of psiRNAs encoded proximal to the leader of a CRISPR locus. Current data indicate that these are the most recently acquired psiRNA sequences within CRISPR loci (Barrangou et al., 2007; Pourcel et al., 2005). Increased levels of these psiRNAs may be important for targeting current invaders. Distal psiRNAs may be produced at lower levels to provide surveillance for past invaders. It is not clear whether the increased level of leader-proximal psiRNAs results from differences in RNA transcription (e.g. partial transcription of the loci), processing, stability or other factors. This is an important new aspect of understanding the regulation of psiRNA expression that remains to be explored.

Methods and Materials

Small RNA cloning and sequencing.

Total RNA was isolated from *Pyrococcus furiosus* cells using the Trizol reagent (Invitrogen) as indicated by the manufacturer. Approximately 300 μg of total RNA was separated on an 8 x 8.5 cm 15% polyacrylamide 7M urea gel using DNA size standards (pGEM markers, Promega) for size determination. RNAs between 20 and 50 nucleotides were isolated and passively eluted overnight in 0.5 M ammonium acetate, 0.1% SDS, 0.5 mM EDTA, followed by ethanol precipitation. In order to remove potential 5' triphosphates or cap structures, RNAs were treated with 50 U tobacco acid pyrophosphatase (TAP) (Epicentre) for 2 hours at 37°C. The eluted RNAs were cloned using standard microRNA cloning protocols (Lau et al., 2001) using the following primers containing EcoRI restriction sites: 3' adapter : 5'–AppTTTAACCGCGAATTCCAGddC–3' (IDT), 5' adapter: 5'–ACGGAATTCCTCACTrArArA–3' (IDT), RT/PCR primer: 5'–GACTAGCTGGAATTCGCGGTTAAA–3 (IDT)', PCR primer: 5'–CAGCCAACGGAATTCCTCACTAAA–3'. An additional PCR was performed in order to add a BanI restriction site (GGYRCC) for concatamerization of the PCR products using the following primers: PCR2 5' primer 5'–GACTAGCTTGGTGCCGAATTCGCGGTTAAA–3', PCR2 3' primer 5'–GAGCCAACAGGCACCGAATTCCTCACTAAA–3'. The products were subject to restriction digestion and DNA ligation by standard methods (Lau et al., 2001). cDNAs were cloned into the pCRII TOPO vector (Invitrogen) and transformed into TOP10 cells (Invitrogen) as described by the manufacturer. Plasmid preparation and sequencing

was performed in a 96-well plate format using standard M13 forward, reverse, and T7 promoter primers. Sequences were analyzed using BLAST (NCBI).

Northern analysis.

Approximately 10 μg of *Pyrococcus furiosus* total RNA was separated on 15% polyacrylamide 7 M urea gels (Criterion, Bio-Rad) alongside [^{32}P]-5'-end radiolabeled RNA markers (Decade, Ambion). The RNAs were transferred onto nylon membranes (Zeta-Probe, Bio-Rad) using a Trans-Blot SD Semi-Dry Cell (Bio-Rad). Membranes were baked at 80°C for at least an hour before pre-hybridization in a ProBlot hybrdization oven (LabNet) for at least 1 hour at 42°C. Pre-hybridization and hybridization was performed in either Oligo-UltraHyb (Ambion) buffer or hybridization buffer containing 5x SSC, 7% SDS, 20 mM sodium phosphate, pH 7.0 and 1x Denhardt's solution. Deoxyribonucleotide probes (MWG) (20 pmol) were 5' end labeled with T4 Polynucleotide Kinase (Ambion) and γ-[^{32}P]-ATP (specific activity > 7,000 Ci/mmol, MP Biomedicals) using standard protocols. Labeled probes were added to the pre-hybridization buffer, followed by hybridization overnight at 42°C. Following hybridization, two washes were performed in 2X SSC, 0.5% SDS for 30 minutes at 42°C. Resulting blots were exposed to a phosphoimager screen for 24-72 hours and scanned.

Probe	Sequence
repeat 1,5,6 (antisense)	CTTTCAATTCTATTTT(AG)GTCTTATTC(GT)AAC
repeat 1,5,6 (sense)	GTT(AC)CAATAAGAC(TC)AAAATAGAATTGAAAG
repeat 2,4,7 (antisense)	CTTTCAATTCTTTTGTAGTCTTATTGGAAC

repeat 2,4,7 (sense)	GTTCCAATAAGACTACAAAAGAATTGAAAG
psiRNA 1.01 (antisense)	GGTCAGATCAGATTGCTTAAGACAAGAAATG
psiRNA 1.01 (sense)	CATTTCTTGTCTTAAGCAATCTGATCTGACC
psiRNA 2.01 (antisense)	GTGGAGCAGAGTCAGAAGAAGAAGTGCG
psiRNA 2.01 (sense)	CGCACTTCTTCTTCTGACTCTGCTCCAC
psiRNA 4,02 (antisense)	TCTGATAGGCTTCAAAGAGTGGCGCTTCAAC
psiRNA 4.02 (sense)	GTTGAAGCGCCACTCTTTGAAGCCTATCAGA
psiRNA 5.02 (antisense)	GGGAATGGTTCACGTAGTACTTGAGGGCGC
psiRNA 5.02 (sense)	GCGCCCTCAAGTACTACGTGAACCATTCCC
psiRNA 6.01 (antisense)	CTAAGGACATTTGTACGTCAAATTCTTCAC
psiRNA 6.01 (sense)	GTGAAGAATTTGACGTACAAATGTCCTTAG
psiRNA 7.01 (antisense)	GCTCTCAGCCGCAAGGACCGCATAC
psiRNA 7.01 (sense)	GTATGCGGTCCTTGCGGCTGAGAGC
psiRNA 7.11 (antisense)	CCTTATATGGGTGTTGTGAAGCAGGATAGAAC
psiRNA 7.11 (sense)	GTTCTATCCTGCTTCACAACACCCATATAAGG
psiRNA 7.21 (antisense)	GGCTCTACCTAATCATCCTCTTGACACAAC
psiRNA 7.21 (sense)	GTTGTGTCAAGAGGATGATTAGGTAGAGCC
psiRNA 8.01 (antisense)	GACTGTGTGTGGAGCAGCTATTTGCTTCGGC
psiRNA 8.01 (sense)	GCCGAAGCAAATAGCTGCTCCACACACAGTC

Chromatography.

15 g of *Pyrococcus furiosus* cells were lysed anerobically in 200 mL 50 mM Tris, pH 8.0 in the presence of 4 mg/L RNase-free DNase (Sigma). The extract was subject

to ultracentrifugation at 113,000 x g for 2 hours (Optima L-90K, Beckman-Coulter). The resulting S100 extract was applied to a 60 mL DEAE Sepharose-FF column and eluted using a 0-500 mM NaCl gradient. The resulting fractions were analyzed by isolating RNAs from 250 μl of each 30 mL fraction using the Trizol LS protocol (Invitrogen). The RNAs were separated on 10% polyacrylamide 7 M urea gels, blotted, and subject to Northern analysis as described above, using the Oligo-UltraHyb hybridization buffer for both pre-hybridization and hybridization. See above for probe sequences.

***Native Northern analysis*.**

40 ul of DEAE fractions from peaks A-C were separated on a 4-20% polyacrylamide gel (Bio-Rad) using SDS-free running buffer (25 mM Tris, 19.2 mM glycine). De-proteininzed samples were analyzed in parallel. Gels were run at 50 V for 3-4 hours at room temperature. The gel was soaked in 5M urea, 45 mM Tris, 45 mM boric acid, 1 mM EDTA for 15 minutes, then subject to blotting and Northern analysis as described above, using Ultra-Hyb Oligo hybridization buffer (Ambion) and a probe against psiRNA 7.01.

Acknowledgements

We thank Lindsay Jones for assistance with Northern analysis, Gerti Schut and Mike Adams (University of Georgia) for *Pyrococcus furiosus* material and technical assistance with chromatography, and Alex Huttenhofer (University of Innsbruck) for encouragement. This work was supported by NIH R01-GM54682 (M.T. and R.T.).

References

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

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

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

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

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

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

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

Hannon GJ. 2002. RNA interference. *Nature 418*:244-251.

Hartig JV, Tomari Y, Forstemann K. 2007. piRNAs--the ancient hunters of genome invaders. *Genes Dev 21*:1707-1713.

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

Jansen R, Embden JD, Gaastra W, Schouls LM. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol 43*:1565-1575.

Kim VN. 2006. Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev 20*:1993-1997.

Klein RJ, Misulovin Z, Eddy SR. 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A 99*:7542-7547.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science 294*:858-862.

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

Lin H. 2007. piRNAs in the germ line. *Science 316*:397.

Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV. 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res 30*:482-496.

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

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

Nishida KM, Siomi MC. 2006. [Molecular mechanisms of RNA silencing by siRNA, miRNA and piRNA]. *Tanpakushitsu Kakusan Koso 51*:2450-2455.

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

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

Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, 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 TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Huttenhofer A. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol 55*:469-481.

Tyson GW, Banfield JF. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol 10*:200-207.

Figure 2.1. ***Pyrococcus furiosus CRISPR loci and distribution of cloned psiRNAs.*** The seven *Pyrococcus furiosus* CRISPR loci are illustrated. The psiRNA numbers associated with each locus are indicated above. Each psiRNA is represented by a box. Shaded psiRNAs were cloned at least once in this work and the number of clones isolated is indicated below the shaded box. The genome coordinates of each *Pyrococcus furiosus* CRISPR locus are as follows: 1: 27091-30618; 2: 260714-262113; 4: 312405-313931; 5: 623119-625176; 6: 695937-698992; 7: 1064076-1065543; 8: 1091089-1091857.

CRISPR Locus 1; psiRNAs 1.01 - 1.51

CRISPR Locus 2; psiRNAs 2.01 - 2.20

CRISPR Locus 4; psiRNAs 4.01 - 4.22

CRISPR Locus 5; psiRNAs 5.01 - 5.30

CRISPR Locus 6; psiRNAs 6.01 - 6.45

CRISPR Locus 7; psiRNAs 7.01 - 7.21

CRISPR Locus 8; psiRNAs 8.01 - 8.11

Table 2.1. **Cloned psiRNA sequences.** Repeat sequences are indicated in bold.

	Sequence	psiRNA
Locus 1	CTGATCTGACCAGAGCTG**GTTCCAATAAGACTAAA**	1.01
	CTGATCTGACCAGAGCTG**GTTCCAAGTAAGACTAA**	1.01
	CTGATCTGACCAGAGCTG**GTTCCAATAAGACTAAA**	1.01
	CAATCTGATCTGACCAGAGCTG**GTTCCAAT**	1.01
	CAATCTGATCTGACCAGAGCTG**GTTCCAAT**	1.01
	ATGATTCATTTCTTGTCTTAAGCAAT	1.01
	TTGTCTTAAGCAATCTGATCT	1.01
	CACTAAAGTCATACTTTACTGCTACAACCCGCTCTGG	1.04
	GTCATACTTTACTGCTACAACCCGCTCTGG	1.04
	TACTGCTACAACCCGCTCTGGGTCGAG	1.04
	TTACTGCTACAACCCGCTCTGGGTCGA	1.04
	ACTGCTACAACCCGCTCTGGGTTGA	1.04
	CTGACACGAACATAAACA**GTTCCAATAAGACTACAGAAGA**	1.06
	CTGACACGAACATAAACA**GTTCCAATAAGACTACAGAAGA**	1.06
	GAAAGGGAAATGTGCGTAAAGGTTTTCTTCCC	1.07
	GAAAGGGAAATGTGCGTAAAGGTTTTCTTCCC	1.07
	TTGACCCACCACCAGCCCT**GTTCCAATAAGAC**	1.08
	GAAAGGCGTGCCGTGTGTTTTTATAA	1.11
	GAAAGGCGTGCCGTGTGTTTTTATAA	1.11
	GTTGCTGCATATCCAGTGTGG	1.12
	GTTGCTGCATATCCAGTGTGG	1.12
	GGCAAGTTCTGGCCTATACTGTCTCCTAATGTCT	1.13
	TAGGAGACAGTATAGGCCAGAACTTGCCCAG	1.13
	GACATTAGCNGACAGTATAGGCCA	1.13
	TTAGCAAATTGCCGATTACTGCACATAAAAAAAATA**G**	1.14
	CTATAAGGGATTGAAAGGTCAAAGGTATAN	1.16
	CTATAANGGATTGAAAGGTCAAGGGTATACT	1.16
	TTCGCGGTTAAACAATCTGATCTGACCAGAGCTG**GTTCCAAT**	1.19
	GGACAGCGTGGACACGGTGAACGGGCTCTGGA	1.21
	CTGATAGAACCTTTGCCACC	1.24
	CTGATAGAACCTTTGCCACC	1.24
	CATACTTGCGGATACGGATCCAGTCAAAACTTGACT**G**	1.26
	TTGCGGATATGGATCCAGTCAAAACTTGACT**G**	1.26
	GCGGATACGGATCCAGTCAAAACTTGACT**G**	1.26
	GAAAGCATACTTGCGGATACGGATCCAGT	1.26
	CTCTGGGTCGTCTATGTTTTTGA	1.27
	GAGTAGAAATGCCCAAATTCCCCTTAGGGACA	1.36

	TTTGTGATAGTGTTCTTTGCAACGAAGTGCTTGCTGGTCA**G**	1.43
	GTTTGTGATAGTGTTCTTTGCAACGAAGAGCTTGCTGG	1.43
	GAGTGCCCCGAGCCGGGGGCT	1.49
Locus 2	CTGACACGAACATAAACA**GTTCCAATAAGACTACAGAAGA**	2.02
	CTGACACGAACATAAACA**GTTCCAATAAGACTACAGAAGA**	2.02
	ATGGCTCGATGGAATTAT**GTTCCAATAAGACTACAAAAG**	2.03
	ATGGCTCGATGGAATTAT**GTTCCAATAAGACTACAAAAG**	2.03
	CTAACTAACATCACCAATAATTAATTGTAAGTTAG	2.10
	GCTACCATGGCCATCACCAATAATTAATTGTAAGT	2.10
	CTGAGCCAACCCACCACTTTGGTAAAACT	2.13
	CTGAGCCAACCCACCACTTTGGTAAAACT	2.13
	CTGAGCCAACCCACCACTTTGGTAAAACT	2.13
	TGAGGCTGGAGAGGGCTTCTTTGTTACTACTTGCGT	2.17
	TTATGTTCATGTTCCACATCTAA	2.18
	TATGTTCATGTTCCACACTA	2.18
Locus 4	**TTGAAAG**GAATGTTGCTCAATGCAAAGGGCTCACCGCTGCTG GT**GTTCCA**	4.01
	CTCAATGCAAAGGGCTCACCGCTGCTGGT**GTTCCAATAAGA**	4.01
	CTCACCGCTGCTGGT**GTTCCAATAAGACTACAAAAGA**	4.01
	TTGAAAGTTGAGTTGAAGCGCCACTCTTTGAA	4.02
	TTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA**G**	4.02
	ATTGAAAGTTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA	4.02
	AGTTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA**GT**	4.02
	GTTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA**GT**	4.02
	TTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA**GTT**	4.02
	AAGTCGGGTCCCTTGGAGTTCCGAACGGGCTCCCGAGGCT**GT TCCA**	4.04
	GGGCTCCCGAGGCT**GTTCCAATAAGAC**	4.04
	GTTGATTCCCTTATAGATGTTCGTTTTCCACA	4.05
	ATGTTCGTTCTCGTTCACTGTTATTCTCTT	4.07
	CTCGTTCACTGTTATTCTCTTT	4.07
	AAAACTAAAAAAAGAAGAGGTGGTGGTGAAGAAT	4.08
	GAAAGTCTCAATTGGGGAGTTGCTTTAATGGCTTTT	4.12
	TCAATCCGAGAATCGAATTTTCCTATACGCTTTT**GTT**	4.21
	TTTGTTTTTGCTCCTGTGTCTTGTGGTGATAAAAT**G**	4.22
	TTTGTTTTTGCTCCTGTGTCTTGTGGTGATAAAAT**G**	4.22
	GTGATAAAAT**GTTACAATAAGACTACAAAAG**	4.22
Locus 5	**ATTGAAAG**GACCATACTCACCAGCAGCGGTGAGCCCTTTGCA TTGA	5.01
	ATTGAAAGGACCATACTCACCAGCAG	5.01
	GTTCACGTAGTACTTGAGGGCGCTCAC**GTTACAATAAGACCA**	5.02

	TTTCANGCAGTACTTGAGGGCGCTCAT**GTTNCANTANGACCAA**	5.02
	AAGAAGGGGAATGGTTCACGTAGCTACTTGAGGGC	5.02
	CAATAATACAGTCCTAATGCTCGT**G**	5.03
	CAATAATACAGTCCTAATGCTCGT**G**	5.03
	TTGAAAACGCTAGCAGGACTAGTGCTTGT	5.04
	CGCTAGCAGGACTAGTGCTTGTG	5.04
	CTTCTCGAATCTATCGAATTCG**GTTACAATAAGACCAAAATAG A**	5.09
	AGCCACATAANACATTGTCATACAAAGTATGACAAAATA	5.11
	CACATAAGACATTGTCATACAAAGTAGGACAAAA	5.11
	AAGACATTGTCATACAAAGTAGGACAAA	5.11
	GTCCTCTTGGAGACCGTTCCT**GTTACAATAAGACCA**	5.12
	GTCACGTAATTCGCCAAGTCCNCNT	5.12
	AATAGTTACAATAAGACCAAAATA	5.15
	CTAGCTTTTCACACACTCT	5.18
	TAAACTANGNTGATTTTGTAAT	5.20
	GAAAGAGTATTCCACCGAGAATTGTGCC	5.29
	GTATTCCACCGAGAATTGTGCCTTTGTACTGGACTG	5.29
Locus 6	TCTATTTTAGTCTTATTGTAAC**GTTCCACTAAGGAC**	6.01
	TTTAGTCTTATTGTAAC**GTTCCACTAAGGAC**	6.01
	AATTTGACGTACAAATGTCCTTAGTGGAAC	6.01
	TTCGGGACCTGTAGGTC**GTTACAATAAGACTAAAATAGA**	6.02
	GTTAATGGTAAA**GTTACAATAAGACTAAA**	6.03
	GTTCTGCCGTCCCTTTTCTCGACG	6.09
	TTCTGCCGTCCCTTTTCTCGACGAACCTCATACCGA	6.09
	TTCTGCCGTCCCTTTTCTCGACGAAC	6.09
	TTCTGCCGTCCCTTTTCTCGACGAAC	6.09
	TATAGGCGGAACTCCCT	6.13
	AAGTGTTTTCGAATATTGTTACTTCTTGTGT	6.15
	CTATAAGACTGAAACTTCACACCT	6.37
	TTAACACTCTTAACCCCA**G**	6.38
	GTCCAAAAAC**GTTACAATAAGACTAAA**	6.39
	TTAAGCTGGGATGGGCTATATACAAAGACA**G**	6.42
	AATTCTGGAAGGTTGTAGAAA	6.44
Locus 7	CGCCCACCTTTGTTACGTTCCAATAAGACT	7.01
	TTGTAGTATGCGGTCCTTGCGGCTGAGAGCACTTCAG	7.01
	TTGTAGTATGCGGTCCTTGCGGCTGAGAGCA	7.01
	TTGTAGTATGCGGTCCTTGCGGCTGAGAGCA	7.01
	GTAGTATGCGGTCCTTGCGGCTGAGAGCA	7.01
	GAAAGTTGTAGTATGCGGTCCTTGC	7.01

	GTCTTCGATTAGTGAAAACA**GTTCCAATAAGACTACAAAAG**	7.02
	GTCGTTATCTCTTACGAAGTCTTCGATTAGT	7.02
	GTTACACGTGAGTGCAA**GNTCCAATAAGACTACAAAAGA**	7.04
	GTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	GTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	GTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	TTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	TTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	TTTACACGTGAGTGCAA**GTTCCAATAAGACTACAAAAGA**	7.04
	ACAAAAGAATTGAAAGTTAACCTCCTT	7.07
	AGTTATCTAAGCTCTGCTTAAATGGGAAAATCTTATAAG	7.14
Locus 8	**GAAG**AGGAAGAAATGCAGACGACGTGATAAACTACGTGAA	8.02
	CAGACGACGTGATAAACTACGTGAAAA**GTT**	8.02
	AACTTTTCAACGTAGTTTATCACGTCGTCTGA	8.02
	GTGCACTAAGGCACCATACGCCCAA	8.03
	ATTGAAGCTTGCCCAACCTCTCTAGAAACGCCCA	8.06
	AATTGAAGNNAAAATCTCTTTTTAAATCTTTGA	8.07
	ATTGAAGCGCANATCTNTTTTTAAATCTTTGA	8.07
Leaders	GTAGGAGTATTGGGGCAAAAAAGCCCCCT**GTTCCAATAAGAC**	before 2 or 7
	GGGGGGAATTGGGGCAAAAAAGCCCCCT**GTTCCAATAAGACT**	before 2 or 7
	GGGGGAATTGGGGCAAAAAAGCCCCCT**GTTCCAATAAGACTA C**	before 2 or 7
	GGGGGAATTGGGGCAAAAAAGCCCCCT**GTTCCAATAAGACTA C**	before 2 or 7
	CCCCCT**GTTCCAATAAGACTACAAAAG**	before 2 or 7
	CCCCCT**GTTCCAATAAGACTACAAAAG**	before 2 or 7
	AAGCCCCCT**GTTCCAATAAGACTACAAA**	before 2 or 7
	GAAAAAGCCCCCT**GTTACAATAAGACCAA**	before 5
	GAAAAAGCCCCCT**GTTACAATAAGACCAAAATAGA**	before 5
	TTAGGAGTATTGGGGCGAAAAAGCCCCCT**GTTACAATAAGAC TA**	before 6
	GGGGGAATTAGGGCAAAAAAGCCCACT**GTTCCAATAAGACT**	before 8
	GGGGAATTAGGGCAAAAAAGCCCACT**GTTCCAATA**	before 8

Table 2.2. Distribution of cloned psiRNAs.

CRISPR locus	# of psiRNAs	# of Clones	% of psiRNAs	% of Clones
1	51	40	25%	30%
2	20	12	10%	9%
4	22	20	11%	15%
5	30	20	15%	15%
6	45	16	22%	12%
7	21	17	10%	13%
8	11	7	5%	5%
Total	200	132		

Figure 2.2. ***Northern analysis of RNAs containing CRISPR repeat and psiRNA 4.02 sequences.*** A and B) Northern blots were performed with 10 μg total RNA using a degenerate oligonucleotide probe designed to detect the repeat sequences for CRISPR loci 1, 5, and 6 (A) or a probe for psiRNA 4.02 sequence (B). Radiolabeled RNA marker sizes are indicated (M). The positions of the initial transcript and large intermediates, 2X intermediate, 1X intermediate and primary mature psiRNA described in the text and Figure 4C are indicated. C) Proposed psiRNA biogenesis pathway. The CRISPR locus is transcribed from a start site within the leader sequence to produce an initial transcript that includes a portion of the leader and the alternating psiRNA and repeat sequences. The initial transcript is cleaved within the repeats to produce intermediates. The endonucleolytic cleavage site may be asymmetrically located within the repeat. The 2X intermediate and 1X intermediates are illustrated. Our results indicate that the 1X intermediate is further processed by an exonuclease to remove most of the repeat sequence, resulting in a primary mature psiRNA species that contains 5-10 nucleotides of the repeat sequence.

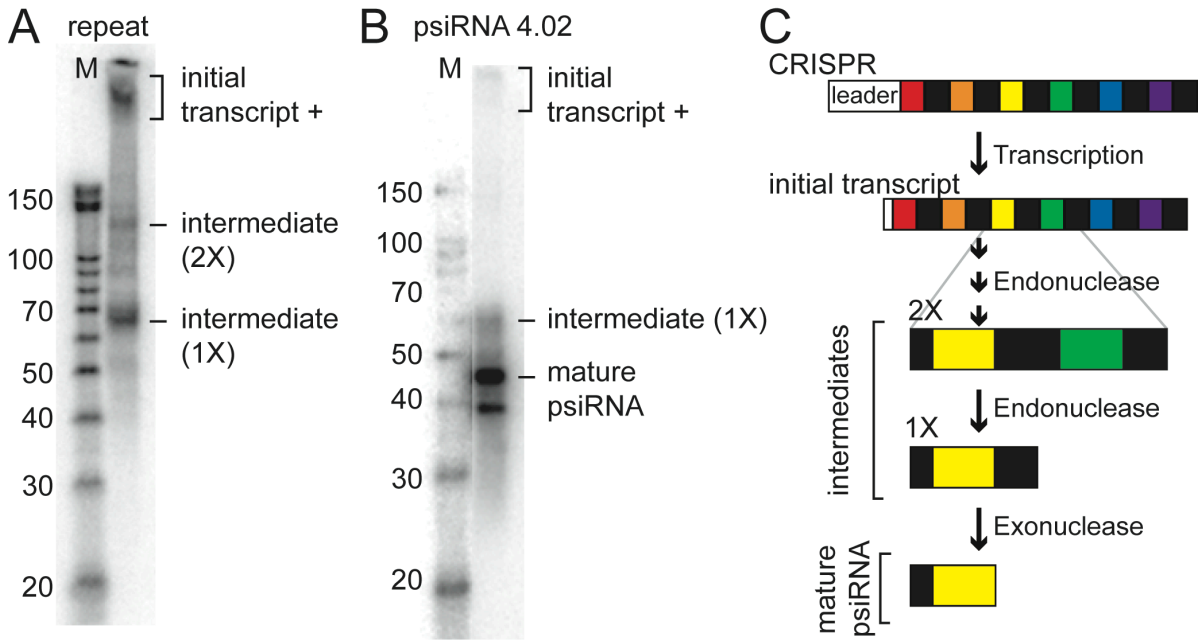

A
repeat
M
initial
transcript +
150
100
70
50
40
30
20
intermediate
(2X)
intermediate
(1X)
B
psiRNA 4.02
M
initial
transcript +
150
100
70
50
40
30
20
intermediate (1X)
mature
psiRNA
C
CRISPR
leader
Transcription
initial transcript
Endonuclease
2X
intermediates
Endonuclease
1X
Exonuclease
mature
psiRNA

Figure 2.3. ***Northern analysis of RNAs from the seven Pyrococcus furiosus CRISPR loci.*** A and B) Northern analysis was performed with 10 μg of total RNA using probes to detect sense (A) (transcribed from the leader sequence) or anti-sense (B) psiRNA or repeat sequence-containing RNAs as indicated. Lanes are approximately aligned on the basis of adjacent marker lanes (not shown except for repeat 1, 5, 6 and psiRNA 8.01 lanes). Dots located to the left of lanes indicate the primary mature psiRNA species. 1X and 2X intermediates detected by repeat probes are indicated.

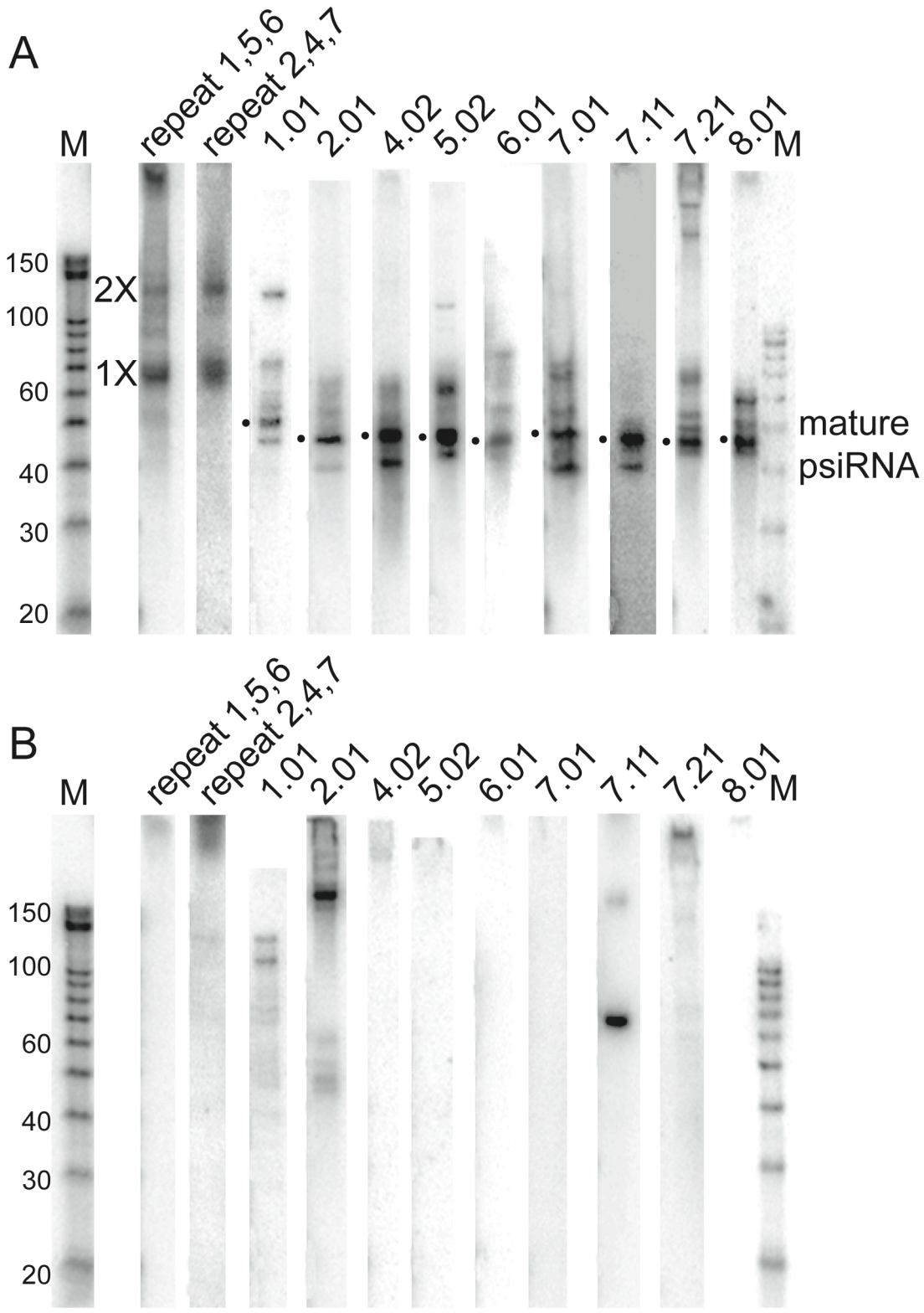
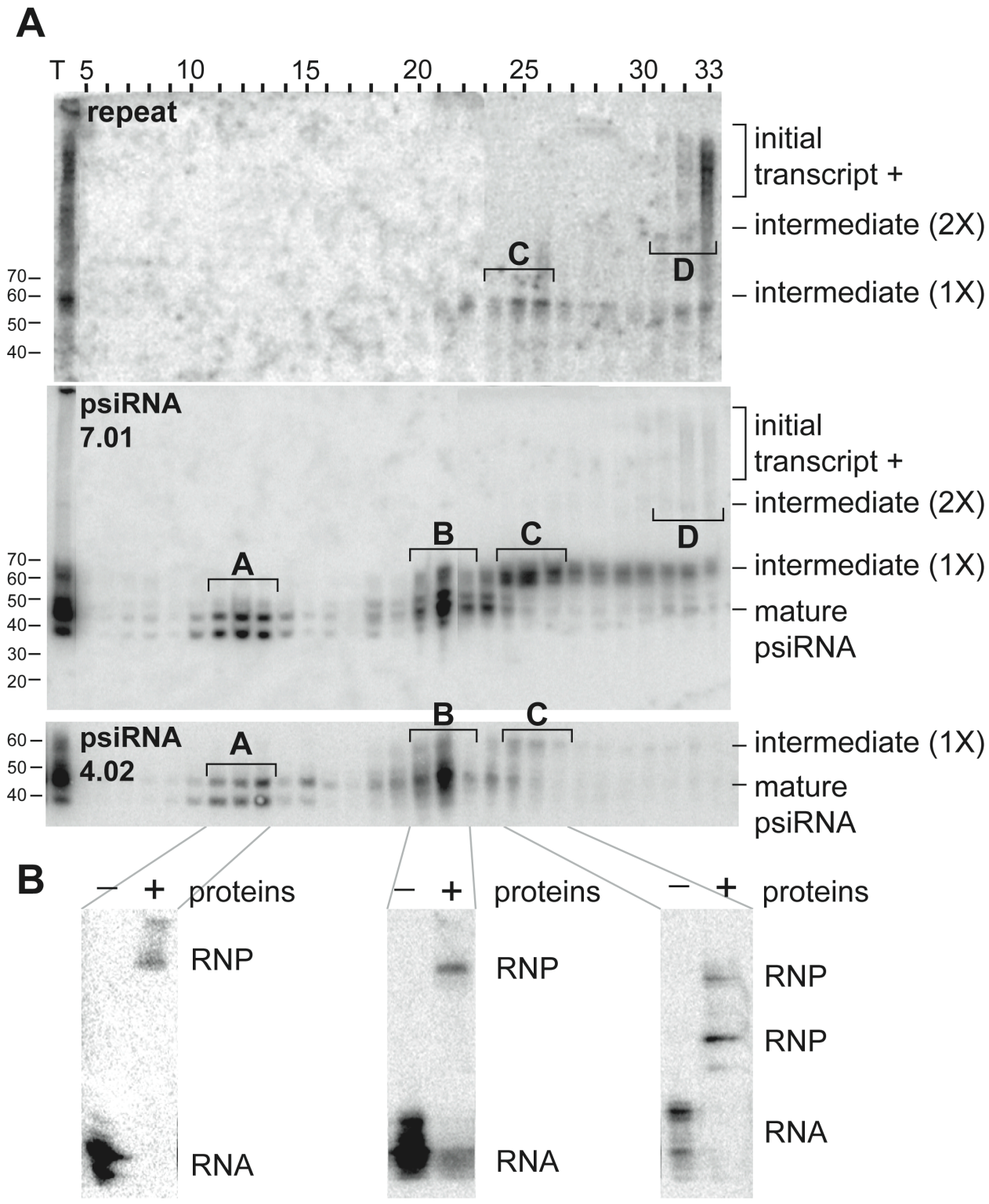

A
M
repeat 1,5,6
repeat 2,4,7
1.01
2.01
4.02
5.02
6.01
7.01
7.11
7.21
8.01
M
150
100
60
40
30
20
2X
1X
mature
psiRNA
B
M
repeat 1,5,6
repeat 2,4,7
1.01
2.01
4.02
5.02
6.01
7.01
7.11
7.21
8.01
M
150
100
60
40
30
20

Figure 2.4. ***CRISPR RNA-protein complexes in fractionated Pyrococcus furiosus cell extract.*** A) *Pyrococcus furiosus* S100 cell extract was separated by DEAE anion exchange chromatography and CRISPR RNAs present in the fractions were examined by Northern analysis using probes against repeat 1, 5, 6, and psiRNAs 7.01 and 4.02 as indicated. Unfractionated extract (T) was co-analyzed for reference. Positions of 1X intermediate, 2X intermediate and primary mature psiRNA, peaks A – D (see text), and markers are indicated. B) Fractions corresponding to peaks A (left), B (center) and C (right) were analyzed by non-denaturing gel electrophoresis and Northern blotting using a probe against psiRNA 7.01. For comparison, proteins were extracted from a portion of each sample and analysis of the RNAs with (+) and without (-) proteins is shown. The positions of the RNAs (- proteins) and potential RNPs (+ proteins) on the native gel are indicated.


A
T 5 10 15 20 25 30 33
repeat
initial transcript +
intermediate (2X)
intermediate (1X)
C
D
70
60
50
40
psiRNA 7.01
initial transcript +
intermediate (2X)
intermediate (1X)
mature psiRNA
A
B
C
D
70
60
50
40
30
20
psiRNA 4.02
intermediate (1X)
mature psiRNA
A
B
C
60
50
40
B
– + proteins
RNP
RNA
– + proteins
RNP
RNA
– + proteins
RNP
RNP
RNA

CHAPTER 3

RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex[1]

[1] Hale, C. R., P. Zhao, et al. (2009). "RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex." Cell **139**(5): 945-56. Reprinted with permission.

Abstract

Compelling evidence indicates that the CRISPR-Cas system protects prokaryotes from viruses and other potential genome invaders. This adaptive prokaryotic immune system arises from the clustered regularly interspaced short palindromic repeats (CRISPRs) found in prokaryotic genomes, which harbor short invader-derived sequences, and the CRISPR-associated (Cas) protein-coding genes. Here we have identified a CRISPR-Cas effector complex that is comprised of small invader-targeting RNAs from the CRISPR loci (termed prokaryotic silencing (psi)RNAs) and the RAMP module (or Cmr) Cas proteins. The psiRNA-Cmr protein complexes cleave complementary target RNAs at a fixed distance from the 3' end of the integral psiRNAs. In *Pyrococcus furiosus*, psiRNAs occur in two size forms that share a common 5' sequence tag but have distinct 3' ends that direct cleavage of a given target RNA at two distinct sites. Our results indicate that prokaryotes possess a unique RNA silencing system that functions by homology-dependent cleavage of invader RNAs.

Introduction

RNAs that arise from the clustered regularly interspaced short palindromic repeats (CRISPRs) found in prokaryotic genomes are hypothesized to guide proteins encoded by CRISPR-associated (*cas*) genes to silence potential genome invaders in prokaryotes (Makarova et al., 2006). CRISPRs consist of multiple copies of a short repeat sequence (typically 25 - 40 nucleotides) separated by similarly-sized variable sequences that are derived from invaders such as viruses and conjugative plasmids (Godde and Bickerton, 2006; Lillestol et al., 2006; Makarova et al., 2006; Mojica et al., 2005; Pourcel et al., 2005; Sorek et al., 2008; Tyson and Banfield, 2008). CRISPR loci

are found in nearly all sequenced archaeal genomes and approximately half of bacterial genomes (Godde and Bickerton, 2006; Haft et al., 2005; Makarova et al., 2006). *cas* genes are strictly found in the genomes of prokaryotes that possess CRISPRs, frequently in operons in close proximity to the CRISPR loci (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). Over 40 *cas* genes have been described, a subset of which is found in any given organism (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). The proteins encoded by the *cas* genes include predicted RNA binding proteins, endo- and exo-nucleases, helicases, and polymerases (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). Recent studies have demonstrated that CRISPRs and *cas* genes function in invader defense in prokaryotes. Exposure of microorganisms that possess the CRISPR-Cas system to a virus results in the appearance of new virus-derived sequences at the leader-proximal end of CRISPR loci in the genomes of surviving individuals (Barrangou et al., 2007; Deveau et al., 2008). Moreover, the acquisition or loss of invader-specific CRISPR elements or of Cas protein genes has been directly correlated with virus and plasmid resistance or sensitivity, respectively (Barrangou et al., 2007; Brouns et al., 2008; Deveau et al., 2008). This rapidly evolving immune system influences the ecology of natural microbial populations (Andersson and Banfield, 2008; Heidelberg et al., 2009; Tyson and Banfield, 2008).

RNAs from the CRISPR loci are hypothesized to guide the CRISPR-Cas defense response based on their potential to base pair with invading nucleic acids. Available data indicate that entire CRISPR loci are transcribed from the leader region, producing primary transcripts containing the full set of CRISPR repeats and embedded invader-derived (or guide) sequences (Hale et al., 2008; Jansen et al., 2002; Lillestol et al.,

2006; Lillestol et al., 2009; Tang et al., 2002; Tang et al., 2005). These large precursor RNAs are processed (or diced) into shorter (~60-70 nucleotide) intermediate RNAs that contain individual invader-targeting sequences (~25-40 nucleotides) by Cas endonucleases that cleave within the repeats (Brouns et al., 2008; Carte et al., 2008). However, the ultimate products of the CRISPR loci appear to be smaller RNAs (Brouns et al., 2008; Hale et al., 2008; Lillestol et al., 2009). In *Pyrococcus furiosus*, the most abundant CRISPR RNAs are two species of ~45 nucleotides and ~39 nucleotides (Hale et al., 2008). These small, abundant products of the CRISPR loci are thought to be the prokaryotic silencing (psi)RNAs of the CRISPR-Cas RNA silencing pathway (Brouns et al., 2008; Hale et al., 2008; Makarova et al., 2006).

Intriguingly, the protein-mediated functions of the CRISPR-Cas system are apparently carried out by distinct sets of Cas proteins in different organisms (Haft et al., 2005). Six “core” CRISPR-associated genes (*cas1* - *cas 6*) are found in many and diverse organisms, however, most organisms have only a subset of these 6 genes and only *cas1* is present in nearly all organisms that appear to possess the system (Haft et al., 2005; Makarova et al., 2006). Furthermore, the core *cas* genes in a given organism are complemented by one or more sets of additional *cas* genes: the *cse*, *csy*, *csn*, *csd*, *cst*, *csh*, *csa*, *csm* and *cmr* genes (Haft et al., 2005). These sets are comprised of 2 to 6 CRISPR-associated genes that co-segregate, and are mostly designated for a prototypical organism (e.g. the *cse* or Cas subtype *Escherichia coli* genes) (Haft et al., 2005). (The *cmr* (Cas module RAMP) gene set is named for its 4 RAMP (repeat-associated mysterious proteins; see below) gene members.) *Escherichia coli* K12, for example, has 3 core *cas* genes and the full set of 5 *cse* genes (which includes the

Escherichia coli subtype member of the core Cas5 gene family, *cas5e*) (Brouns et al., 2008). Phylogenetic analyses suggest that the *cas* genes are distributed by lateral gene transfer (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2002). The functional consequences of the differences in the complement of Cas proteins found among organisms are not yet known.

Functional classes have been predicted for many of the Cas proteins based on sequence, but very few of the proteins have been characterized. Only one of the core Cas proteins, Cas6, has a clearly established function which is to process precursor CRISPR RNAs to release individual invader-targeting RNAs (Carte et al., 2008). Cas1 was recently shown to be a DNA-specific endonuclease with properties consistent with a role in processing invader DNA into fragments that become incorporated into CRISPR loci (Wiedenheft et al., 2009). The five *Escherichia coli* subtype Cas proteins (Cse1-4 and Cas5e (Haft et al., 2005)) have been shown to form a complex that processes precursor CRISPR RNAs in *Escherichia coli* (which lacks Cas6) (Brouns et al., 2008). Many of the Cas proteins are members of the large superfamily of RAMP proteins, which have features of RNA binding proteins (Haft et al., 2005; Makarova et al., 2002; Makarova et al., 2006). At least a few of the RAMPs (including for example Cas6) have been found to possess previously unpredicted nuclease activity (Beloglazova et al., 2008; Brouns et al., 2008; Carte et al., 2008). The Cas proteins are expected to function in various aspects of maintenance of CRISPR gene loci (including addition of new invader-derived elements in response to infection) as well as psiRNA biogenesis and psiRNA-mediated resistance to invaders.

While there is very strong evidence that CRISPR RNAs and Cas proteins function to silence potential invaders in prokaryotes (Barrangou et al., 2007; Brouns et al., 2008; Deveau et al., 2008), the effector complexes and silencing mechanisms of the CRISPR-Cas pathway remain unknown. Recent studies in *Staphylococcus* species and *Escherichia coli* (Brouns et al., 2008; Marraffini and Sontheimer, 2008) indicate that the CRISPR-Cas systems present in those organisms (comprised of the Csm or Cse proteins and several core Cas proteins, respectively) target invader DNA rather than RNA, but the effectors and mechanisms of silencing in these organisms remain unknown. The results presented here demonstrate that the Cmr or RAMP module proteins function with mature psiRNAs to cleave target RNAs. These findings define psiRNA-guided RNA cleavage as a mechanism for the function of the CRISPR-Cas system in organisms that possess the RAMP module of Cas proteins.

Results

Isolation of a complex containing mature psiRNAs and a subset of Cas proteins

PsiRNAs are hypothesized to guide Cas proteins to effect invader silencing in prokaryotes (Brouns et al., 2008; Hale et al., 2008; Makarova et al., 2006). *Pyrococcus furiosus* is a hyperthermophilic archaeon whose genome encodes 200 potential psiRNAs (organized in seven CRISPR loci) and at least 29 potential Cas proteins (largely found in 2 gene clusters), including members of all 6 core Cas protein families and 3 sets of additional Cas proteins: the Cmr, Cst and Csa proteins (see Figure 3.1F). In *Pyrococcus furiosus*, most psiRNAs are processed into 2 species of ~45 nucleotides and ~39 nucleotides (Hale et al., 2008). To gain insight into the functional components of the CRISPR-Cas invader defense pathway, we isolated complexes containing the

mature psiRNA species from *Pyrococcus furiosus* cellular extract on the basis of psiRNA fractionation profiles (Figure 3.1). The doublet of psiRNAs, detectable both by Northern blotting of an individual psiRNA and total RNA staining (SYBR), was purified away from larger CRISPR-derived RNAs (including the 1X intermediate; Hale, 2008) as well as other cellular RNAs (Figure 3.1C).

To determine whether the psiRNAs are components of RNA-protein complexes in the purified fraction (Figure 3.1C), we performed native gel northern analysis. The mobility of the psiRNAs on native gel electrophoresis was reduced in the purified fraction relative to a sample from which proteins were extracted (Figure 3.1D), indicating the presence of psiRNA-protein complexes in the purified fraction. We gel purified the psiRNA-containing complex from the native gel and analyzed the sample by mass spectrometry. The sample contained a mixture of proteins that included seven Cas proteins identified with 99% confidence: Cmr1-1, Cmr1-2, Cmr2, Cmr3, Cmr4, Cmr5, and Cmr6 (Figure 3.1E).

The identities of the non-Cas proteins found in the sample are listed in Table 3.1. Analysis of a native gel-purified psiRNP obtained by an alternate chromatography scheme revealed a similar Cas protein profile (Cmr2, Cmr3, Cmr4, and Cmr6), but few common non-Cas proteins (Table 3.1). The five common co-purifying non-Cas proteins are denoted in Table S1. None of these proteins has any known link to the CRISPR-Cas system.

Remarkably, the seven Cas proteins associated with the complex are all encoded by the tightly linked RAMP module or *cmr* genes (Haft et al., 2005). Moreover, the identified proteins comprise the complete set of Cmr proteins (Haft et al., 2005).

(The independently defined “polymerase cassette” is closely related to the RAMP module (Makarova et al., 2006).) There are 6 *cmr* genes: *cmr2* encodes a predicted polymerase with HD nuclease domains, and *cmr1*, *cmr3*, *cmr4*, and *cmr6* encode repeat-associated mysterious proteins (RAMPs) (Haft et al., 2005; Makarova et al., 2002). The *Pyrococcus furiosus* genome contains two *cmr1* genes and a single representative of each *cmr2* – *cmr6*, and all seven corresponding proteins were found in the purified psiRNP complex (Figure 3.1E). The organization of the genes encoding the seven identified proteins is shown in Figure 3.1F. Six of the seven identified Cas proteins are encoded in a nearly contiguous region of one of the two major *cas* gene loci in *Pyrococcus furiosus*. This locus is located directly adjacent to CRISPR locus 7, and also encodes core Cas proteins Cas1 - Cas4, Cas5t and Cas6. The striking correlation between the evolutionary co-segregation and physical association of the 6 Cmr proteins strongly supports the co-function of the proteins. Our findings indicate that the two mature psiRNA species are components of complexes containing the RAMP module or Cmr proteins in *Pyrococcus furiosus*.

psiRNAs possess a 5’ psiRNA-tag sequence

In order to better understand the nature of the two psiRNA species that are components of the purified complexes, each of the two RNA bands present in the final chromatography sample (Figure 3.2A) was extracted and cloned. We obtained sequences of 51 RNAs (20 from the upper band and 31 from the lower band) that included psiRNAs from all seven *Pyrococcus furiosus* CRISPR loci (Table 3.2). Six RNAs with the same guide sequence were represented in both the upper and lower

bands, consistent with Northern analysis that has shown that most psiRNAs exist in both size forms (Hale et al., 2008).

The cloned psiRNAs consisted primarily of an individual guide (invader-targeting or “spacer”) sequence, however, all of the clones retained a portion of the common repeat sequence at the 5’ end. Indeed, the majority (~70%) of the RNAs in both bands contained an identical 5’ end consisting of an 8-nucleotide segment of the repeat sequence (Figure 3.2A). The difference between the two psiRNA size forms was found at the 3’ ends. Downstream of the repeat sequence, the majority of the clones from the top band contained 37 nucleotides of guide sequence (the full length of a typical guide element in *Pyrococcus furiosus*) (Figure 3.2A, top panel). The 3’ ends of most of the clones from the bottom band were located within the guide sequence. The majority of these RNAs contained 31 nucleotides of guide sequence downstream of the repeat sequence (Figure 3.2A, bottom panel).

The psiRNAs are processed from long CRISPR locus transcripts (Brouns et al., 2008; Hale et al., 2008; Lillestol et al., 2006; Lillestol et al., 2009; Tang et al., 2002; Tang et al., 2005) (Figure 3.2B). In *Pyrococcus furiosus*, the Cas6 endoribonuclease cleaves CRISPR RNAs at a site within the repeat element located 8 nucleotides upstream of the guide sequence, generating the precise 5’ end observed in the two psiRNA species found in the complex (Figure 3.2B; (Carte et al., 2008)). Our results indicate that the 5’ end generated by the Cas6 endoribonuclease is maintained in the mature psiRNAs, but that the RNAs undergo further processing at the 3’ end to generate psiRNAs that contain either ~37 or ~31 nucleotides of guide sequence (Figure 3.2B). The mechanism that defines the two distinct 3’ end boundaries is not known. The

larger ~45-nucleotide mature psiRNA species is generally more abundant than the smaller ~39-nucleotide species ((Hale et al., 2008), Figures 3.1 and 3.2A).

The short repeat sequence that remains at the 5' end of mature psiRNAs in *Pyrococcus furiosus* provides a common identifying sequence tag for the psiRNAs that could function in recognition of the RNAs by the proteins in the CRISPR-Cas pathway. In order to more rigorously delineate the potentially important psiRNA-tag or "psi-tag", we purified small RNAs from *Pyrococcus furiosus*, performed deep sequencing and obtained the sequences of the 5' ends of more than 10,000 CRISPR-derived RNAs (from loci 1-7). The 5' ends of the majority of the RNAs mapped 8 nucleotides upstream of the guide sequence (Figure 3.2C), verifying the presence of a discrete psi-tag on small CRISPR-derived RNAs in *Pyrococcus furiosus*.

The sequences of CRISPR repeats (from which psi-tags are derived) are generally conserved within groups of organisms, but can vary widely (Godde and Bickerton, 2006; Kunin et al., 2007). Thus, while the sequence of the psi-tag found on most *Pyrococcus furiosus* psiRNAs (AUUGAAAG) can be found in the repeat sequence of numerous organisms, psi-tags of distinct sequence and length would be expected in others. We found evidence to support this prediction in the psiRNAs from *Pyrococcus furiosus* CRISPR locus 8, which contains a single nucleotide deletion in the psi-tag region of the repeat. The majority (60%) of the 640 sequenced RNAs that mapped to CRISPR locus 8 possessed a 7-nucleotide AUUGAAG psi-tag. In *Escherichia coli*, CRISPR transcripts are cleaved by a different endoribonuclease (Cse3 of the Cse complex), which nonetheless appears to generate RNAs with an 8-nucleotide AUAAACCG repeat sequence at the 5' end (Brouns et al., 2008). An 8-nucleotide

ACGAGAAC repeat sequence is also present at the 5' termini of CRISPR RNAs in *S. epidermidis* (Marraffini and Sontheimer, 2008), suggesting that the psi-tag is a general feature of the psiRNAs. Interestingly, the distinct CRISPR repeat sequences found in various genomes are accompanied by distinct subsets of Cas proteins (Kunin et al., 2007), which may reflect coupling of specific series of Cas proteins with the psi-tagged RNAs that they recognize.

Homology-dependent cleavage of a target RNA

One hypothesis for the mechanism by which CRISPR RNAs and Cas proteins mediate genome defense is psiRNA-guided cleavage of invader nucleic acids (Makarova et al., 2006). Therefore, we tested the ability of the isolated psiRNP complexes to recognize and cleave a labeled RNA and DNA target complementary to endogenous *Pyrococcus furiosus* psiRNA 7.01 (first psiRNA encoded in CRISPR locus 7, which Northern analysis indicated is present in the native complexes, see Figure 3.1). The 5' end-labeled 7.01 target RNA was cleaved at two sites (site 1 indicated with green vertical line and site 2 indicated with blue vertical line, substrate 1, Figure 3.3B) yielding 5' end-labeled products of 27 and 21 nucleotides (indicated with corresponding green and blue arrowheads, substrate 1, Figure 3.3A). The single-stranded DNA 7.01 target sequence was not cleaved (substrate 3, Figure 3.3).

Further characterization of the cleavage activity revealed that the psiRNP complexes cleave the target RNA on the 5' side of the phosphodiester bond. The 3' end generated by the complex is not a substrate for polyadenylation (Figure 3.4A), indicating the presence of a 3' phosphate (or 2', 3' cyclic phosphate) end. In addition, cleavage activity is lost in the presence of 0.1 mM EDTA indicating that the enzyme depends on

divalent cations (Figure 3.4B). Activity was restored by the addition of 1 mM Mg^{2+}, Mn^{2+}, Ca^{2+}, Zn^{2+}, Ni^{2+} or Fe^{2+} with no detectable change in cleavage sites with any of the metals, but was not supported by Co^{2+}or Cu^{2+} (Figure 3.4B). Cleavage of the target RNA did not require sequences extending beyond the 37-nucleotide region of complementarity with the psiRNA, and occurred at the same two sites in the target RNA lacking sequence extensions (substrate 6, Figure 3.3). No activity was observed toward RNAs that lacked homology with known *Pyrococcus furiosus* psiRNAs, including the reverse 7.01 target sequence, antisense 7.01 target sequence, and a box C/D RNA (substrates 2, 7 and 8, Figure 3.3). Pre-annealing a synthetic psiRNA 7.01 to the 7.01 target RNA (to form a double-stranded RNA target) blocked cleavage by the psiRNPs (substrate 5, Figure 3.3). Finally, we tested a target for endogenous *Pyrococcus furiosus* psiRNA 6.01 and observed cleavage that generates 2 products of the same sizes observed for the 7.01 target RNA (substrate 4, Figure 3.3).

These results demonstrate the presence of cleavage activity in *Pyrococcus furiosus* that is specific for single-stranded RNAs that are complementary to psiRNAs. The activity is associated with a purified fraction that contains 2 mature psiRNA species and 7 RAMP module (Cmr) proteins.

Cleavage of the target RNA occurs a fixed distance from the 3' end of the psiRNA

To investigate the mechanism of psiRNA-directed RNA cleavage, we analyzed the results of cleavage assays with a series of truncations of the 7.01 target RNA (Figure 3.5A). We found that the target RNA truncations analyzed did not affect the locations of the two cleavage sites. The full-length 7.01 target RNA is cleaved at sites 1 and 2 to generate 14- and 20-nucleotide 5' end-labeled products, respectively (Figures

3.3 and 3.5A). The 3' end-truncated target RNAs were cleaved at the same two sites to yield the same two 5' end-labeled cleavage products (except where truncation eliminated cleavage site 2, Δ20-37, Figure 3.5A). On the other hand, in the case of the 5' end-truncated target RNAs, cleavage at the same sites would be expected to generate shorter 5' end-labeled cleavage products. The 14-nucleotide product that results from cleavage of the Δ1-6 target RNA at site 2 was observed (Figure 3.5A), but cleavage at site 1 could not be assessed because the size of the product is below that which could be detected in the experiment. If the twelve- and eighteen-nucleotide 5' end-truncated target RNAs were cleaved at the same two sites, the products would also be outside the range of detection, however, interestingly, very little cleavage of these RNAs was observed (Figure 3.5, Δ1-18 and Δ1-12, compare substrate band +/- complex).

Strikingly, the difference in the sizes of the two cleavage products observed with the various substrates is the same as the difference in the sizes of the two endogenous psiRNA species (6 nucleotides in both cases, Figure 3.3). This size difference as well as the specific product sizes suggest that the two cleavages occur a fixed distance (14 nucleotides) from the 3' ends of the two psiRNAs. Figure 3.5B illustrates the proposed mechanism by which the 45- and 39-nucleotide psiRNAs guide cleavage at target sites 1 and 2, respectively, for each of the target RNAs analyzed here. For example, using the full-length 7.01 target RNA we observed 20- and 14-nucleotide cleavage products (Figure 3.3, panel 5) suggesting cleavage of the bound target RNA 14 nucleotides from the 3' end of the 39- and 45-nucleotide psiRNAs, respectively (Figure 3.5B, F.L.). In addition, a 7-nucleotide extension at the 5' end of the target RNA resulted in a pair of 5'

end-labeled products 27 and 21 nucleotides in length (Figure 3.3, panel 1), consistent with cleavage of the substrate 14 nucleotides from the ends of the two psiRNAs (Figure 3.5B, F.L.+ext). The anchor for this counting mechanism is the 3' end of the psiRNA. While reductions in the extent of duplex formation between the 5' end of the psiRNA and the cleavage site (3' truncations to within 6 nucleotides of the cleavage site) did not have an observable effect on cleavage efficiency, truncations that reduced duplex formation between the 3' end of the psiRNA and the cleavage site had a strong negative impact, suggesting that basepairing of the last 14 nucleotides of the psiRNA with the target is critical for cleavage activity.

The results of these studies indicate that both of the mature psiRNA species are active in guiding target RNA cleavage by a mechanism that depends upon the distance from the 3' end of the psiRNA.

Analysis of reconstituted Cmr-psiRNA complexes

Identification of the Cmr proteins in the purified psiRNP complex (Figure 3.1) along with the evolutionary evidence for their co-function with the CRISPRs (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2002) strongly suggests that the Cmr proteins and psiRNAs function as a complex to cleave target RNAs (Figure 3.3). In order to determine whether the Cmr proteins and psiRNAs are sufficient for function (independent of other co-purifying *Pyrococcus furiosus* components), we tested the ability of purified recombinant Cmr proteins and synthetic psiRNAs to cleave target RNAs (Figure 3.6). A reconstituted set of six *Pyrococcus furiosus* Cmr proteins (Cmr1-1, Cmr2 – Cmr6) and two mature psiRNA species (45- and 39-nucleotide psiRNA 7.01, found in the native complex based on Northern analysis (Figure 3.1) and activity of the

native complex against the 7.01 target (Figure 3.3)) cleaved the target RNA at 2 sites generating the same size products as those observed with the isolated native complex (Figure 3.6A). While both *Pyrococcus furiosus* isoforms of the Cmr1 protein are present in the isolated complexes (Figure 3.1), we found that only one of the two proteins (Cmr1-1) was required for a functional reconstituted complex (Figure 3.6A), suggesting that the isoforms may perform redundant functions. No activity was observed in the absence of the psiRNAs or in the absence of the Cmr proteins (Figure 3.6A), indicating that both are necessary. These results demonstrate that the RAMP module Cas proteins and psiRNAs function together to cleave complementary target RNAs.

In order to determine whether all of the six Cmr proteins are essential for psiRNA-guided RNA cleavage, we assayed cleavage activity in the absence of each of the individual proteins (Figure 3.6B). Omission of Cmr5 did not observably affect the activity of the complex (Figure 3.6B). However, cleavage was significantly reduced in the absence of any one of the other proteins (Figure 3.6B), indicating that 5 of the 6 RAMP module proteins are required for activity of the psiRNA-Cmr protein complex.

Finally, we had reconstituted the cleavage activity profile observed for the native complexes using the two psiRNA species (45- and 39-nucleotides) (e.g. Figure 3.6A). Our model for the mechanism of cleavage predicts that each of the psiRNAs guides a distinct cleavage: the 45-nucleotide psiRNA at site 1, and the 39-nucleotide psiRNA at site 2 (see Figure 3.5B). To determine whether both psiRNAs are required for activity, and whether each guides the distinct cleavage that is predicted by the model, we tested the activity of complexes reconstituted with a single psiRNA. As predicted, we found that the 45-nucleotide psiRNA guided cleavage at site 1 producing a 14-nucleotide 5’

end-labeled product, and the 39-nucleotide psiRNA guided cleavage at site 2 producing a 20-nucleotide 5' end-labeled product (Figure 3.6C). Based on our truncation analysis (Figure 3.5, D20-37), the larger product of the cleavage guided by the 39-nucleotide psiRNA could act as a substrate for cleavage guided by the 45-nucleotide psiRNA, and consistent with this, we often obtain more of the smaller cleavage product in cleavage assays where both guide RNAs are present with either the native complex or the reconstituted complex (e.g. Figure 3.6A). The results of these experiments demonstrate that each of the psiRNA species is competent to form functional psiRNPs and guides cleavage 14 nucleotides from its 3' end.

Discussion

The findings presented here reveal the mechanism of action of an RNA-protein complex implicated in a novel RNA silencing pathway that functions in invader defense in prokaryotes. Previous work had shown that both invader-specific sequences within CRISPRs and Cas protein genes are important in virus and plasmid resistance in prokaryotes (Barrangou et al., 2007; Brouns et al., 2008; Deveau et al., 2008; Marraffini and Sontheimer, 2008). The results presented here establish how small RNAs from CRISPRs and the RAMP module Cas proteins function together to destroy RNAs recognized by the CRISPR RNAs. The major findings and models established in this work are summarized in Figure 3.7.

Our findings indicate that the RAMP module of the CRISPR-Cas system silences invaders by psiRNA-guided cleavage of invader RNAs (Figure 3.7). Specifically, the results indicate that psiRNAs present in complexes with the Cmr proteins recognize and bind an invader RNA such as a viral mRNA (via the psiRNA guide sequence co-opted

from the invader by another branch of the CRISPR-Cas system), and that the complex then cleaves the invader RNA, destroying the message and presumably blocking the viral life cycle. The psiRNA-Cmr complexes cleave complementary RNAs (Figures 3.3 and 3.6). Five of the six Cmr proteins are required for target RNA cleavage (Figure 3.6) and the component of the complex that provides catalytic activity remains to be determined. Cmr2 contains a predicted nuclease domain (Makarova et al., 2002; Makarova et al., 2006), however the other four essential proteins (Cmr1, 3, 4 and 6) belong to the RAMP superfamily, members of which have been found to be ribonucleases (Beloglazova et al., 2008; Brouns et al., 2008; Carte et al., 2008). It will be important in future work to identify the catalytic component(s) of the psiRNA-Cmr protein complex. Our data indicate that the Cmr ribonuclease generates products with 3' phosphate (or 2', 3' cyclic phosphate) and 5' hydroxy termini and requires divalent metal ions for activity (Figure 3.4).

Our results also establish a simple model for the mechanism of cleavage site selection by the psiRNA-Cmr effector complex - a 14-nucleotide ruler anchored by the 3' end of the psiRNA (Figure 3.7). We found that *Pyrococcus furiosus* psiRNAs occur in two lengths that share a 5' psi-tag (derived from the CRISPR repeat) and contain either ~37 or ~31 nucleotides of guide sequence (Figures 3.1 and 3.2). Both psiRNA species are associated with the Cmr effector complex (Figure 3.1) and each guides cleavage at a distinct site (Figure 3.6C). Analysis of the cleavage products of both psiRNAs and of a series of substrate RNAs (Figures 3.3, 3.5 and 3.6) indicates that the complex cleaves based on a 14-nucleotide counting mechanism anchored by the 3' end of the psiRNA.

The results suggest that the 3' end of the psiRNA places the bound target RNA relative to the enzyme active site (Figure 3.7).

The activity of the psiRNA-Cmr protein complex (RNA-guided RNA cleavage) bears an interesting resemblance to that of Argonaute 2 (a.k.a. Slicer) (Liu et al., 2004), an enzyme with an analogous function in the eukaryotic RNAi pathway, however there is little similarity between the enzymes. There is no significant sequence homology between the Cmr proteins and Argonaute 2 (or between any of the Cas proteins and known components of the eukaryotic RNAi pathway). Both the psiRNA-Cmr complex and Argonaute 2 employ a ruler mechanism for cleavage site selection; however, in the case of Argonaute 2, the site of cleavage is located ~10-11 nucleotides from the 5' end of the siRNA (Elbashir et al., 2001a; Elbashir et al., 2001b). The activity of both enzymes requires divalent metal ions (Figure 3.4 and (Schwarz et al., 2004)), however for the psiRNA-Cmr RNP, it is not yet clear whether the metal is involved in cleavage catalysis or is required for some other essential aspect of the functionality of this multi-component complex. Finally, Argonaute 2 cleaves target RNAs on the 3' side of the phosphodiester bond, leaving 3' OH and 5' phosphate termini (Martinez and Tuschl, 2004). It is interesting that eukaryotes and prokaryotes exploit distinct small RNA-guided gene silencing pathways to combat viruses and other mobile genetic elements that they encounter (Ghildiyal and Zamore, 2009; Malone and Hannon, 2009).

Figure 3.7 also illustrates the Cmr-psiRNA effector complex model that arises from the findings presented here. Both size classes of psiRNAs and all seven Cmr proteins are found in complexes in active, purified fractions (Figure 3.1), however accurate RNA-guided cleavage activity can be reconstituted with either psiRNA species

and with a single Cmr1 isoform (Figure 3.6). We hypothesize that each psiRNA associates with a single set of six Cmr proteins, and that Cmr1-1 and Cmr1-2 function redundantly in *Pyrococcus furiosus*. Five unrelated proteins that co-purified with the complexes (Table 3.1) are not essential for reconstitution of cleavage activity *in vitro* (Figure 3.6) and are not included in our model, but could play a role in function *in vivo*. Recognition of the psiRNAs by the Cmr proteins and psiRNA-Cmr complex assembly likely depend upon conserved features of the RNAs that could include 5' and 3' end groups and folded structure as well as the psi-tag. Our data reveal that the psiRNA-Cmr complex can utilize psiRNAs of different sizes to cleave a target RNA at distinct sites (Figure 3.6C). Thus, the two size forms of psiRNAs present in *Pyrococcus furiosus* may provide more certain and efficient target destruction.

Our data indicate that the function of the RAMP module of Cas proteins is psiRNA-guided destruction of invading target RNA. The widespread occurrence of the *cmr* genes in diverse archaea (including *Sulfolobus* and *Archaeoglobus* species) and bacteria (including *Bacillus* and *Myxococcus* species) indicates that invader RNA cleavage is a mechanism utilized by many prokaryotes for viral defense (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). However, not all prokaryotes with the CRISPR-Cas system possess the RAMP module (Cmr) proteins. In these numerous other organisms, it is possible that a different set of Cas proteins mediates psiRNA-guided RNA cleavage or that Cas proteins effect invader resistance by another mechanism. Indeed, very recent work indicates that the CRISPR-Cas system targets invader DNA in a strain of *Staphylococcus epidermidis* and perhaps *Escherichia coli* (Brouns et al., 2008; Marraffini and Sontheimer, 2008), which possess the Mtube (Csm)

and Ecoli (Cse) subtype Cas protein modules, respectively (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006). The prokaryotes include evolutionarily distant and very diverse organisms. Diversity in the core components of the eukaryotic RNAi machinery has led to a tremendous variety of observed RNA-mediated gene silencing pathways that can act at post-transcriptional or transcriptional levels (Chapman and Carrington, 2007; Farazi et al., 2008; Hutvagner and Simard, 2008; Zaratiegui et al., 2007). The diversity of Cas proteins found in CRISPR-containing prokaryotes may reflect significantly different mechanisms of CRISPR element integration, CRISPR RNA biogenesis, and invader silencing.

Experimental Procedures

Chromatography

Pyrococcus furiosus S100 extract was prepared from approximately 4 grams of cells. Cells were resuspended in 20 mL of 50 mM Tris (pH 7.0), 100 U RNase-free DNase (Promega), and 0.5 mM phenylmethanesulphonyl fluoride (PMSF) at room temperature by stirring. The resulting whole cell extract was subject to ultracentrifugation at 100,000 x g for 1.5 hours using an SW 41 Ti rotor (Beckman). The resulting S100 extract was loaded onto a 5 mL Q-sepharose Fast Flow (GE) pre-packed column. Proteins were eluted using a 0-1 M NaCl gradient. Fractions were analyzed by Northern analysis by isolating RNA from 100 ul of each fraction using Trizol LS (Invitrogen, following manufacturer's instructions). The RNAs were separated on 15% TBE-urea gels (Criterion, Bio-Rad), blotted and analyzed for the presence of a single guide sequence as described previously (Hale et al., 2008). Peak fractions containing the psiRNA doublet were further separated on a second 5 mL Q-sepharose column,

eluted with 220-430 mM NaCl. Fractions were analyzed as described above. Peak fractions were pooled, diluted in 50 mM sodium phosphate buffer, pH 7.0, and loaded onto a 5 mL S-sepharose column (GE). Bound proteins were eluted with a gradient of 0-1 M NaCl. Native gel northern analysis was performed as described previously (Hale et al., 2008). The secondary data shown in Table S1 was obtained from S100 extract fractionated on a DEAE column as previously described (Hale et al., 2008) followed by a hydroxyapatite column eluted with a gradient of 5-500 mM sodium phosphate buffer, pH 6.5, and further purified by native gel electrophoresis.

Protein assignment by tandem mass spectrometry

In-gel and in-solution tryptic digests were performed as previously described (Lim et al., 2008; Wells et al., 2002). Desalted tryptic peptides were analyzed by nLC-MS/MS on a linear ion-trap (LTQ, ThermoFisher) as previously described (Lim et al., 2008). Acquired data was searched against a *Pyrococcus furiosus*-specific database (forward and inverted) using the TurboSEQUEST algorithm (ThermoFisher). Data was collated and filtered to obtain a 1% false discovery rate at the protein level using the ProteoIQ software package (BioInquire) that is based on the PROVALT algorithm (Weatherly et al., 2005).

Cloning and sequencing of psiRNAs from the purified complexes

RNAs from S-column fractions (isolated as described above for Northern analysis) were treated with 1 U calf intestinal alkaline phosphatase (Promega) for 1 hour at 37°C, followed by extraction with phenol:chloroform:isoamyl alcohol (PCI; pH 5.2, Fisher) and ethanol precipitation. The resulting RNAs were separated by 15% polyacrylamide, TBE-urea gels (Criterion, Bio-Rad), visualized by SYBR Gold staining

(Invitrogen) and the visible bands were excised. RNAs were passively eluted overnight in 0.5 M ammonium acetate, 0.1% SDS, 0.5 mM EDTA, followed by ethanol precipitation. A 5'-phosphorylated, 3' capped oligonucleotide (5'-pCTCGAGATCTGGATCCGGG-ddC3'; IDT) was ligated with T4 RNA ligase to the 3' end of the RNAs. The ligated RNAs were PCI extracted, ethanol precipitated, gel purified, and subject to reverse transcription using Superscript III (Invitrogen) RT (as described by the manufacturer), followed by gel purification. The gel-purified cDNAs were polyA-tailed for 15 minutes at 37°C using terminal deoxynucleotide transferase (Roche) using manufacturer's recommendations. PCR was performed to amplify the cDNA libraries using the following primers: 5'-CCCGGATCCAGATCTCGAG-3', 5'-GCGAATTCTGCAG$(T)_{30}$-3'). cDNAs were cloned into the TOPO pCRII (Invitrogen) cloning vector and transformed into TOP10 cells. White and light-blue colonies were chosen for plasmid DNA preparation, and sequencing using the M13 Reverse and T7 promoter sequencing primers was performed by the University of Georgia Sequencing and Synthesis Facility.

Small RNA deep sequencing

Small RNA libraries were prepared using the Illumina small RNA Sample preparation kit as described by the manufacturer (Illumina). Briefly, total RNA was isolated from *Pyrococcus furiosus* and fractionated on a 15% polyacrylamide/urea gel, and small RNAs 18-65 nt in length were excised from the gel. 5' and 3' adapters were sequentially ligated to the small RNAs and the ligation products were gel-purified between each step. The RNAs were then reverse-transcribed and PCR-amplified for 16 cycles. The library was purified with a Qiagen QuickPrep column and quantitated using

an Agilent Bioanalyzer and a nanodrop. The sample was diluted to a concentration of 2 pM and subjected to 42 cycles of sequencing on the Illumina Genome Analyzer II.

Small RNA Analysis

Sequence data was extracted from the images generated by the Illumina Genome Analyzer II using the software applications Firecrest and Bustard. The adapter sequences were then trimmed from the small RNA reads, which were then mapped to the *Pyrococcus furiosus* genome using btbatchblast. Only reads that mapped perfectly to the genome over their entire length were used for further analysis. The location and number of reads that initiate within the CRISPR repeats were determined using a perlscript. As the maximal read length of the sequences was 42 nt, it was not possible to be certain that the 3' end of a read represented the actual 3' end of the small RNA. Therefore, the deep sequencing data was only used to determine the 5' ends.

Nuclease assays

To detect target RNA cleavage, 2 µL of the peak S-column fractions (Figure 3.1C) or 500 nM each of recombinant proteins was incubated with 0.05 pmoles of ^{32}P-5' end-labeled synthetic target RNAs (Figures 3.3, 3.5 and 3.6) and 0.5 pmoles of each unlabeled psiRNA (Figure 3.6) for 1 hour at 70°C in 20 mM HEPES pH 7.0, 250 mM KCl, 1.5 mM $MgCl_2$, 1 mM ATP, 10 mM DTT, in the presence of 1 unit of SUPERase-In ribonuclease inhibitor (Applied Biosystems). For assays with recombinant proteins, the psiRNAs were first incubated with the proteins for 30 minutes at 70°C prior to the addition of target RNA. Reaction products were isolated by treatment with 800 ng of proteinase K for 30 minutes at room temperature, followed by PCI extraction and ethanol precipitation. The resulting RNAs were separated on 15% polyacrylamide, TBE

7M urea gels and visualized by phosphorimaging. 5' end-labeled RNA size standards (Decade Markers, Applied Biosystems) were used to determine the sizes of the observed products. Annealed RNAs were prepared by mixing equimolar amounts of RNAs in 30 mM HEPES pH 7.4, 100 mM potassium acetate, 2 mM magnesium acetate and incubating for 1 minute at 95°C, followed by 1 hour at 37°C. Annealing was confirmed by non-denaturing 8% PAGE.

For analysis of the chemical ends of the cleavage products, cleavage reactions were performed using 5'-end labeled target as described above. The resulting RNA products were isolated by PCI extraction and ethanol precipitation, and subject to polyadenylation by incubation with 5 U *Escherichia coli* polyA polymerase (NEB) for 15 minutes at 37°C as described by the manufacturer. The reaction was stopped by PCI extraction, followed by ethanol precipitation. The resulting products were analyzed on 15% polyacrylamide, TBE 7M Urea gels as described above.

In order to determine the divalent metal requirements of the purified complex, cleavage reactions were performed for 1 hour at 70°C in 50 mM HEPES pH 7.0, 250 mM KCl, 1 mM ATP, 10 mM DTT, 0.1 mM EDTA, and 1 mM metal (if applicable) in the presence of 1 unit of SUPERase-In ribonuclease inhibitor (Applied Biosystems). Certified metal reference solutions (Spex CertiPrep except calcium obtained from Fisher Scientific) were added to 1 mM final concentration. The resulting products were isolated and analyzed as described above.

Expression and purification of recombinant proteins

The genes encoding *Pyrococcus furiosus* Cmr1-1 (PF1130), Cmr2 (PF1129), Cmr3 (PF1128), Cmr4 (PF1126), Cmr5 (PF1125) and Cmr6 (PF1124) were amplified by

PCR from genomic DNA or existing constructs and cloned into a modified version of pET24d (PF1124, PF1125 and PF1126) or pET200D (PF1128, PF1129 and PF1130). The recombinant proteins were expressed in *Escherichia coli* BL21-RIPL cells (DE3, Stratagene). The cells (400 mL cultures) were grown to a OD_{600} of 0.7, and expression of the proteins was induced with 1 mM isopropyl-b-D-thiogalactopyranoside (IPTG) overnight at room temperature. The cells were pelleted, resuspended in 20 mM sodium phosphate buffer (pH 7.6), 500 mM NaCl and 0.1 mM phenylmethylsulfonyl fluoride (PMSF), and disrupted by sonication. The sonicated sample was centrifuged at 4,500 rpm for 15 min at 4°C. The supernatant was heated at 75-78°C for 20 min, centrifuged at 4,500 rpm for 20 min at 4°C, and filtered (0.8 µm pore size Millex filter unit, Millipore). The recombinant histidine-tagged proteins were purified by batch purification using 50 µl Ni–NTA agarose beads (Qiagen) equilibrated with resuspension buffer. Following 3 washes (resuspension buffer), the bound proteins were eluted with resuspension buffer containing 500 mM imidazole. The protein samples were dialyzed at room temperature against 40 mM HEPES (pH 7.0) and 500 mM KCl prior to performing activity assays.

Synthetic psiRNAs

The 45- and 39-nucleotide psiRNAs were chemically synthesized (Integrated DNA Technologies). The sequence of the 45-nucleotide psiRNA 7.01 is: AUUGAAAGUUGUAGUAUGCGGUCCUUGCGGCUGAGAGCACUUCAG. The sequence of the 39-nucleotide psiRNA 7.01 is:
AUUGAAAGUUGUAGUAUGCGGUCCUU
GCGGCUGAGAGCA.

Acknowledgements

We extend special thanks to Tim Davies (Director, University of Georgia Bioexpression & Fermentation Facility) for *Pyrococcus furiosus* cells, the University of Connecticut Health Center Translational Genomics Core Facility for use of the Illumina Genome Analyzer, Frank Sugar (University of Georgia Bioexpression Facility) for expert advice on chromatography, Mike Adams and the Southeast Collaboratory for Structural Genomics (University of Georgia) for constructs, Lindsay Jones, Joshua Elmore and Sonali Majumdar (Terns Lab, University of Georgia) for generation of protein expression constructs, and Claiborne Glover (University of Georgia) for critical reading. L.W. is a Georgia Cancer Coalition Distinguished Scientist. This work was supported by National Institutes of Health grants RO1GM54682 (M.T. and R.T.) and R01GM062516 (B.R.G.).

References

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

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.

Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E.V., Edwards, A.M., Savchenko, A., and Yakunin, A.F. (2008). A Novel Family of Sequence-specific Endoribonucleases Associated with the Clustered Regularly Interspaced Short Palindromic Repeats. J Biol Chem *283*, 20361-20371.

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.

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.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet *8*, 884-896.

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.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001a). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001b). Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. Embo J *20*, 6877-6888.

Farazi, T.A., Juranek, S.A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development *135*, 1201-1214.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.

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

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

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

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol *9*, 22-32.

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.

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

Lillestol, R.K., Redder, P., Garrett, R.A., 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.

Lim, J.M., Sherling, D., Teo, C.F., Hausman, D.B., Lin, D., and Wells, L. (2008). Defining the regulated secreted proteome of rodent adipocytes upon the induction of insulin resistance. J Proteome Res *7*, 1251-1263.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

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

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.

Malone, C.D., and Hannon, G.J. (2009). Small RNAs as guardians of the genome. Cell *136*, 656-668.

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

Martinez, J., and Tuschl, T. (2004). RISC is a 5' phosphomonoester-producing RNA endonuclease. Genes Dev *18*, 975-980.

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.

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.

Schwarz, D.S., Tomari, Y., and Zamore, P.D. (2004). The RNA-induced silencing complex is a Mg2+-dependent endonuclease. Curr Biol *14*, 787-791.

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.

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.

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

Weatherly, D.B., Atwood, J.A., 3rd, Minning, T.A., Cavola, C., Tarleton, R.L., and Orlando, R. (2005). A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics *4*, 762-772.

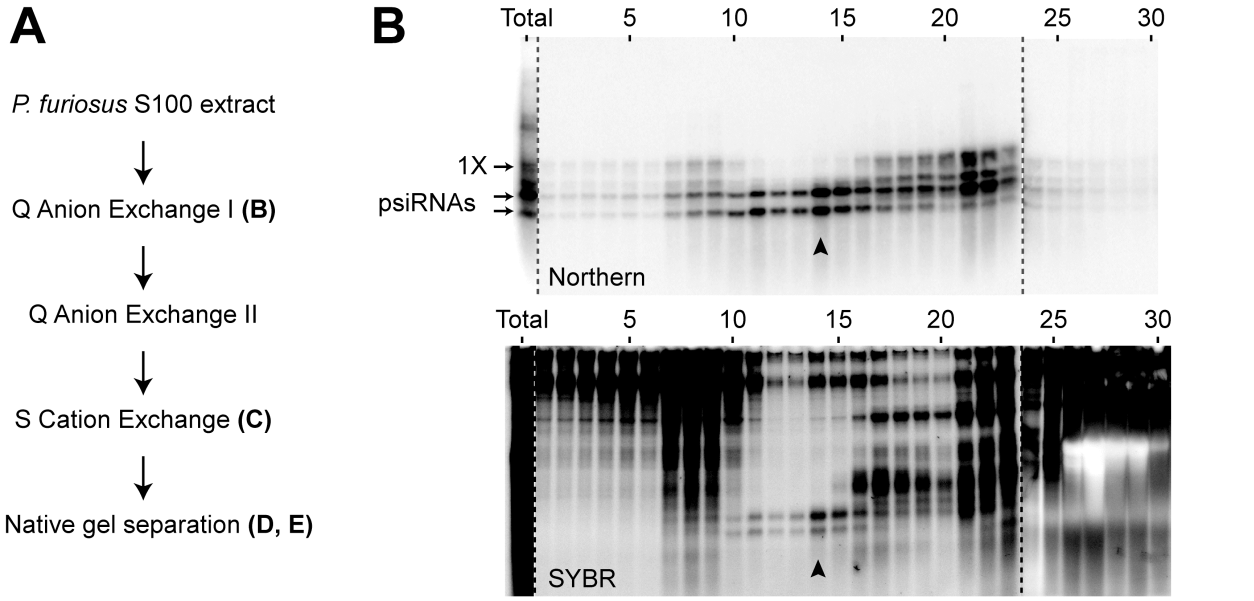
Wells, L., Vosseller, K., Cole, R.N., Cronshaw, J.M., Matunis, M.J., and Hart, G.W. (2002). Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. Mol Cell Proteomics *1*, 791-804.

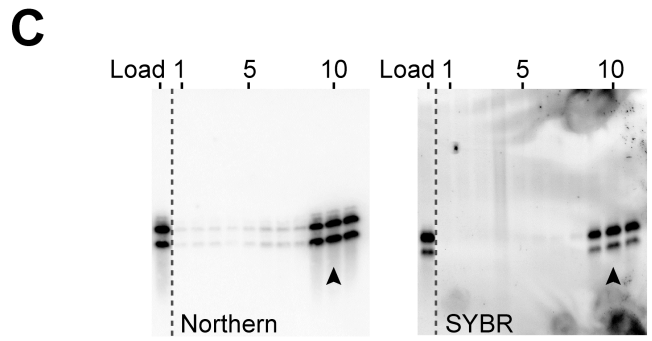
Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure *17*, 904-912.

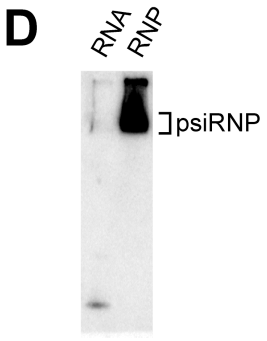
Zaratiegui, M., Irvine, D.V., and Martienssen, R.A. (2007). Noncoding RNAs and gene silencing. Cell *128*, 763-776.

Figure 3.1. Identification of a ribonucleoprotein complex containing psiRNAs and Cas proteins. A) psiRNP purification scheme. Letters indicate the location of corresponding data within Figure 3.1. B) psiRNA (Northern panel) and total RNA (SYBR panel) profiles across the initial Q-sepharose anion exchange fractions and an unfractionated sample (total). Northern analysis (top panel) was performed for *Pyrococcus furiosus* psiRNA 7.01. The positions of the mature psiRNAs and 1X intermediate RNA (Hale et al., 2008) are indicated. The SYBR panel shows all RNAs detected by SYBR Gold staining. The peak fraction is indicated by an arrow in each panel. Non-contiguous lanes from the same gel (total sample) and a second gel (fractions 24-30) are indicated by dashed lines. C) psiRNA (Northern analysis of psiRNA 7.01) and total RNA (SYBR staining) profiles across the S-sepharose cation exchange fractions and starting material (load). The peak fraction is indicated by an arrow in each panel. Non-contiguous lanes from the same gel are indicated by dashed lines. D) Native gel Northern analysis of the psiRNP. The peak S-sepharose fraction (arrow, C) was fractionated by native gel electrophoresis and analyzed by Northern blotting for psiRNA 7.01. RNA extracted from the same fraction was co-analyzed. The position of the psiRNP is indicated. E) Cas proteins identified by tandem mass spectrometry. The isolated psiRNP (D) was subject to in-gel trypsin digestion and tandem mass spectrometry. Sequence coverage and the number of unique peptides for Cas proteins identified with 99% confidence are shown. *Pyrococcus furiosus cas* gene names are as given (Haft et al., 2005) and proposed functions are as predicted (Haft et al., 2005; Makarova et al., 2006). See also Table S1. F) Genome organization of predicted *Pyrococcus furiosus cas* genes. Operon organization and COG assignments were

adapted from NCBI database. Core *cas* genes (*cas*) and Cas module-RAMP (*cmr*), Cas subtype Apern (*csa*) and Cas subtype Tneap (*cst*) genes are indicated. Proteins identified by mass spectrometry are indicated in black.







E

Protein	Unique Peptides	Coverage	COG	Gene name	Proposed Function
PF1129	54	55.9%	1353	*cmr2*	Polymerase/Nuclease
PF1128	13	40.4%	1769	*cmr3*	RNA-binding protein (RAMP)
PF1126	14	58.3%	1336	*cmr4*	RNA-binding protein (RAMP)
PF1124	5	16.2%	1604	*cmr6*	RNA-binding protein (RAMP)
PF1125	5	24.9%	3337	*cmr5*	ND
PF0352	5	28.1%	1367	*cmr1-2*	RNA-binding protein (RAMP)
PF1130	1	3.6%	1367	*cmr1-1*	RNA-binding protein (RAMP)

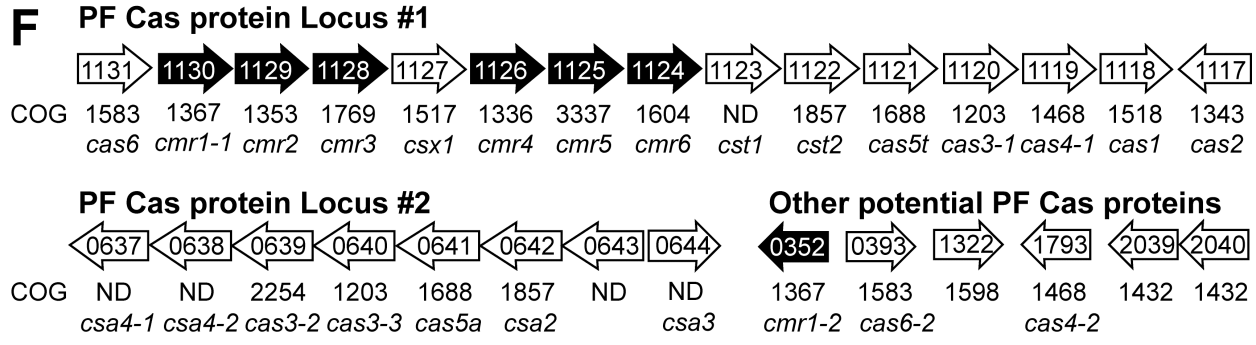
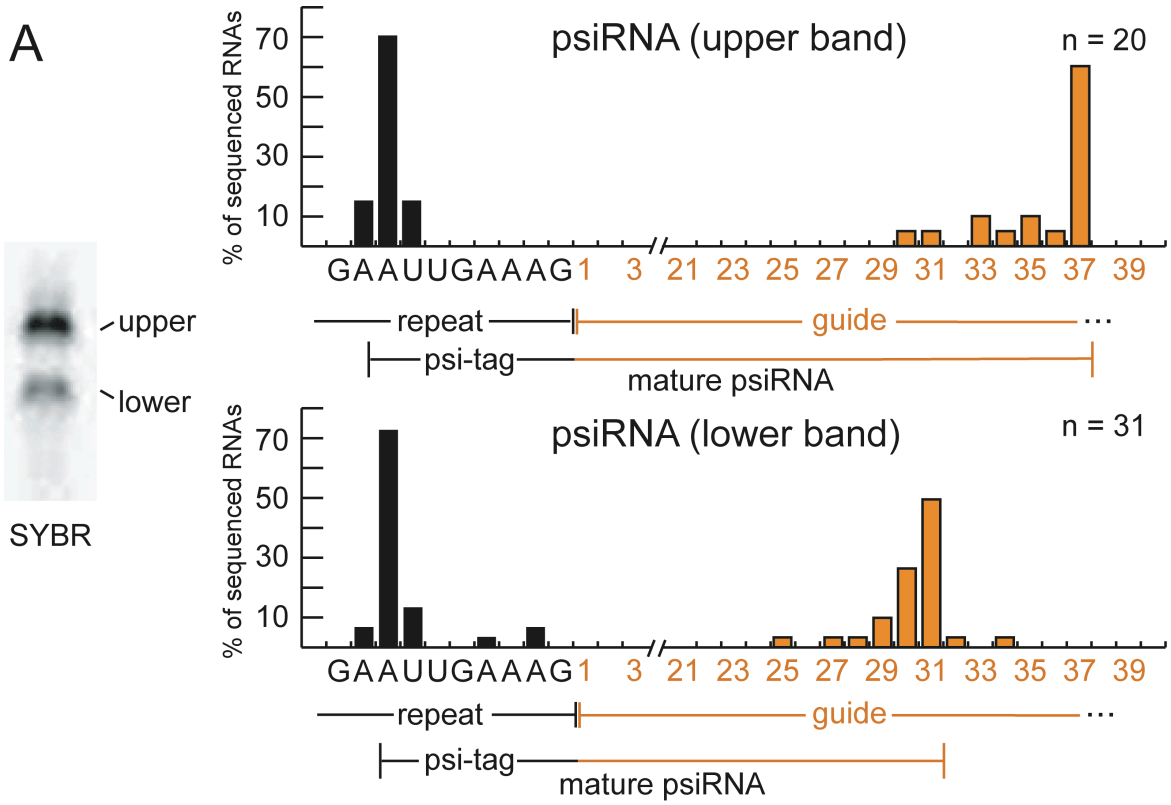


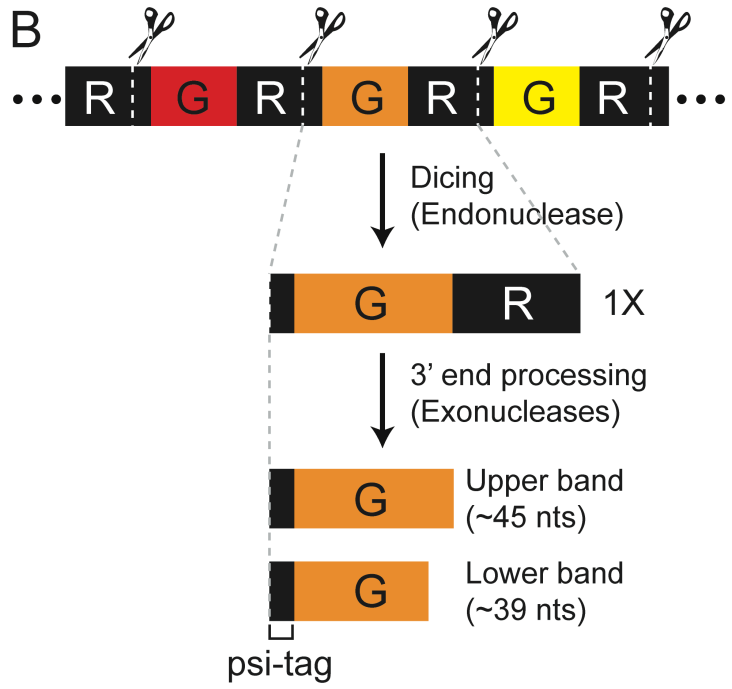
Table 3.1. **Proteins identified by tandem mass spectrometry of native gel-purified RNA-protein complexes, expanded from Figure 3.1E.** All proteins that were identified in the gel-purified complex shown in Figure 3.1D (Preparation 1) and in the native gel-purified complex obtained by an alternate chromatography scheme (Preparation 2; see experimental procedures) are listed. Numbers represent the % amino acid sequence coverage, with the number of unique peptides in parentheses. The annotated function is from the NCBI database.

	Protein	Prep 1	Prep 2	Annotated Function
Cas proteins	PF1129	55.9 (54)	26.9 (20)	hypothetical protein PF1129
	PF1128	40.4 (13)	20.5 (5)	hypothetical protein PF1128
	PF1126	58.3 (14)	17.6 (3)	hypothetical protein PF1126
	PF1124	16.2 (5)	11.8 (4)	hypothetical protein PF1124
	PF0352	28.1 (5)		hypothetical protein PF0352
	PF1125	24.9 (5)		hypothetical protein PF1125
	PF1130	3.6 (1)		hypothetical protein PF1130
Non-Cas proteins	PF1717	76.2 (28)		translation initiation factor IF-2 gamma subunit
	PF1683	73.6 (19)		N-acetyl-gamma-glutamyl-phosphate reductase
	PF0990	60.6 (26)		phenylalanyl-tRNA synthetase beta subunit
	PF1685	59.0 (20)		acetylornithine/acetyl-lysine aminotransferase
	PF0481	55.7 (7)		translation initiation factor IF-2 beta subunit
	PF1827	53.8 (14)	20.8 (4)	hypothetical protein PF1827
	PF1881	51.6 (4)		chromatin protein
	PF0989	45.1 (22)		phenylalanyl-tRNA synthetase alpha subunit
	PF0124	34.3 (14)		hypothetical protein PF0124
	PF1140	30.2 (7)		translation initiation factor IF-2 alpha subunit
	PF0495	29.7 (34)		reverse gyrase
	PF1204	29.2 (11)		seryl-tRNA synthetase
	PF1264	26.1 (3)		translation initiation factor IF-5A
	PF0351	25.6 (8)		hypothetical protein PF0351
	PF1238	23.9 (14)		putative ABC transporter
	PF1615	23.1 (18)		hypothetical protein PF1615
	PF0496	21.2 (5)		hypothetical protein PF0496
	PF0594	18.4 (4)	14.3 (2)	Ornithine carbamoyltransferase
	PF1405	16.6 (10)	12.9 (7)	Cleavage and polyadenylation specifity factor protein
	PF0547	15.8 (5)		hypothetical protein PF0547
	PF0969	14.7 (4)		2-ketovalerate ferredoxin oxidoreductase subunit alpha
	PF0220	14.1 (6)	13.6 (7)	Hexulose-6-phosphate synthase
	PF1375	13.3 (6)		elongation factor Tu
	PF1976	12.1 (6)		L-aspartate oxidase
	PF0666	11.5 (6)		nol1-nop2-sun family putative nucleolar protein IV
	PF1746	11.0 (6)		hypothetical protein PF1746
	PF0251	11.0 (4)		hypothetical protein PF0251
	PF1579	10.4 (7)		DNA topoisomerase VI subunit B

	PF0966	10.4 (4)		2-oxoglutarate ferredoxin oxidoreductase
	PF0533	10.1 (7)		indolepyruvate ferredoxin oxidoreductase subunit a
	PF1578	9.2 (3)		DNA topoisomerase VI subunit A
	PF0026	8.8 (4)		tRNA nucleotidyltransferase
	PF1540	8.7 (5)		ADP forming acetyl coenzyme A synthetase
	PF1203	8.1 (5)		formaldehyde:ferredoxin oxidoreductase
	PF1046	7.9 (3)		queuine trna-ribosyltransferase
	PF0464	7.5 (4)		glyceraldehyde-3-phosphate:ferredoxin oxidoreductase
	PF1768	5.1 (2)		2-oxoglutarate ferredoxin oxidoreductase
	PF0440	3.9 (5)		ribonucleotide-diphosphate reductase alpha subunit
	PF1843	1.7 (2)	7.3 (6)	chromosome segregation protein smc
	PF0102		76.6 (15)	hypothetical protein PF0102
	PF1883		74.9 (13)	small heat shock protein
	PF1548		63.3 (24)	hypothetical protein PF1548
	PF1931		27.9 (6)	hypothetical protein PF1931
	PF0162		11.2 (2)	hypothetical protein PF0162
	PF0204		6.0 (2)	hypothetical protein PF0204
	PF1871		3.7 (1)	"N(2),N(2)-dimethylguanosine tRNA methyltransferase"
	PF1245		2.2 (1)	hypothetical d-nopaline dehydrogenase
	PF1167		1.5 (1)	chromosome segregation protein

Figure 3.2. psiRNA species in the RNP contain a common 5' sequence element and distinct 3' termini. A) Sequence analysis of RNAs associated with the complex. RNA species present in the S-sepharose fraction (visualized by SYBR Gold staining) are shown in SYBR panel. RNAs in the upper and lower bands were isolated, cloned, and sequenced. Graphs show the percentage of sequenced RNAs with 5' ends located at specific positions within the repeat sequence (black), and with indicated numbers of guide sequence nucleotides downstream of the repeat sequence (orange). The average guide sequence is 37 nucleotides in *Pyrococcus furiosus*. A consensus for each psiRNA species is diagrammed under each graph. The 8-nucleotide repeat sequence found at the 5' end of the majority of the psiRNAs is indicated as the psi-tag. See Table S2 for sequences of the cloned psiRNAs. B) Model for biogenesis of the two psiRNA species in *Pyrococcus furiosus*. CRISPR locus transcripts containing alternating repeat (R, black segments) and guide (G, colored segments) elements are cleaved at a specific site within the repeat by the Cas6 endoribonuclease (Carte et al., 2008), ultimately producing 1X intermediate RNAs that contain a full invader-targeting sequence flanked on both sides by segments of the repeat. The mature RNAs retain the 5' end repeat sequence (psi-tag). Uncharacterized 3' end processing of the 1X intermediate by endo- and/or exo-nucleases forms the two major mature psiRNAs: a 45-nucleotide species that contains the 8-nucleotide psi-tag and a full guide sequence, and a 39-nucleotide species that contains a shorter 31-nucleotide guide sequence. C) Deep sequencing of small RNAs from *Pyrococcus furiosus* confirms the presence of the psi-tag. The 5' ends of the sequenced psiRNAs are graphed as in A. The number of total clones analyzed (n) is indicated in the graphs of panels A and C.


A
psiRNA (upper band)
n = 20
% of sequenced RNAs
70
50
30
10
GAAUUGAAAG 1 3 21 23 25 27 29 31 33 35 37 39
repeat
guide
psi-tag
mature psiRNA
upper
lower
SYBR
psiRNA (lower band)
n = 31
% of sequenced RNAs
70
50
30
10
GAAUUGAAAG 1 3 21 23 25 27 29 31 33 35 37 39
repeat
guide
psi-tag
mature psiRNA


B
R G R G R G R
Dicing
(Endonuclease)
G R
1X
3' end processing
(Exonucleases)
G
Upper band
(~45 nts)
G
Lower band
(~39 nts)
psi-tag

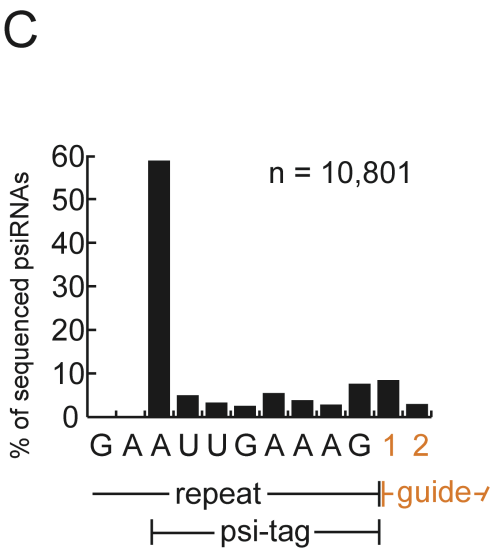

C
n = 10,801
% of sequenced psiRNAs
60
50
40
30
20
10
0
G A A U U G A A A G 1 2
repeat
guide
psi-tag

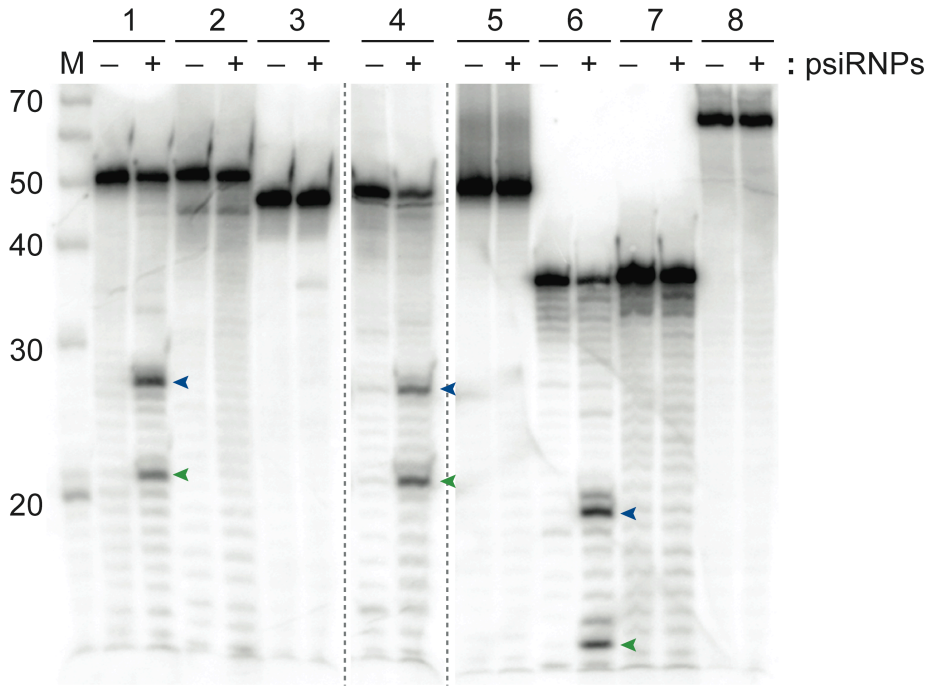
Table 3.2. **Cloned psiRNAs from Figure 3.2A.** The sequences of the CRISPR-derived clones obtained from the upper and lower psiRNA bands (Figure 3.2A) are shown. The length and psiRNA designation of each sequence is indicated. (Some CRISPR loci encode indistinguishable psiRNAs.) Repeat sequence is shown in bold and the sequences are aligned by the 5' repeat sequences. The assignment of sequences (TT) shaded in grey (and corresponding clone lengths) is ambiguous due to the addition of polyA in the cloning procedure. One rRNA fragment was also cloned (not shown).

	psiRNA sequence	length	psiRNA
Upper Band	**ATTGAAAG**TTAGCAAATTGCCGATTATTGCACATAAAAAAAATA**G**	45	1.14
	ATTGAAAGTTAGCAAATTGCCGATTATTGCACATAAAAAAAATA**G**	45	1.14
	ATTGAAAGACTGGATTGAGAGCAACTTGTCGAATTATGTCGTCAA	45	1.40
	AATTGAAAGTGTTCATCAGCACTTCTTCTTCTGACTCTGCTCC	43	2.01
	AATTGAAAGTGTTCATCGCACTTCTTCTTCTGACTCTGCTCC	42	2.01
	ATTGAAAGCTAATTTACGCTTTAGCTCGTGATCAACCCTAATC	43	2.19
	ATTGAAAGCTAATTTACGCTTTAGCTCGTGATCAACCC	38	2.19
	ATTGAAAGTTGAGTTGAAGCGCCACTCTTTGAAGCCTATCAGA**GT**	45	4.02
	ATTGAAAGGCTTCAGGTCTTCAATATTCAATCCCGGTCCCTTTCA	45	4.03
	ATTGAAAGGCTTCAGGTCTTCAATATTCAATCCCGGTCCCTTTCA	45	4.03
	TTGAAAGTCTCTACCCTTACAAGCTTCTCGAATCTATCGAATTC	44	5.09
	TTGAAAGGTCACGTAATTCGCCAAGTTCTCTTGGATACCGTTC	43	5.12
	ATTGAAAGGTGGATAATATAATCCCTGTTTTTCCCAAGA	39	5.13
	ATTGAAAGTGGAACTCTATCAAGGTTTGCAACACCTTGCTCCCGC	45	5.24
	TTGAAAGGACAAAGAACTCCCTAGCGTCCCTCCCCGTGTA	38	6.07
	ATTGAAAGTGGGGTCTCGTCGCAATCGGTGCAGTATTCCTAAGCC	45	6.26
	ATTGAAAGATCTCCATCATACCAATGCTGTGCAAAATCAATCTT**G**	45	6.40
	ATTGAAAGTAAACTTAAGCTGGGATGGGCTATATACAAAGACAGA	45	6.42
	AATTGAAAGTCAAGAGTTCTATCCTGCTTCACAACACCCATATAA	46	7.11
	ATTGAAAGGCGTTAATGAACAATAAGCCTGACACGAACATAAA	43	1.06, 2.02
Bottom Band	**ATTGAAAG**CCGGTTCTGCACCCGAAACTTTCATACCAAA	39	1.03
	AATTGAAAGCCGGTTCTGCACCCGAAACTTTCATACCAA	39	1.03
	ATTGAAAGGTAGTGAGGCGTTGAACTTGACCCACCACCA	39	1.08
	ATTGAAAGTGAGTTGTTTAGTCTAACTCTTACACCATC	38	1.19
	ATTGAAAGTGCGCTATTCTCGGGTCAAGCCTCCCAGCCT	39	1.22
	TTGAAAGCACCACCACGATGAAGGTACCGTTTTCAAC	37	1.37
	TTGAAAGCACCACCACGATGAAGGTACCGTTTTCAAC	37	1.37

	ATTGAAAGTGTTCATCGCACTTCTTCTTCTGAC	33	2.01
	ATTGAAAGCTTCTTCGAAGTCGTAGTTTAGTGTGTCAAG	39	2.05
	ATTGAAAGTTCTAGAAGTTCTCTTGCGAGAGCCAGGAGC	39	2.06
	TTGAAAGCTAATTTATGCTTTAGCTCGTGATCAACCCTA	39	2.19
	TTGAAAGCTAATTTACGCTTTAGCTCGTGATCAACC	36	2.19
	AGGAATGTTGCTCAATGCAAAGGGCTCACCGCT	33	4.01
	AAAGTCTCAATTGGGGAGTGCTTTAATGGCTTTT	34	4.12
	ATTGAAAGGGAACTCCTCGATTTTAGTACCTGTGTC	36	5.05
	ATTGAAAGCCACATAAGACATTGTCATACAAAGTAGG	37	5.11
	ATTGAAAGGTCACGTAATTCGCCAAGTCCTCTTGGAGA	38	5.12
	ATTGAAAGGTGGATAATATAATCCCTGTTTTTCCCAAGA	39	5.13
	ATTGAAAGGTGGATAATATAATCCCTGTTTTTCCCAAGA	39	5.13
	ATTGAAAGGTGGATAATATAATCCCTGTTTTTCCCAAGA	39	5.13
	ATTGAAAGGTGGATAATATAATCCCTGTTTTTCCCAAGA	39	5.13
	ATTGAAAGCAGTTCTACTTTGATAAGACTGTGGTGGTTA	39	6.03
	ATTGAAAGGACAAAGAACTCCCTAGCGTCCCTCCCCGTG	39	6.07
	AATTGAAAGTTCTGCCGTCCCTTTCTCGACGAACCTCAT	39	6.09
	ATTGAAAGGCACCTTCTTCACCATCGCCGTCTGGATTGC	39	6.14
	AGTTGTAGGCTCGTGGACTTGGCTTCCACACAACTA	36	6.24
	ATTGAAAGTATCTATTGTACAGGTACTTGTTACACGT	37	7.14
	ATTGAA GCTTGCCCAACCTCTCTAGAAACGCCCAC	35	8.06
	ATTGAAAGGCGTTAATGAACAATAAGCCTGACACGAAC	38	1.06, 2.02
	ATTGAAAGTAATCTCAATAACTTTGGCTTCTTTTCTGTG	39	4.14, 5.19
	ATTGAAAGTAATCTCAATAACTTTGGCTTCTTTTCTGTG	39	4.14, 5.19

Figure 3.3. Specific cleavage of complementary target RNAs. The indicated 5' end-labeled substrates were incubated in the presence (+) or absence (-) of the native psiRNPs (Figure 3.1C). Products were resolved by denaturing gel electrophoresis. The primary cleavage products are indicated by green and blue arrows in panel A, and the corresponding sites of cleavage are indicated with green (site 1) and blue (site 2) vertical lines in the substrate sequences shown in panel B. Non-contiguous lanes from the same gel are indicated by dashed lines, and the sizes of RNA markers (M) are indicated in panel A. "Target" substrates (1, 3, 4, 5, 6) contain regions of perfect complementarity to the guide sequence of the indicated *Pyrococcus furiosus* psiRNA. Grey bars demarcate the guide sequences in the panel B. "+ ext" substrates (1, 2, 3, 4, 5) contain 5' and 3' polyA extensions. For substrate 5, a synthetic psiRNA (sequence shown in grey) was pre-annealed to the 7.01 target RNA + ext. Substrate 2 is a reverse target sequence substrate and substrate 7 is an antisense target substrate. Substrate 3 is DNA; all other substrates are RNA. Substrate 8 is unrelated RNA sR2. See Figure 3.4 for further characterization of the cleavage activity.

A



B

#	Substrate	Sequence (5' - 3')
1	7.01 target + ext	AAAAAAA CUGAAGUGCUCUCA | GCCGCA | AGGACCGCAUACUACAA AAAAAA
2	Reverse 7.01 target + ext	AAAAAAA AACAUCAUACGCCAGGAACGCCGACUCUCGUGAAGUC AAAAAA
3	DNA target + ext	AAAAAAA CTGAAGUGCTCTCAGCCGCAAGGACCGCATACTACAA AAAAAA
4	6.01 target + ext	AAAAAAA GUUCCACUAAGGAC | AUUUGU | ACGUCAAAUUCUUCACU AAAAAA
5	Annealed 7.01 target + ext	3′ GACUUCACGAGAGUCGGCGUUCCUGGCGUAUGAUGUU 5′ AAAAAAA CUGAAGUGCUCUCAGCCGCAAGGACCGCAUACUACAA AAAAAA
6	7.01 target	CUGAAGUGCUCUCA | GCCGCA | AGGACCGCAUACUACAA
7	Antisense 7.01 target	UUGUAGUAUGCGGUCCUUGCGGCUGAGAGCACUUCAG
8	sR2 C/D RNA	GGGGATGATGAGTTTTTCCCTCACTCTGATTAGTGATGAGGAGCCGATGCACTGACC

Site 1 Site 2

Figure 3.4. **Characterization of the activity of the psiRNA-Cmr protein complex, related to Figure 3.3. A)** Determination of the end groups of the cleavage products generated by the psiRNA-Cmr complex. 5' end-labeled target RNA (- psiRNPs) and psiRNA-Cmr complex cleavage products (+ psiRNPs) were subject to polyadenylation by *Escherichia coli* polyA polymerase (+) or no treatment (–). The position of the target RNA and polyadenylated target RNA are indicated. Arrows indicate the locations of the cleavage products. No polyadenylation of the cleavage products was observed, indicating the lack of 3' hydroxy groups on these RNAs. **B)** Divalent metal dependence of the psiRNA-Cmr complex. Cleavage reactions were performed in the presence of 0.1 mM EDTA and 1 mM of the indicated divalent metals or no metal (–). Non-contiguous lanes from a second gel are indicated by a dashed line. The cleavage products are indicated by arrows.

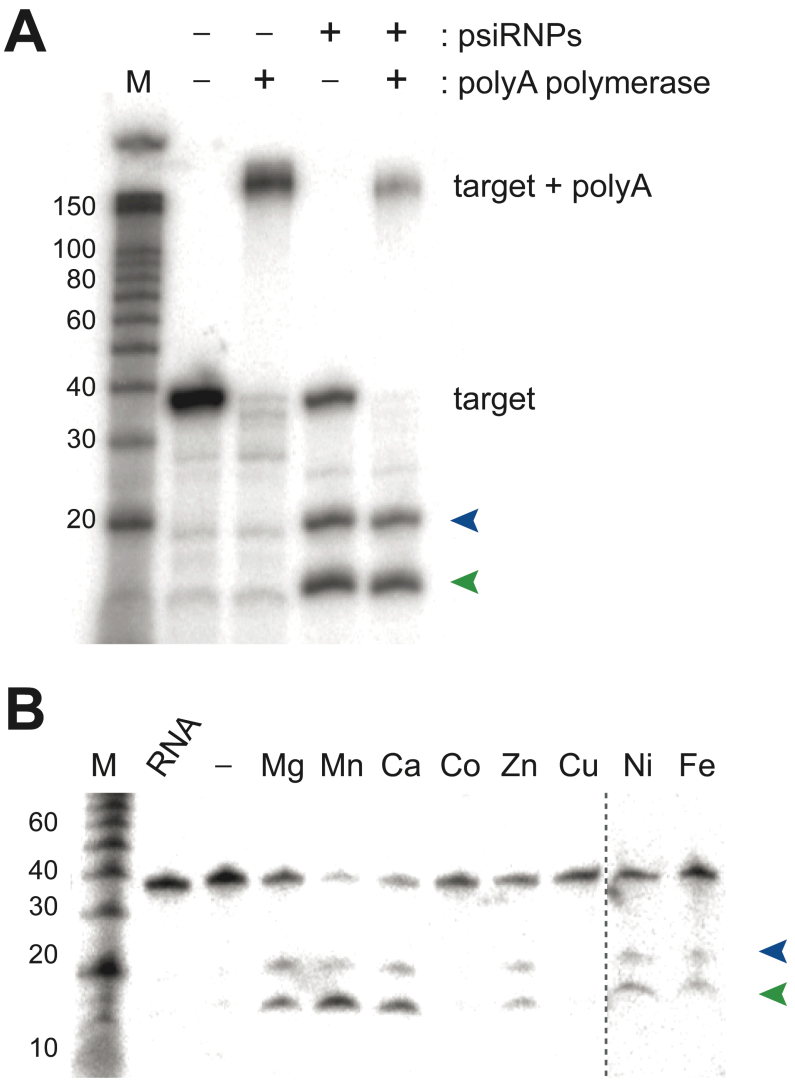
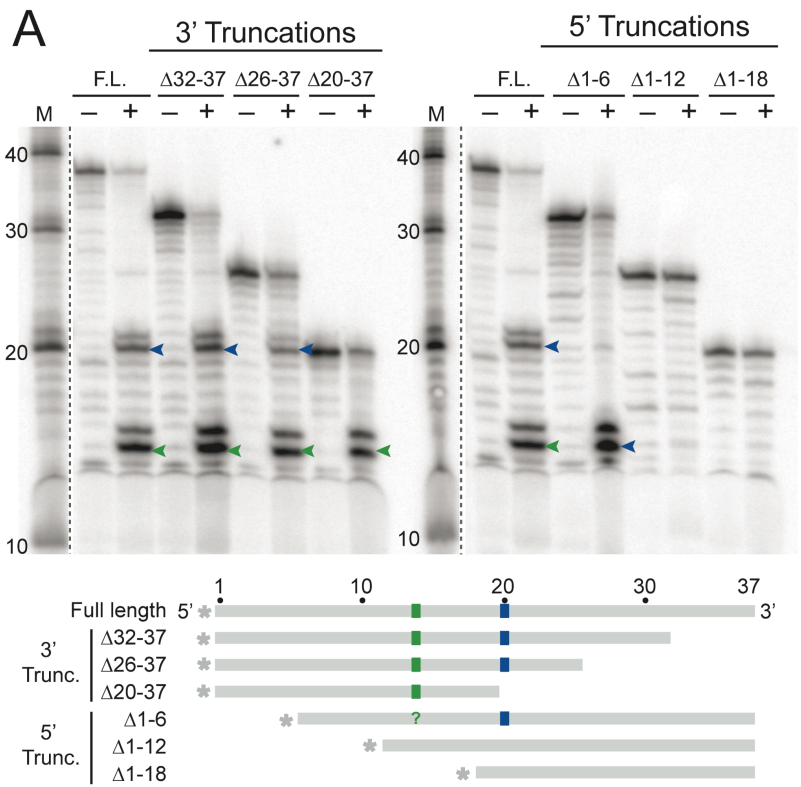

A
– – + + : psiRNPs
M – + – + : polyA polymerase
target + polyA
target
150
100
80
60
40
30
20
B
M RNA – Mg Mn Ca Co Zn Cu Ni Fe
60
40
30
20
10

Figure 3.5. Cleavages occurs 14 nucleotides from the 3' ends of the psiRNAs. A) The indicated 5' end-labeled (*) substrates were incubated in the presence (+) or absence (-) of the psiRNP (Figure 3.1C). The substrates were full-length 7.01 target RNA (F.L.), and the indicated truncations, and are diagramed below the gel. As in Figure 3.3, the locations of observed cleavages at sites 1 (green) and 2 (blue) are indicated on the diagrams of the substrate RNAs and the corresponding cleavage products are indicated with green and blue arrows on the gel. The question mark on the diagram of the Δ1-6 target RNA indicates that this cleavage could not be assessed. Non-contiguous lanes from the same gel are indicated by dashed lines. B) Model for cleavage at two sites directed by two psiRNAs. The 45-nucleotide psiRNA species guides cleavage at site 1 and the 39-nucleotide psiRNA guides cleavage at site 2 on each of the substrate RNAs as indicated. In both cases, cleavage occurs 14 nucleotides from the 3' end of the psiRNA. Observed products are shown in green (site 1) and blue (site 2) and correspond to products in Figures 3.3 and 3.5A.


A
3' Truncations
5' Truncations
F.L.
Δ32-37
Δ26-37
Δ20-37
Δ1-6
Δ1-12
Δ1-18
M
40
30
20
10
Full length 5'
3'
3' Trunc.
5' Trunc.
1
10
20
30
37
?

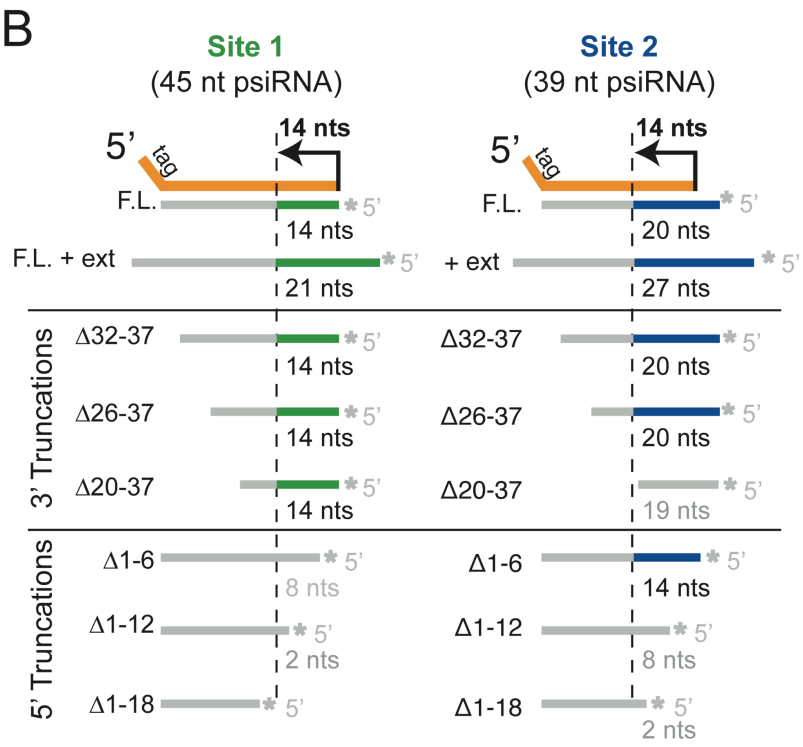

B
Site 1
(45 nt psiRNA)
Site 2
(39 nt psiRNA)
5'
tag
14 nts
F.L.
F.L. + ext
+ ext
21 nts
20 nts
27 nts
3' Truncations
5' Truncations
Δ32-37
Δ26-37
Δ20-37
Δ1-6
Δ1-12
Δ1-18
8 nts
2 nts
19 nts

Figure 3.6. Target RNA cleavage requires five Cmr proteins and a single psiRNA species. A) 5' end-labeled 7.01 target RNA was incubated in the absence of added psiRNAs or proteins (-), in the presence of synthetic psiRNAs (+) or purified recombinant *Pyrococcus furiosus* Cmr proteins (R), or in the presence of purified native psiRNPs (N) as indicated. The synthetic psiRNAs were 45- and 39-nucleotide forms of psiRNA 7.01. The six added recombinant Cmr proteins were *Pyrococcus furiosus* Cmr1-1, Cmr2, Cmr3, Cmr4, Cmr5 and Cmr6. Products were resolved by denaturing gel electrophoresis. The products corresponding to cleavage at site 1 and site 2 (see Figure 3.3) are indicated by green and blue arrows, respectively. Non-contiguous lanes from the same gel are indicated by a dashed line. The sizes of RNA markers (M) are indicated. B) The 7.01 target RNA (Target) was incubated with the synthetic 7.01 psiRNAs (both 45- and 39-nucleotide species) in the absence (+ psiRNAs) and presence of the purified recombinant *Pyrococcus furiosus* Cmr proteins (all), and also with combinations of proteins lacking individual Cmr proteins as indicated (e.g. - Cmr6). C) Cleavage activity of the recombinant psiRNP (R) reconstituted with either the individual 7.01 psiRNA species (45- or 39-nt) or both. Cleavage by the native psiRNP (N) is included for comparison.

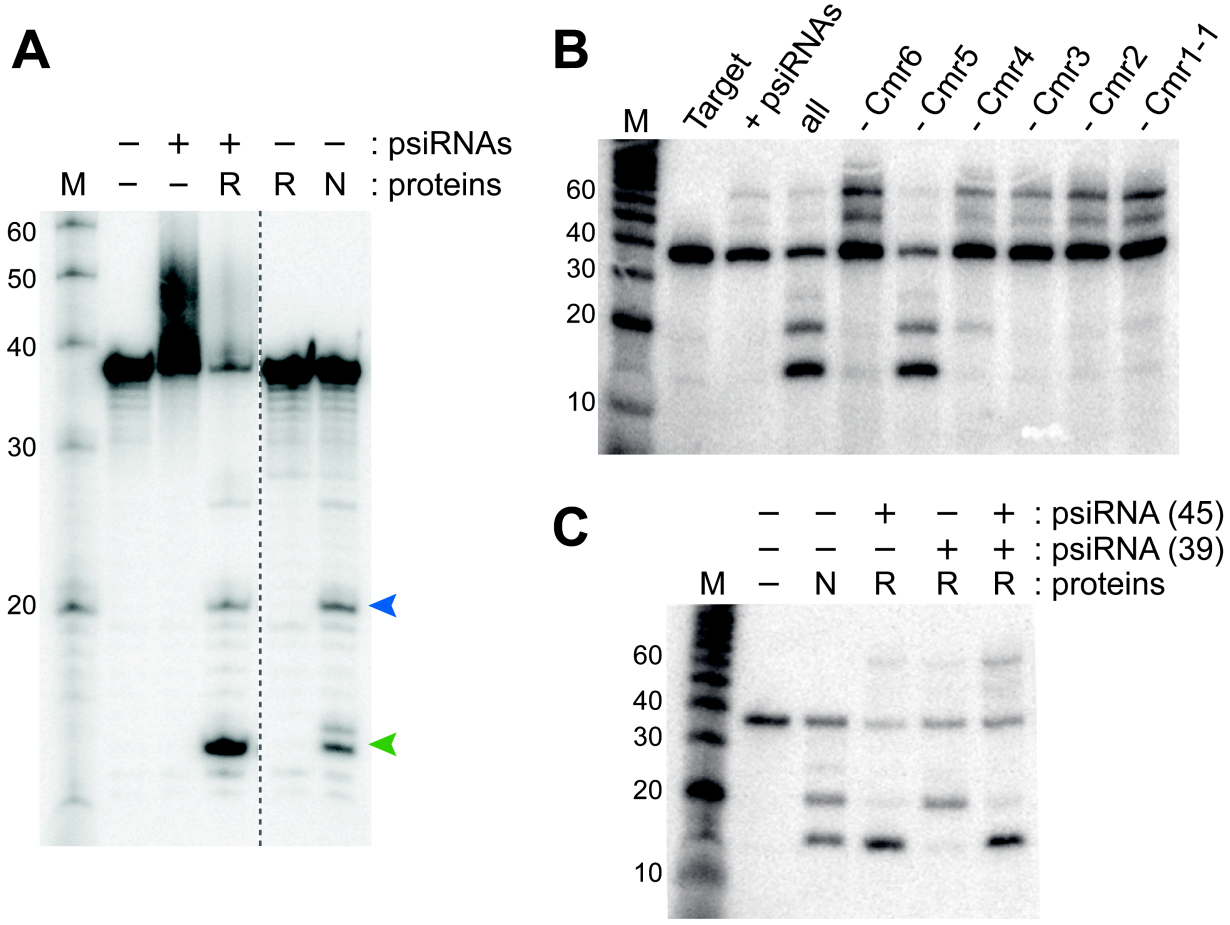
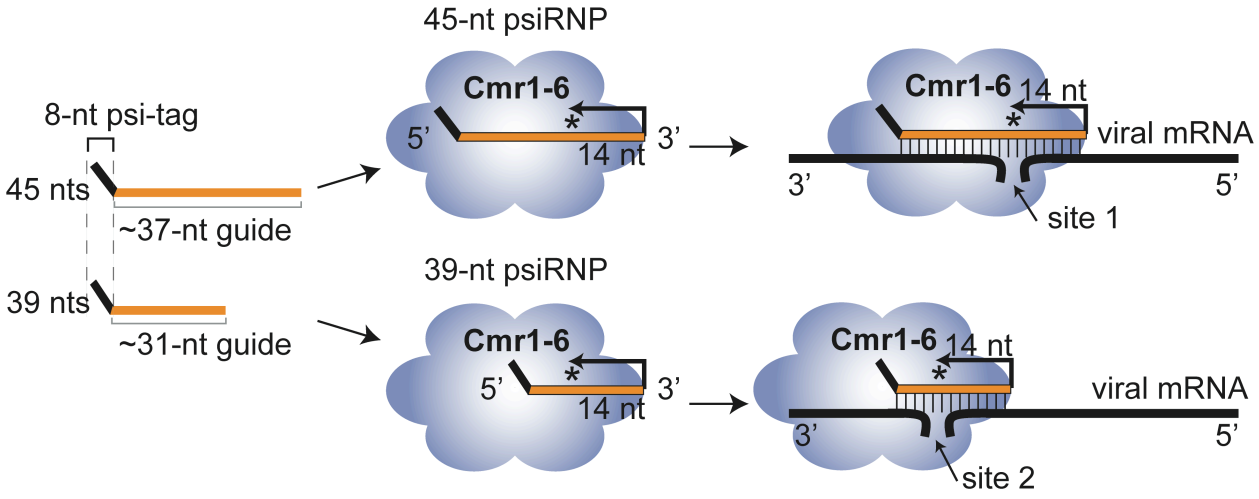

A
– + + – – : psiRNAs
M – – R R N : proteins
60
50
40
30
20
B
M
Target
+ psiRNAs
all
- Cmr6
- Cmr5
- Cmr4
- Cmr3
- Cmr2
- Cmr1-1
60
40
30
20
10
C
– – + – + : psiRNA (45)
– – – + + : psiRNA (39)
M – N R R R : proteins
60
40
30
20
10

Figure 3.7. Model for the function of psiRNA-Cmr protein complexes in silencing molecular invaders. Based on the results of this study, a psiRNA with a conserved 5' sequence element derived from the CRISPR repeat (psi-tag) and a region of invader-targeting sequence assembles with six Cas module-RAMP proteins (Cmr1-6). The assembled psiRNP interacts with an invader RNA (e.g., viral mRNA) through base pairing between the psiRNA and invader RNA, positioning the region of the RNA-RNA duplex 14 nucleotides from the 3' end of the psiRNA in proximity to the active site (star) of the enzyme. In *Pyrococcus furiosus*, there are two prominent size forms of each psiRNA with different 3' ends that guide cleavage of viral mRNAs at two distinct sites. There are also two Cmr1 proteins in *Pyrococcus furiosus* that are both found in purified preparations and likely function redundantly.


45-nt psiRNP
8-nt psi-tag
45 nts
~37-nt guide
Cmr1-6
5’
*
14 nt
3’
Cmr1-6
14 nt
viral mRNA
3’
5’
site 1
39-nt psiRNP
39 nts
~31-nt guide
Cmr1-6
5’
*
14 nt
3’
Cmr1-6
14 nt
viral mRNA
3’
5’
site 2

CHAPTER 4

DISCUSSION

The CRISPR/Cas system is a recently discovered genome defense system in prokaryotic organisms (reviewed in (Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010a; Sorek et al., 2008; van der Oost et al., 2009)). In this system, short sequences from invader nucleic acids such as virus, phages and transposons are inserted into a CRISPR locus upon infection. The presence of these sequences confers immunity from future infections by the corresponding invader. Little is known about the biochemical mechanisms involved in this system. In this work, we sought to understand the components of the CRISPR/Cas system in *Pyrococcus furiosus*, a hyperthermophilic archeaon. In particular, we sought to understand the nature of the small RNAs that are transcribed from the CRISPR loci, and the role of psiRNA-Cas protein complexes in the interference stage of the CRISPR/Cas system of genome defense. Figure 4.1 outlines the discoveries described in this work. In short, we found that the CRISPRs of *Pyrococcus furiosus* are constitutively and unidirectionally transcribed from the leader sequence, and processed into short psiRNAs that contain a full spacer sequence and 8 nucleotides of the 5' repeat, which is referred to as the psi-tag. At least a fraction of these psiRNAs are incorporated into a Cmr protein-containing complex, which has the ability to sequence-specifically recognize and cleave complementary target RNAs (Figure 4.1).

Expression of CRISPR loci

In our investigation of the expression of CRISPR loci in *Pyrococcus furiosus*, we observed at least some level of constitutive expression of the CRISPRs. This indicates that expression is not dependent on the presence of an invader, although there may be some level of up-regulation upon infection or other types of stress. We also observed a preference for the accumulation of leader-proximal spacer-derived psiRNAs. It was previously observed that insertion of new spacers into the CRISPRs occurs at the leader sequence (Barrangou et al., 2007; Pourcel et al., 2005). We observed that the psiRNAs directed against the most recent invaders are the most abundant.

We observed unidirectional transcription from the leader portion of the CRISPR locus (Figure 2.3). Unidirectional transcription was also seen in *Archaeoglobus fulgidus* (Tang et al., 2002*)*, *Solfolobus solfataricus* (Tang et al., 2005), *Eschericia coli* (Brouns et al., 2008) and *Staphylococcus epidermis* (Marraffini and Sontheimer, 2010b). One study, which more extensively studied transcription of the CRISPR loci in *Sulfolobus solfataricus*, found abundant transcription from both strands of the CRISPR locus (Lillestol et al., 2006; Lillestol et al., 2009). RNAs from both strands incorporated the entire locus, and both seemed to be processed into small RNAs, although the leader strand RNAs were smaller (about 40 nts) than the complement strand RNAs (about 55 nts). These RNAs were found to be about equal in abundance, and were not found to form dsRNA with the leader strand spacers, which may indicate presence in a distinct RNA-protein complex (Lillestol et al., 2009). It is possible that transcription from the complementary strand occurs only under particular growth conditions, such as stress, and was therefore not detected in the *Pyrococcus furiosus* total RNA that was tested. It

is also possible that this is yet another example of diversity in the mechanism of action of CRISPR/Cas systems in various organisms, and only certain organisms under certain circumstances may produce antisense psiRNAs.

The Cmr complex

It is predicted that mature psiRNAs form RNPs with Cas proteins. In Chapter 2, we show here that at least four distinct psiRNA-containing complexes are found in *Pyrococcus furiosus* cell extract (Figure 2.4), including two (A and B) that contained the mature-sized psiRNAs. Further purification of one of these complexes (A) in Chapter 3 revealed that it contained a complete set of the RAMP module proteins, designated Cmr1-6, as described by Haft et al. (Haft et al., 2005). The *Pyrococcus furiosus* genome contains two copies of the *cmr1* gene, and proteins from both genes were identified. We presume that these two proteins are functionally redundant, and indeed, only Cmr1-1 was needed for the reconstituted activity. However, it will be important in the future to understand whether Cmr1-2 can substitute for Cmr1-1, or whether it plays an additional role in the Cmr complex *in vivo*.

We show here that five of the six Cmr proteins (Cmr1,2,3,4, and 6) are required for cleavage activity. However, we do not currently understand the role that each plays in the cleavage activity. Cmr2 contains an N-terminal HD nuclease-like domain (Makarova et al., 2002) and may be responsible for the actual cleavage event. However, this domain is not well conserved among Cmr2 proteins, whereas the polymerase-like thumb and palm domains are much more highly conserved, which may implicate this protein in a more polymerase-like activity, although no such activity has been observed to date. Four of the Cmr proteins, (1,3,4 and 6) are predicted to be

RAMPs, or Repeat Associated Mysterious Proteins (Haft et al., 2005; Makarova et al., 2006). This family of proteins has a conserved ferredoxin-like fold and a G-rich C-terminal loop, and is predicted to contain a family of RNA-binding proteins. However, several members of this family have been seen to possess nuclease activity, including the core protein Cas6 and the Ecoli subtype specific protein Cse3 (Brouns et al., 2008; Carte et al., 2008; Makarova et al., 2006). It is therefore possible that one or more of the RAMP proteins present in the complex are responsible for the sequence-specific cleavage activity.

The psi-tag

Cloning and sequencing of the RNAs found in the Cmr complex revealed the presence of two forms of psiRNAs, both having 8 nucleotides of repeat on the 5' end and either a full invader-targeting sequence, or about 6 nucleotides less than a full invader-targeting sequence. We refer to the 5' 8 nucleotide sequence as the psi-tag. Deep sequencing revealed that the psi-tag is seen on the majority of psiRNAs in *Pyrococcus furiosus* total RNA (Figure 3.2). A similar sequence has been seen in *Escherichia coli* (Brouns et al., 2008) and *Staphylococcus epidermis* (Marraffini and Sontheimer, 2010a). The generation of the psi-tag is mediated by Cas6 in *Pyrococcus furiosus* and Cse3 in *Escherichia coli* (Brouns et al., 2008; Carte et al., 2008).

We predict that the psi-tag is an important recognition signal for recognition of psiRNAs by the Cas proteins. It will be important to understand if the psi-tag is necessary and/or sufficient for targeting the CRISPR system against nucleic acids. This information could greatly benefit the world of prokaryotic biology. For example, introducing artificial, tagged psiRNAs into the cell may allow researchers to target and

sequence-specifically degrade any cellular RNA, which would create an RNAi-like system in prokaryotes, the benefits of which would be numerous. In addition, the CRISPR system may also be able to be used to create novel antibiotics that would turn the cell's own defense system against itself by expressing tagged psiRNAs that target important cellular genes, such as ribosomal RNAs.

Additionally, it has recently been shown that the psi-tag is responsible for distinguishing between foreign DNA and the genomic CRISPR locus (Marraffini and Sontheimer, 2010b). In *Staphylococcus epidermis*, it was seen that mutation of the proto-spacer of a psiRNA-targeted plasmid in a manner that allows complementarity to the tag sequence eliminates resistance to that plasmid. This psi-tag-mediated property protects the DNA in the CRISPR locus from being targeted by the CRISPR machinery (Marraffini and Sontheimer, 2010b).

Comparison of the CRISPR/Cas and eukaryotic RNAi systems

The CRISPR/Cas system was originally proposed to be the functional analog of the eukaryotic RNAi pathway in prokaryotes. There is no sequence similarity between any of the Cas proteins and any of the proteins involved the eukaryotic RNAi pathways. However, the discovery of an RNA targeting Cas protein-psiRNA complex brings to light functional similarities, especially between the CRISPR/Cas system and the piRNA pathway. The canonical RNAi pathway begins with the processing of dsRNA into siRNAs by the RNase III-like Dicer protein (Elbashir et al., 2001; Song and Joshua-Tor, 2006). However, biogenesis of piRNAs has been found to be Dicer-independent. Instead, it appears that there are piRNA clusters within the genome that are transcribed as long transcripts and processed down to shorter RNAs by an unknown pathway

(Hartig et al., 2007; Klattenhoff and Theurkauf, 2008). This model bears a striking resemblance to the CRISPR/Cas expression system, where the locus is transcribed and processed into psiRNAs. The major difference in the piRNA pathway is the absence of repeats, which are extremely important in the processing of psiRNAs in the CRISPR/Cas system.

Once processed, piRNAs are bound by Argonaute proteins Piwi and Aubergine. The resulting complex is able to recognize and cleave complementary RNAs. In the *Pyrococcus furiosus* CRISPR/Cas system, it appears that mature psiRNAs are sorted by an unknown mechanism in at least two distinct Cas protein-containing complexes, one of which contains the RAMP module proteins. In contrast to the piRNA pathway, psiRNAs are bound by multi-protein complexes, which may not all guide RNA cleavage.

In this work, we describe single-stranded RNA cleavage that occurs by a ruler mechanism that counts 14 nucleotides from the 3' end of the psiRNA. The Slicer activity of the Argonaute proteins involved in the RNAi pathway also utilizes a ruler-like mechanism, but counts from the 5' end of the siRNA instead of the 3' end, as seen by the Cmr complex.

RNA vs. DNA-targeting by the CRISPR/Cas system

This work presents the first evidence of RNA targeting by the CRISPR/Cas system. The Cmr proteins are present in numerous CRISPR-containing organisms, including both archaea and bacteria. Studies in two bacterial organisms have shown evidence of DNA targeting. However, neither of these organisms contain the Cmr proteins. It is likely then, that this gene module encodes components of an RNA-targeting complex in a diverse range of organisms. It is possible that the "subtype

specific" genes encode proteins that allow DNA-targeting, other systems for RNA-targeting, or other silencing mechanisms. This would allow some organisms to have multiple lines of defense against invaders. As mentioned previously, the diversity of the components of the CRISPR/Cas system among organisms is remarkable. Therefore, only increased research into various CRISPR-containing organisms will elucidate the true nature of the complex in each system.

Future Directions/Applications

This work presents *in vitro* evidence of a Cmr protein-containing complex. It will be important in the future to confirm this activity *in vivo*, and also to understand if any other proteins (Cas or non-Cas) are required for viral targeting. Although none of the non-Cas protein identified by mass spectrometry in this work were necessary for RNA targeting *in vitro*, they may play important roles in the cell, and this needs to be elucidated.

As mentioned above, there is much more to understand about the role of the psi-tag in the cleavage activity, as this could potentially be extremely useful in manipulation of the CRISPR/Cas system for molecular biology purposes, and could have potential industrial or clinical applications.

One major application for the CRISPR/Cas system is in the dairy industry, where microbial cultures are very carefully maintained. Much attention has been spent on keeping bacteriophages from infecting the many strains of bacteria that are used for culturing milk, yogurts and cheeses. Artificial CRISPR loci may be able to be inserted into these organisms to allow for a better-protected microbe.

In conclusion, it seems that we have merely scraped the surface of fully understanding the mechanisms involved in the CRISPR/Cas system. This work revealed a great deal about the mechanism of expression and interference in Pyrococcus furiosus, but there is still much to be learned about the mechanisms in other organisms, which contain much different complements of Cas proteins.

References

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.

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.

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.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev *15*, 188-200.

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.

Hartig, J.V., Tomari, Y., and Forstemann, K. (2007). piRNAs--the ancient hunters of genome invaders. Genes Dev *21*, 1707-1713.

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

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

Klattenhoff, C., and Theurkauf, W. (2008). Biogenesis and germline functions of piRNAs. Development *135*, 3-9.

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.

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

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.

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

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

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.

Song, J.J., and Joshua-Tor, L. (2006). Argonaute and RNA--getting into the groove. Curr Opin Struct Biol *16*, 5-11.

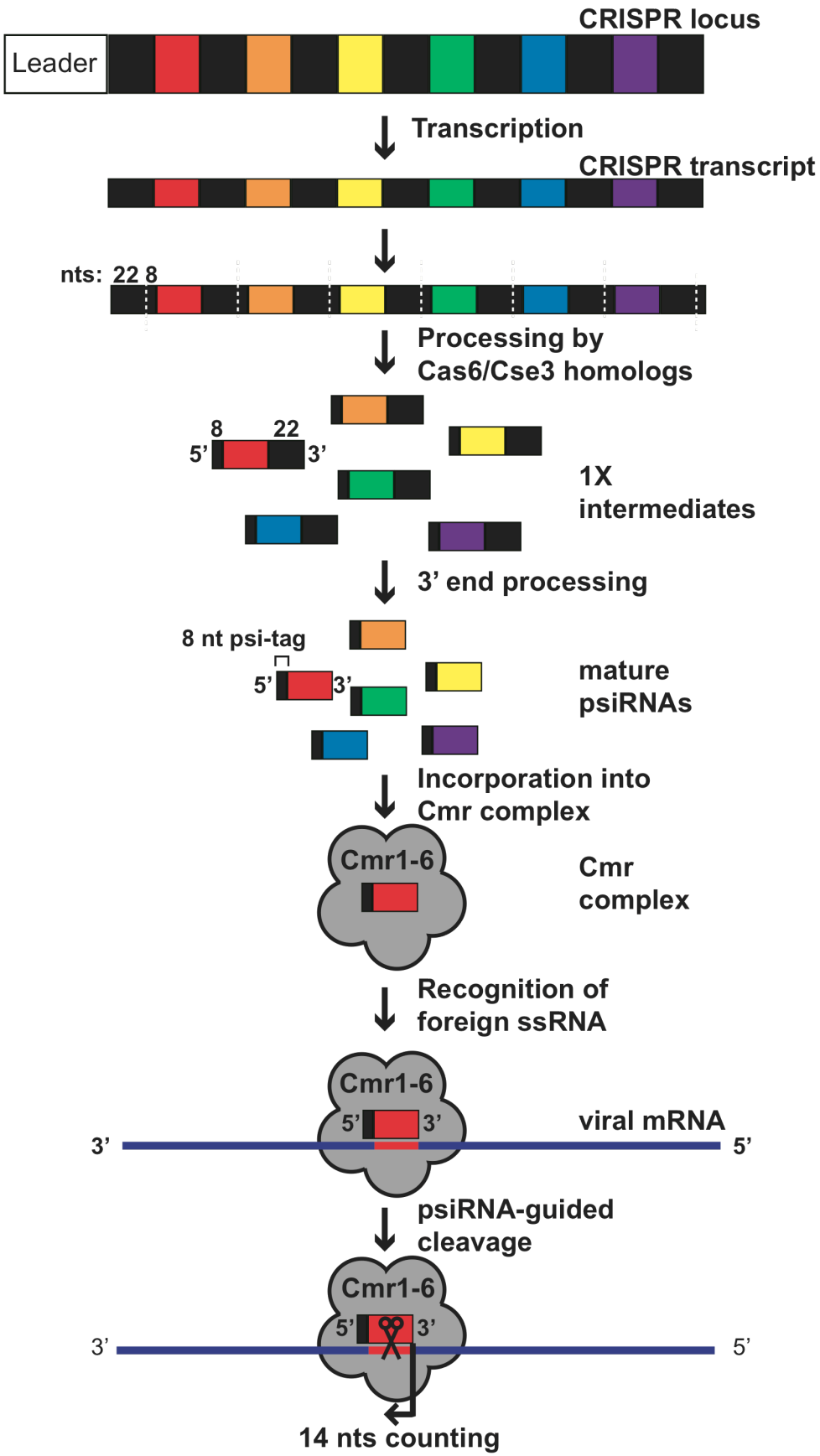
Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol *6*, 181-186.

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.

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.

Figure 4.1. **Overview of findings in this work**. In *Pyrococcus furiosus*, the CRISPR locus is transcribed as a long transcript and processed into 1X intermediates by Cas6. These intermediates contain a full spacer sequence, 8 nucleotides of repeat on the 5' end (dubbed the psi-tag), and 22 nucleotides of repeat on the 3' end. The 1X intermediates undergo 3' processing by an unknown protein, generating a mature psiRNA which contains a full repeat sequence and the psi-tag sequence. At least a fraction of these psiRNAs are incorporated into the Cmr complex, which contains Cmr1-6. This RNA protein complex has the ability to recogize and cleave single stranded RNAs that contain sequences that are complementary to the psiRNA. This cleavage was shown to occur by a ruler mechanism, which counts 14 nucleotides from the 3' end of the incorporated psiRNA.


CRISPR locus
Leader
Transcription
CRISPR transcript
nts: 22 8
Processing by
Cas6/Cse3 homologs
8
22
5'
3'
1X
intermediates
3' end processing
8 nt psi-tag
5'
3'
mature
psiRNAs
Incorporation into
Cmr complex
Cmr1-6
Cmr
complex
Recognition of
foreign ssRNA
Cmr1-6
5'
3'
viral mRNA
3'
5'
psiRNA-guided
cleavage
Cmr1-6
5'
3'
3'
5'
14 nts counting

APPENDIX A

FORMATION OF THE CONSERVED PSEUDOURIDINE AT POSITION 55 IN ARCHAEAL tRNA[1]

[1] Roovers, M.*, Hale, C.*, et al. (2006). "Formation of the conserved pseudouridine at position 55 in archaeal tRNA" *Nucleic Acids Research*. 34(15): 4293:301. Reprinted with permission. *authors contributed equally to this work

Abstract

Pseudouridine (Ψ) located at position 55 in tRNA is a nearly universally conserved RNA modification found in all three domains of life. This modification is catalyzed by TruB in bacteria and by Pus4 in eukaryotes, but so far the Ψ55 synthase has not been identified in archaea. In this work, we report the ability of two distinct pseudouridine synthases from the hyperthermophilic archaeon *Pyrococcus furiosus* to specifically modify U55 in tRNA *in vitro*. These enzymes are $_{pfu}$Cbf5, a protein known to play a role in RNA-guided modification of rRNA, and $_{pfu}$PsuX, a previously uncharacterized enzyme that is not a member of the TruB/Pus4/Cbf5 family of pseudouridine synthases. $_{pfu}$PsuX is hereafter renamed $_{pfu}$Pus10. As would be expected for an enzyme that serves to modify U55 in archaeal tRNAs, orthologs of both $_{pfu}$Cbf5 and $_{pfu}$Pus10 are predicted in all archaeal organisms whose genomes have been sequenced. Both enzymes specifically modify tRNA U55 *in vitro* but exhibit differences in substrate recognition. In addition, we find that in a heterologous *in vivo* system, $_{pfu}$Pus10 efficiently complements an *Escherichia coli* strain deficient in the bacterial Ψ55 synthase TruB. These results indicate that it is likely that $_{pfu}$Cbf5 or $_{pfu}$Pus10 (or both) is responsible for the introduction of pseudouridine at U55 in tRNAs in archaea. While we cannot unequivocally assign the function from our results, both possibilities represent unexpected functions of these proteins as discussed herein.

Introduction

Pseudouridine (Ψ) is the most abundant modified nucleoside in RNA. It has been identified in tRNAs, rRNAs, snRNAs and/or snoRNAs of all organisms that have been analyzed to date (Massenet et al., 1998; McCloskey and Rozenski, 2005;

Rozenski et al., 1999; Sprinzl and Vassilenko, 2005). Ψ is formed post-transcriptionally by enzymatic isomerization of uridine by a class of enzymes known as pseudouridine synthases. Most Ψ synthases have the ability to recognize and catalyze isomerization of one (or a few) unique target uridines (Ferre-D'Amare, 2003).

For many years, there were thought to be only four families of Ψ synthases, corresponding to TruA, TruB, RluA and RsuA in bacteria (Koonin, 1996). Members of these families were found in all 3 domains of life. Recently however, Ψ synthases that share little sequence homology to any of these families (such *Escherichia coli* TruD) have been found, leading to the recognition of additional groups of Ψ synthases, as well as the idea that there may be many more Ψ synthases that have yet to be discovered (Kaya and Ofengand, 2003).

Ψ55 in tRNA is a quasi-universally conserved modification found in all forms of life analyzed to date. The Ψ synthases responsible for this very highly conserved modification have been identified in eukaryotes and bacteria, and are members of the TruB/Pus4/Cbf5 family (Cluster of Orthologous Groups of proteins (COG) 0130 (Ofengand et al., 2001)). In the yeast *Saccharomyces cerevisiae*, Pus4 was shown to be the enzyme responsible for this modification (Becker et al., 1997a), and in *Escherichia coli*, TruB is the enzyme responsible for formation of Ψ55 (Nurse et al., 1995). Although Ψ55 is known to exist in archaeal tRNAs and the activity has been identified in *Pyrococcus furiosus* cell extract (Constantinesco et al., 1999), no archaeal Ψ55 synthase has been identified to date. The goal of the present work was to identify the enzyme responsible for isomerization of U55 to Ψ55 in *Pyrococcus furiosus*, our

model archaeal system. We have tested two candidate Ψ synthases, $_{pfu}$Cbf5 and $_{pfu}$Pus10 (previously known as PsuX).

The only identified archaeal protein within the TruB/Pus4/Cbf5 family was recently shown to function in RNA-guided pseudouridylation of rRNA (Baker et al., 2005). While most Ψ synthases are dedicated to modification of a single target uridine, in eukaryotes and archaea members of the Cbf5 sub-family modify multiple specific sites by a mechanism that depends on guide RNAs and a set of essential accessory proteins (Baker et al., 2005). (Thus in most eukaryotes, the paralogous proteins Pus4 and Cbf5 function in tRNA U55 and rRNA pseudouridylation, respectively). The functional complex is comprised of Cbf5 (the enzyme) plus Nop10, Gar1 and L7Ae (essential accessory proteins) and an H/ACA guide RNA (Baker et al., 2005). The guide RNA mediates substrate recognition by basepairing with nucleotides surrounding the target uridine. Multiple guide RNAs provide for pseudouridylation of numerous individual rRNA sites. Once recruited to the substrate rRNA by the guide RNA, the complex catalyzes formation of Ψ at the targeted uridine residue (Baker et al., 2005). There is no existing evidence that Cbf5 functions in the absence of a guide RNA and accessory proteins, or that Cbf5 is involved in pseudouridylation of tRNA.

The second candidate Ψ55 synthase that we tested has been called PsuX (Watanabe and Gray, 2000). The *psuX* gene (hereafter the *pus10* gene) was identified by sequence similarity to an internal region (catalytic domain; residues G66 to A271) of the *Euglena gracilis CBF5* gene (Watanabe and Gray, 2000). Pus10 (PF1139) belongs to a different Ψ synthase family than TruB, Pus4, and Cbf5 (COG 1258 for Pus10, and

COG 0130 for the three other Ψ synthases (Ofengand et al., 2001)) and its function and specificity are unknown.

In this work, the activities of recombinantly expressed and purified $_{pfu}$Cbf5 and $_{pfu}$Pus10 enzymes were characterized and compared in *in vitro* assays. The results suggest that both enzymes are capable of selective modification of tRNA U55 *in vitro* under similar reaction conditions. We found that $_{pfu}$Pus10 can pseudouridylate a truncated tRNA substrate (miniS) and a tRNA lacking the 3'-CCA, but that $_{pfu}$Cbf5 functions only on the full-length tRNA substrate. Moreover, $_{pfu}$Pus10 efficiently complements tRNA Ψ55 formation in an *Escherichia coli truB* knockout strain. These results suggest two novel and interesting candidates for the pseudouridylation of tRNA U55 in archaea.

Experimental procedures

Strains and plasmids.

The ML 100 strain (BL21(DE3) *truB*) was constructed by transduction of the BL21(DE3) strain (Novagen) by phage 363 lysates of the GOB113 strain (Bylund et al., 1998). The GOB113 strain (Hfr P4X *sdr*$^{+}$ *truB*2422::mini-Tn*10*Cm) carrying a mutation in the *truB* gene making this strain defective for TruB activity was kindly provided by Michael Wikström (University of Umeå, Sweden). The construction of the pET21d plasmid allowing T7-expression in *Escherichia coli* of the *Pyrococcus furiosus* Cbf5 protein ($_{pfu}$Cbf5; ORF=PF1785) and purification of the recombinant protein was described previously (Baker et al., 2005).

The plasmid pML156 allowing *in vitro* transcription of *Pyrococcus furiosus* tRNA^{Asp} by T7 RNA polymerase is a pUC18 derivate with the fragment containing the

Pyrococcus furiosus $tRNA^{Asp}$ gene downstream of a T7 promoter cloned in its *Bam*HI/*Hin*dIII sites. This plasmid was kindly provided by Sylvie Auxilien (CNRS, Gif-sur-Yvette, France). The corresponding *Pyrococcus furiosus* $tRNA^{Asp}$(U55C) mutant was obtained with the QuickChange site-directed mutagenesis kit (Stratagene) using pML156 as a template. The oligonucleotides used were 5'-CGACCCGGGTCCAAATCCCGG-3' and 5'-CCGGGATTTGGACCCGGGTCG-3'. The sequence of the mutant was verified.

Cloning the Pyrococcus furiosus pus10 gene, expression and purification of the corresponding $_{pfu}$Pus10 protein.

The gene encoding *Pyrococcus furiosus* Pus10 (ORF=PF1139), was amplified by PCR from genomic DNA and cloned into the *Nde*I/*Xho*I sites of pET28b. The primers used in this PCR reaction were 5'-GTCATATGATACTTGAAAAAGCCA GAG-3' and 5'-GTCTCGAGTCAATTATCTCC CTCAACATCG-3'. The resulting plasmid (pET28b-*pus10*) allowed T7-expression in *Escherichia coli* (strain Rosetta(DE3)) of the *Pyrococcus furiosus* Pus10 protein ($_{pfu}$Pus10) bearing an N-terminal His-tag. The transformed cells were grown to an optical density at 660 nm of 0.5, and the expression of the protein was induced with 1 mM isopropylthio-β-D-galactoside (IPTG) for 3 h at 37 °C. Cells were then harvested and resuspended in buffer A (20 mM Na phosphate pH 7.0, 1 M NaCl) and lyzed by a 30 min sonication at 4 °C using a Vibracell 75041 sonicator (40 % amplitude). The lysate was cleared by centrifugation (20,000 g for 30 min) and was applied to a column of Chelating Sepharose Fast Flow (1 x 30 cm; Amersham Biosciences) charged with Ni^{2+}. The column was washed with buffer A and adsorbed material was eluted with a linear gradient (from 0 to 1 M) of imidazole in buffer

A. Eluted fractions were analyzed by SDS-PAGE. Fractions containing the $_{pfu}$Pus10 protein were pooled, concentrated to 3 mg/ml in buffer containing 1M NaCl and 200 mM imidazole and kept at –80°C until used.

Preparation of radiolabeled tRNA substrates.

The general procedure for generating [^{32}P] radiolabeled tRNA transcripts as substrates for enzymatic reactions is based on the method described previously (Reyes and Abelson, 1987). Radioactive [^{32}P] *in vitro* transcripts were prepared as described elsewhere (Roovers et al., 2004) using *Mva*I (for generating tRNA^{Asp} substrate) or *Nar*I (for generating tRNA^{Asp}-3'CA substrate) digested plasmids as templates. α-^{32}P-Radiolabelled nucleotide triphosphates (400 Ci/mmol) were from ICN Biomedicals and T7 RNA polymerase was from Roche Diagnostics. Radioactive transcripts were purified by 10% polyacrylamide gel electrophoresis. The radioactive band corresponding to full length transcript was eluted from the gel with water and finally precipitated by ethanol in the presence of 0.3 M Na acetate pH 5. The *Pyrococcus furiosus* tRNA^{Asp} minisubstrate (TΨ-loop & stem - see Figure A2A) was prepared by *in vitro* transcription of a single stranded DNA template (minus stand), following the procedure described previously (Becker et al., 1997b).

tRNA pseudouridine synthase assay*.*

Purified recombinant $_{pfu}$Cbf5 (5 µg or 2 µg) or $_{pfu}$Pus10 (5 µg or 0.5 µg) was incubated with the *in vitro* transcribed tRNA [^{32}P]-labeled transcript for one hour at 70°C in buffer B (20 mM HEPES pH 7.0, 500 mM KCl, 1.5 mM $MgCl_2$). The reaction was terminated by the addition of an equal volume of phenol and the RNA recovered by ethanol precipitation. The modified tRNA was then treated and analyzed for its Ψ

content as described previously (Grosjean et al., 2004). In brief, each RNA sample was hydrolyzed by nuclease P1 or RNase T2 and the resulting nucleotides were analyzed by 1D-(20 cm) or 2D-(10 cm x 10 cm) thin layer chromatography (TLC) on cellulose plates (Merck). First dimension was developed with solvent A (isobutyric acid/concentrated NH_4OH/water; 66/1/33; v/v/v); the second dimension was developed with solvent B (concentrated HCl/isopropanol/water; 18/68/14; v/v/v). When only one dimension was performed as in Figure A3, to reduce background, the incubated tRNA was digested with RNase T1, and the fragments were separated on a 20% polyacrylamide gel and the fragment corresponding to the TΨ-loop was extracted from the gel, eluted, ethanol precipitated, and then subjected to nuclease P1 digestion and TLC, as outlined above, except using an acidic solvent for TLC (concentrated HCl/isopropanol/water; 15/70/15; v/v/v).

rRNA pseudouridine synthase assay.

Recombinant $_{pfu}$Cbf5 (2 µg) was incubated with 0.75 pmol unlabeled guide (Pf9) sRNA and 0.05 pmol UTP-[^{32}P]- labeled-target rRNA as described previously (Baker et al., 2005). For experiments with accessory proteins present, the additional proteins were added at stoichiometric amounts equivalent to 2 µg $_{pfu}$Cbf5. Incubation was at 70°C for one hour. Resulting RNA was phenol extracted, digested with nuclease P1 for one hour at 37˚C and run on a 1D (20 cm) TLC plate in acidic solvent (concentrated HCl/isopropanol/water; 15/70/15; v/v/v). The results were analyzed by autoradiography.

CMC/RT assay for mapping pseudouridine residues.

Total tRNA was isolated from *Escherichia coli* cells that were either BL21(DE3), ML100, or ML100 transformed with the pET plasmids containing respectively the *cbf5* or

pus10 genes. 100 ml of cells were grown to OD 0.4, then protein expression was induced with 0.05 mM IPTG for 3 h at 37°C. Cells were then rapidly transferred from 37 to 50°C and incubated for an additional 2 hours. tRNA was isolated according to (Ofengand et al., 2001). CMC modification was performed on 2 µg of total tRNA as described previously (Ofengand et al., 2001). The resulting modified tRNA samples were then subjected to primer extension using 5'-end labeled primers. Primer for $_{eco}$tRNA^{Cys} was 5' –TGGAGGCGCGTCCGG – 3'. Primer for $_{eco}$tRNA^{Phe} was: 5' – TGGTGCCCGGACTCGG – 3'. Primer (0.5 pmol) was annealed to tRNA in 50 mM Tris-Cl pH 8.6, 60 mM NaCl and 10 mM DTT for 3 min at 70˚C, 5 min at 37˚C and at least 2 min at 0˚C. Primer extension was performed using 2 µl of the above annealing reaction in a 5 µl reaction mixture containing 50 mM Tris-Cl pH 8.6, 60 mM NaCl, 10 mM DTT and 2.4 mM $MgCl_2$, 330 µM dNTPs and 20U reverse transcriptase (Promega). Reaction was carried out for 30 min at 37˚C. For sequencing reactions, 10 µg of unmodified wild type total tRNA was used in the same annealing reaction conditions as before, except using 1.5 pmol of labeled primer. Extension was performed in the same buffer as above, except with the addition of 1 mM individual ddNTPs in four different reactions. The resulting fragments were separated on a 15% sequencing acrylamide gel (7M urea). Results were analyzed by autoradiography.

Results

Both $_{pfu}$Cbf5 and $_{pfu}$Pus10 can form Ψ55 in tRNA in vitro

As described above, bioinformatic searches yielded two candidate enzymes for U55 modification in archaea. $_{pfu}$Cbf5 is a member of the TruB/Pus4/Cbf5 family of enzymes that are responsible for Ψ55 modification in bacteria and eukaryotes. $_{pfu}$Pus10

(previously PsuX) is a member of another family of pseudouridine synthases with unknown specificity. To test whether these enzymes are tRNA Ψ55 synthases, we cloned the *Pyrococcus furiosus cbf5* and *pus10* genes, expressed and purified the $_{pfu}$Cbf5 and $_{pfu}$Pus10 proteins and assayed their activity. Both enzymes catalyzed the formation of Ψ *in vitro* (Figure A1). More extensive tests were then performed to further analyze the activity of $_{pfu}$Cbf5 and $_{pfu}$Pus10. Both enzymes were found to work under the same reaction conditions, with the optimal activity observed at 1 hour at 70˚C (Figure A1C-F). $_{pfu}$Pus10 efficiently produced 1 mol of Ψ for each mol of tRNA in the reaction, suggesting complete modification of a single site in the tRNA. The maximum extent of modification observed with $_{pfu}$Cbf5 in this assay was ~0.6 mol Ψ per mol of substrate RNA (Figure A1A,C,E). This limitation could reflect the sensitivity of Cbf5 to alterations in the substrate RNA or the requirement of Cbf5 for accessory proteins (see below).

We determined the sites of tRNA modification for each enzyme by performing a combination of nearest neighbor analysis and mutational analysis. Wildtype *in vitro* transcribed $_{pfu}$tRNA^{Asp} was [^{32}P]-radiolabeled with one of the four nucleotides. The RNA was incubated with the protein under the optimal reaction conditions (see Figure A1). The modified tRNA was isolated and digested with either RNase T2 or nuclease P1 as indicated (Figure A2C and see Figure A2B). The resulting mononucleotides were analyzed by two-dimensional thin layer chromatography (2D-TLC) followed by autoradiography. For both $_{pfu}$Cbf5 and $_{pfu}$Pus10, Ψ is only detected when tRNA is either UTP-labeled and digested with nuclease P1, or CTP-labeled and digested with RNase T2 (Figure A2C, panels 1-4 and 9-12 and data not shown). These results indicate that

the modified uridine is 5'-adjacent to a cytidine. This limits the modified uridine to positions 22, 35, 39, 55 and/or 60 (numbered according to yeast $tRNA^{Phe}$ (Ofengand et al., 2001)). To test whether the modified uridine is formed at position 55, a U55C mutant substrate was constructed in which U55 was replaced with C55, which cannot be modified. For $_{pfu}$Cbf5 (Figure A2C, panels 5-6), this mutation completely abolished activity. For $_{pfu}$Pus10 (Figure A2C, panels 13-14), this mutation also dramatically reduced the amount of Ψ produced. The small amount of Ψ seen with $_{pfu}$Pus10 may be due to inefficient modification of other sites on the tRNA under the *in vitro* reaction conditions. These results clearly indicate that both $_{pfu}$Cbf5 and $_{pfu}$Pus10 can specifically modify U55 to Ψ55 *in vitro*.

Previously, we identified U55 modification activity in a crude *Pyrococcus furiosus* cell extract (Constantinesco et al., 1999). This extract was active on both a wildtype tRNA substrate and a truncated mini-substrate (miniS) that contains only the TΨ-arm and the acceptor stem (outlined in Figure A2A) (Constantinesco et al., 1999). To test the ability of the two candidate proteins to act on this substrate, [^{32}P]-UTP-radiolabeled transcripts of the miniS were incubated with the $_{pfu}$Cbf5 or $_{pfu}$Pus10 proteins, and analyzed as described above. As observed in Figure A2C panel 7, $_{pfu}$Cbf5 was unable to modify any uridine residue in the mini-substrate. In contrast, $_{pfu}$Pus10 was active on the mini-substrate (Figure A2C, panel 15).

The binding of $_{pfu}$Cbf5 to H/ACA guide RNAs (for rRNA modification) was previously found to depend on the presence of an intact ACA sequence near the 3' terminus of the guide RNA (box ACA) and mutations in this element prevented Cbf5 activity toward rRNA substrate (Baker et al., 2005; Charpentier et al., 2005). tRNAs

contain a CCA sequence at the 3' end. To determine whether the 3'CCA plays an important role in recognition of tRNA by $_{pfu}$Cbf5, we generated a tRNA that lacked the terminal CA (see dashed outline in Figure A2A). As shown in Figure A2C (panels 8 and 16), $_{pfu}$Cbf5 did not modify the tRNA -3'CA mutant. At the same time, the -3'CA mutation did not affect the ability of $_{pfu}$Pus10 to modify the substrate tRNA.

Ψ55 formation by $_{pfu}$Cbf5 in the absence and presence of accessory proteins

While the known tRNA pseudouridylations are introduced by dedicated, single-subunit enzymes, we know that in order to function with guide RNAs to modify rRNA in archaea and eukaryotes, Cbf5 requires accessory proteins: Gar1, L7Ae, and Nop10 (Baker et al., 2005; Charpentier et al., 2005). In order to determine whether the accessory proteins could also be important in tRNA U55 modification, pseudouridylation assays were performed in the absence and presence of these proteins.

As has been shown previously (Baker et al., 2005; Charpentier et al., 2005) and is demonstrated in Figure A3 (right panel, rRNA), in rRNA $_{pfu}$Cbf5 converts a radiolabeled uridine (targeted by a specific guide RNA) to Ψ only in the presence of the 3 accessory proteins (C+ lane).

Uniformly [^{32}P]-UTP-radiolabeled archaeal $_{pfu}$tRNA^{Asp} was also incubated with $_{pfu}$Cbf5 in the absence (C) and presence (C+) of stoichiometric amounts of the accessory proteins (Figure A3, left panel). After incubation, the tRNA was isolated and digested with RNase T1, and the TΨ-arm fragment was obtained by gel purification. The TΨ-arm fragment consists of nucleotides 54-65 and contains 3 uridines. The isolated fragment was digested with nuclease P1 and the mononucleotides were separated by single dimension TLC. As shown in Figure A3 (left panel), $_{pfu}$Cbf5

produced about 0.6 mol of Ψ per mol of tRNA in the absence of the accessory proteins, and about 0.9 mol of Ψ per mol of tRNA in the presence of the other proteins. The same extent of enhancement (~50% increase) was observed in two experiments.

In order to determine which of the 3 uridines within the TΨ arm of the tRNA was modified in this assay, a mutant $_{pfu}$tRNA^{Asp} in which U55 was replaced by C55 was tested (Figure A3, middle panel, tRNAU55C). No pseudouridylation was observed in the tRNA lacking U55, indicating that the pseudouridylation activity observed is directed specifically at uridine 55.

Our findings indicate that $_{pfu}$Cbf5 can modify tRNA U55, but not rRNA, in the absence of Gar1, Nop10 and L7Ae. However, the activity of $_{pfu}$Cbf5 toward tRNA U55 is enhanced in the presence of the accessory proteins, suggesting that a multi-protein Cbf5 complex may function in Ψ55 synthesis in *Pyrococcus furiosus*.

$_{pfu}$Pus10 can efficiently form Ψ55 in tRNA in Escherichia coli.

Our results indicate that both $_{pfu}$Cbf5 and $_{pfu}$Pus10 can catalyze tRNA Ψ55 formation *in vitro*. Unfortunately, because *Pyrococcus furiosus* is not currently amenable to genetic analysis, we are unable to unequivocally determine which protein is responsible for U55 modification in this organism. Therefore we tested for the ability of the enzymes to modify tRNA U55 in a heterologous system – an *Escherichia coli* strain harboring a null allele of the *truB* gene (Ofengand et al., 2001). It has been demonstrated that TruB is necessary (and sufficient) for the formation of Ψ55 *in vivo,* affording us the opportunity to test if expression of either of the two archaeal enzymes would complement U55 modification of *Escherichia coli* tRNAs (Ofengand et al., 2001). The *truB* strains were transformed with plasmids containing the *Pyrococcus furiosus*

pus10 or *cbf5* gene and the effect of the expression of each of these proteins on tRNA modification was assayed by primer extension after CMCT treatment (Ofengand et al., 2001). The cells were incubated for one hour at 50˚C (a compromise between the high temperature optimum for enzymatic activity (see Figure A1E-F) and optimal *Escherichia coli* growth temperature). tRNA was isolated from these cells and Ψ residues within $tRNA^{Cys}$ and $tRNA^{Phe}$ were identified using the CMC/RT assay system (Ofengand et al., 2001). Similar results were obtained with both tRNAs. As expected, Ψ55 was observed in the wildtype (wt) strain but not in the *truB* (KO) strain (Figure A4). *truB* cells transformed with a plasmid containing the *cbf5* gene showed no significant modification of U55 (data not shown), but because we were not able to detect expression of the Cbf5 protein in the *Escherichia coli*, no conclusion can be drawn from the lack of activity. Nearly wildtype levels of U55 modification were observed in the *truB* strains transformed with a plasmid containing the *pus10* gene (+Pus10). These results indicate that $_{pfu}$Pus10 can efficiently catalyze formation of tRNA Ψ55 in this heterologous system.

Discussion

Identification of archaeal enzymes that modify U55 in tRNA

Until the present work, no protein was attributed to the role of pseudouridylation of tRNA U55 in archaea. U55 modification activity was demonstrated in a cell-free extract of *Pyrococcus furiosus*, but no protein was implicated as the source of the activity (Constantinesco et al., 1999). We have identified two archaeal proteins capable of site-specific modification of U55: $_{pfu}$Cbf5, the only member of the TruB/Pus4/Cbf5 family (COG 0130) found in archaea, and $_{pfu}$Pus10 (previously PsuX, belonging to COG 1258), which was previously uncharacterized. The evidence reported here indicates

that both of these archaeal enzymes are capable of specifically performing this modification. It is possible that either or both of these proteins introduces the conserved Ψ55 in tRNA in *Pyrococcus furiosus*.

Pus10 (alias PsuX) is a previously uncharacterized enzyme. The data presented here clearly show that this enzyme can efficiently modify U55 both *in vitro* within a variety of substrates (wildtype, -3'CA mutant and miniS $tRNA^{Asp}$), and *in vivo* within *Escherichia coli* tRNAs. While Pus10 is not a member of the (TruB/Pus4/Cbf5) family of Ψ synthases that carry out tRNA U55 pseudouridylation in bacteria and eukaryotes, Pus10 orthologs are found in all sequenced archaeal genomes, consistent with the notion that this may be the conserved archaeal enzyme responsible for tRNA Ψ55 (Watanabe and Gray, 2000). Pus10 homologs are not present in bacteria (which employ TruB for Ψ55 synthesis), but do also appear to exist in certain higher eukaryotes (which employ Pus4 for Ψ55 synthesis) (Watanabe and Gray, 2000). Pus10 is renamed here as the 10^{th} Ψ synthase acting on tRNA (for review on Pus1-9, see (Ofengand et al., 2001)). The implication of Pus10 in tRNA Ψ55 formation (based on our findings) represents the first evidence for function of a protein outside the TruB/Pus4/Cbf5 family in introduction of this highly conserved modification.

The results of our studies indicate that $_{pfu}$Cbf5 also catalyzes specific tRNA Ψ55 formation *in vitro*. Cbf5 is the only enzyme in archaea that bears significant sequence similarity to the known Ψ55 synthases in bacteria and eukaryotes. Therefore, it is the most logical choice for the relevant enzyme. However, a role for Cbf5 in RNA-guided rRNA modification has been firmly established in archaea (Ofengand et al., 2001) as well as eukaryotes (Ofengand et al., 2001). Our finding that Cbf5 may also function in

tRNA U55 modification suggests an unexpected dual function for Cbf5 in archaea. There is some precedent for this concept in the dual specificity of Pus1p for tRNA and U2 snRNA in yeast (Ofengand et al., 2001).

Pseudouridylation of rRNA by Cbf5 depends upon the accessory proteins Gar1, Nop10 and L7Ae (Ofengand et al., 2001). In this work we found that the activity of Cbf5 toward tRNA is enhanced in the presence of the accessory proteins. Thus, in both cases the relevant enzyme may be a multi-protein complex.

The dual function of archaeal Cbf5 in tRNA and rRNA modification may account for structural similarities that have been noted between components of these pseudouridylation reactions from different systems. The X-ray crystal structures of bacterial TruB and archaeal Cbf5 are remarkably similar (Ofengand et al., 2001). Moreover, the structure of the regions of the substrate RNAs recognized and modified by TruB and Cbf5, the TΨ-arm of tRNA and guide RNA/rRNA duplex respectively, are strikingly comparable (Ofengand et al., 2001). Indeed this similarity was recently used to model the interaction between $_{pfu}$Cbf5 and the guide RNA/rRNA complex (Rashid et al., 2006). Our findings suggest that $_{pfu}$Cbf5 may also recognize a common feature of tRNAs and H/ACA guide RNAs outside the area of modification – the CCA and ACA at the 3' ends of tRNAs and H/ACA RNAs, respectively (Figure A3 and (Ofengand et al., 2001)). These conserved structural features between tRNA and rRNA modification systems may be a consequence of the dual function of Cbf5 in both processes in archaea. Alternatively, the observed activity of Cbf5 toward tRNA may reflect general conservation between Ψ synthases.

Ψ55 synthases in the three domains of life

Modification of U55 in tRNA to Ψ55 occurs in all three domains of life. In bacteria, TruB is responsible for this modification (Ofengand et al., 2001). In eukaryotes (with few exceptions including *Drosophila melanogaster* and *C. elegans*) the tRNA Ψ55 synthase is Pus4 (Ofengand et al., 2001). The results presented here implicate Cbf5 or Pus10 (or both) as the enzyme responsible for this modification in archaea. Cbf5 also functions in H/ACA RNA-guided modification of rRNA in archaea (Ofengand et al., 2001). rRNA modifications are also introduced by Cbf5 (by an RNA-guided mechanism) in eukaryotes (Ofengand et al., 2001). (Pseudouridylation of rRNA is carried out by a set of site-specific enzymes in bacteria (Ofengand et al., 2001)). Pus10 is more distantly related to these other Ψ synthases.

The overall domain organization of Pus10 differs significantly from the other Ψ55 synthases (Figure A5). Sequence analysis of Pus10 reveals the presence of several distinct domains (Ofengand et al., 2001). CX_2C Zn binding motifs are found in the N-terminal region of Pus10 proteins (Ofengand et al., 2001). A similar motif is present in Pus1, a distinct tRNA Ψ synthase in *S. cerevisiae*, which contains one atom of Zn essential for the native conformation and tRNA recognition (Ofengand et al., 2001). A THUMP domain (present in some THioUridine synthases, Methylases and Pseudouridine synthases) is located adjacent to the CX_2C Zn binding motifs of Pus10 (Ofengand et al., 2001). This characteristic motif is proposed to be involved in RNA binding (Ofengand et al., 2001). A recent study indicates that the THUMP motif alone cannot bind RNA, but facilitates the interaction of the catalytic domain of the archaeal Trm-G10 methyltransferase enzyme with tRNA substrates (Ofengand et al., 2001). The

Pus10-related proteins of *Drosophila* and *Caenorhabditis* lack this THUMP domain (Ofengand et al., 2001). The catalytic domain of Pus10 appears to be found in the C-terminal region of the protein (Ofengand et al., 2001).

TruB and Cbf5 (but not Pus4 of *S. cerevisiae*) contain a characteristic PUA domain (so called because it is present in PseudoUridine synthases and Archaeosine transglycosylase) within the C-termini of the proteins (Ofengand et al., 2001) (Figure A5). In archaeal Cbf5, the PUA domain plays a crucial role in binding H/ACA guide RNAs (Ofengand et al., 2001). This RNA binding depends on the presence of an ACA sequence at the 3' end of the guide RNA (Ofengand et al., 2001). We have found that archaeal Cbf5 also requires an intact 3' terminal CCA sequence for activity toward tRNA (Figure A2).

Evolutionary Considerations

The observations made in this study may provide some insight on the evolutionary origin of the present-day RNA Ψ synthases, especially the tRNA Ψ55 synthases. The finding that in archaea Cbf5 may function in tRNA U55 pseudouridylation as well as RNA-guided pseudouridylation of rRNA supports the idea that Cbf5 is a direct descendant of a primordial TruB/Pus4-like tRNA Ψ synthase. This primordial Ψ synthase may have had the ability to also act on rRNA (and some dependence on guide RNAs and accessory proteins) in a common ancestor of eukaryotes and archaea. After the divergence of eukaryotes and archaea, a gene duplication in eukaryotes may have allowed separation of the functions in Cbf5 and Pus4. An analogous situation exists in the cases of the RNA-guided machinery

catalyzing the formation of 2'*O*-methylribose in both rRNA and tRNA (Ofengand et al., 2001).

Archaeal Pus10, as a member of a completely different family of Ψ synthases appears to be an independent invention of a novel type of tRNA Ψ synthase. We suggest that these tRNA Ψ synthases could have resulted from convergent evolution.

As we have indicated, our results suggest the possibility that Cbf5 and Pus10 provide redundant function in formation of the highly conserved tRNA U55 modification in archaea. Therefore it is also possible that the Pus10-related proteins present in higher eukaryotes also function redundantly in tRNA Ψ55 synthesis (in this case with Pus4).

Acknowledgments

We thank S. Auxilien (CNRS, Gif-sur-Yvette, France) for the kind gift of the pML156 plasmid and P. M. Wikström (University of Umeå, Sweden) for providing the strain GOB113. This work was supported by a research grant from the Fonds pour la Recherche Fondamentale Collective (FRFC) to LD, a research grant from the CNRS (GEOMEX program) to HG, and NIH grant RO1 GM54682 to MT and RT.

References

Baker, D.L., Youssef, O.A., Chastkofsky, M.I., Dy, D.A., Terns, R.M., and Terns, M.P. (2005). RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. Genes Dev *19*, 1238-1248.

Becker, H.F., Motorin, Y., Planta, R.J., and Grosjean, H. (1997a). The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of

psi55 in both mitochondrial and cytoplasmic tRNAs. Nucleic Acids Res *25*, 4493-4499.

Becker, H.F., Motorin, Y., Sissler, M., Florentz, C., and Grosjean, H. (1997b). Major identity determinants for enzymatic formation of ribothymidine and pseudouridine in the T psi-loop of yeast tRNAs. J Mol Biol *274*, 505-518.

Bylund, G.O., Wipemo, L.C., Lundberg, L.A., and Wikstrom, P.M. (1998). RimM and RbfA are essential for efficient processing of 16S rRNA in Escherichia coli. J Bacteriol *180*, 73-82.

Charpentier, B., Muller, S., and Branlant, C. (2005). Reconstitution of archaeal H/ACA small ribonucleoprotein complexes active in pseudouridylation. Nucleic Acids Res *33*, 3133-3144.

Constantinesco, F., Motorin, Y., and Grosjean, H. (1999). Transfer RNA modification enzymes from Pyrococcus furiosus: detection of the enzymatic activities in vitro. Nucleic Acids Res *27*, 1308-1315.

Ferre-D'Amare, A.R. (2003). RNA-modifying enzymes. Curr Opin Struct Biol *13*, 49-55.

Grosjean, H., Keith, G., and Droogmans, L. (2004). Detection and quantification of modified nucleotides in RNA using thin-layer chromatography. Methods Mol Biol *265*, 357-391.

Kaya, Y., and Ofengand, J. (2003). A novel unanticipated type of pseudouridine synthase with homologs in bacteria, archaea, and eukarya. Rna *9*, 711-721.

Koonin, E.V. (1996). Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res *24*, 2411-2415.

Massenet, S., Mougin, A., and Branlant, C. (1998). Posttranscriptional modifications in the U small nuclear RNAs. In Modification and Editing of RNA, G. H, and B. R, eds. (Washington, DC, American Society of Microbiology).

McCloskey, J.A., and Rozenski, J. (2005). The Small Subunit rRNA Modification Database. Nucleic Acids Res *33*, D135-138.

Nurse, K., Wrzesinski, J., Bakin, A., Lane, B.G., and Ofengand, J. (1995). Purification, cloning, and properties of the tRNA psi 55 synthase from Escherichia coli. Rna *1*, 102-112.

Ofengand, J., Del Campo, M., and Kaya, Y. (2001). Mapping pseudouridines in RNA molecules. Methods *25*, 365-373.

Rashid, R., Liang, B., Baker, D.L., Youssef, O.A., He, Y., Phipps, K., Terns, R.M., Terns, M.P., and Li, H. (2006). Crystal structure of a Cbf5-Nop10-Gar1 complex and implications in RNA-guided pseudouridylation and dyskeratosis congenita. Mol Cell *21*, 249-260.

Reyes, V.M., and Abelson, J. (1987). A synthetic substrate for tRNA splicing. Anal Biochem *166*, 90-106.

Roovers, M., Wouters, J., Bujnicki, J.M., Tricot, C., Stalon, V., Grosjean, H., and Droogmans, L. (2004). A primordial RNA modification enzyme: the case of tRNA (m1A) methyltransferase. Nucleic Acids Res *32*, 465-476.

Rozenski, J., Crain, P.F., and McCloskey, J.A. (1999). The RNA Modification Database: 1999 update. Nucleic Acids Res *27*, 196-197.

Sprinzl, M., and Vassilenko, K.S. (2005). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res *33*, D139-140.

Watanabe, Y., and Gray, M.W. (2000). Evolutionary appearance of genes encoding proteins associated with box H/ACA snoRNAs: cbf5p in Euglena gracilis, an early diverging eukaryote, and candidate Gar1p and Nop10p homologs in archaebacteria. Nucleic Acids Res *28*, 2342-2352.

Figure A1: Ψ **formation by $_{pfu}$Cbf5 and $_{pfu}$Pus10 in $_{pfu}$tRNA^{Asp}.** (A, B) Increasing molar amounts of $_{pfu}$Cbf5 and $_{pfu}$Pus10 were incubated at 70°C for 60 min. (C, D) 0.12 nmol (5 μg) of $_{pfu}$Cbf5 and 0.10 nmol (5 μg) of $_{pfu}$Pus10 were incubated with T7-transcribed $_{pfu}$tRNA^{Asp} for increasing time intervals at 70°C. (E, F) 0.12 nmol (5 μg) of $_{pfu}$Cbf5 and 0.10 nmol (5 μg) of $_{pfu}$Pus10 were incubated at increasing temperatures for 60 minutes. After incubation, the tRNA was recovered, hydrolyzed by nuclease P1 and the resulting nucleotides were separated by 2D-TLC chromatography (see Materials and Methods). Mol of Ψ produced per mol of tRNA was measured after counting the radioactivity in the spots corresponding to UMP and ΨMP on the chromatogram and taking into account the nucleotide composition of the tRNA substrate. Precision is estimated to be about 10%. The results shown are representative of two similar experiments.

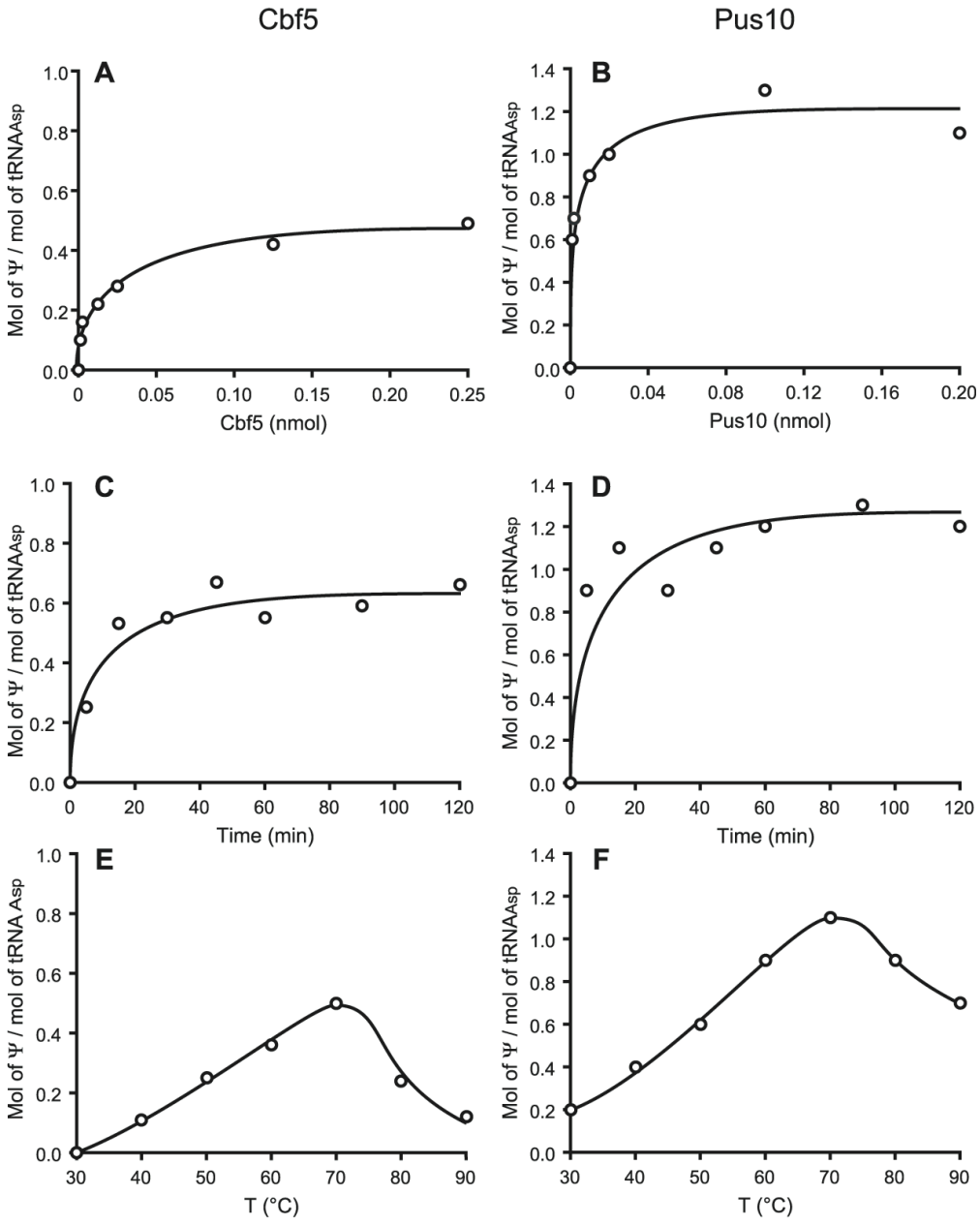

Cbf5
Pus10
A
B
C
D
E
F
Mol of Ψ / mol of tRNAAsp
Cbf5 (nmol)
Pus10 (nmol)
Time (min)
T (°C)

Figure A2: Analysis of $_{pfu}$Cbf5 and $_{pfu}$Pus10 modification of various tRNA substrates. (A) Cloverleaf structure of $tRNA^{Asp}$ of *Pyrococcus furiosus* used as the wildtype tRNA substrate in this work. The universal numbering system for nucleotides in tRNA corresponds to that of (Ofengand et al., 2001). U55 is indicated. C75 and A76 (missing in –3'CA substrates) are indicated by a dashed box. The portion of the tRNA sequence included in the miniS substrate is indicated by the plain box. (B) The expected patterns of nuclease P1 and RNase T2 cleavage in the U55 region of a [^{32}P] UTP- and [^{32}P] CTP-labeled tRNA, respectively. Nuclease P1 generates 5'phosphate-nucleosides while RNase T2 generates 3'phosphate-nucleosides. (C) 2D-TLC analysis of $_{pfu}$Cbf5 and $_{pfu}$Pus10 modification of various tRNA substrates. WT indicates wildtype $_{pfu}tRNA^{Asp}$ (panels 1-4, 9-12); U55C indicates U55 replacement mutant (panels 5,6,13,14); miniS indicates mini-substrate (panels 7,15) and –3'CA indicates 3' terminal CA deletion (panels 8,16). See text for more details. UTP/CTP/ATP and GTP refer to the [^{32}P]-labeled nucleotide incorporated at transcription. Incubation was for 1 hour at 70°C in the presence of 0.12 nmol (5 μg) of $_{pfu}$Cbf5 (panels 1-8) and 0.01 nmol (0.5 μg) of $_{pfu}$Pus10 (panels 9-16). After incubation, the RNA was digested by nuclease P1 or RNase T2 (as indicated in each panel) and the resulting nucleotides were analyzed by 2D-TLC on cellulose plates and autoradiography. Circles in dotted lines show the migration of the canonical nucleotides used as UV markers.

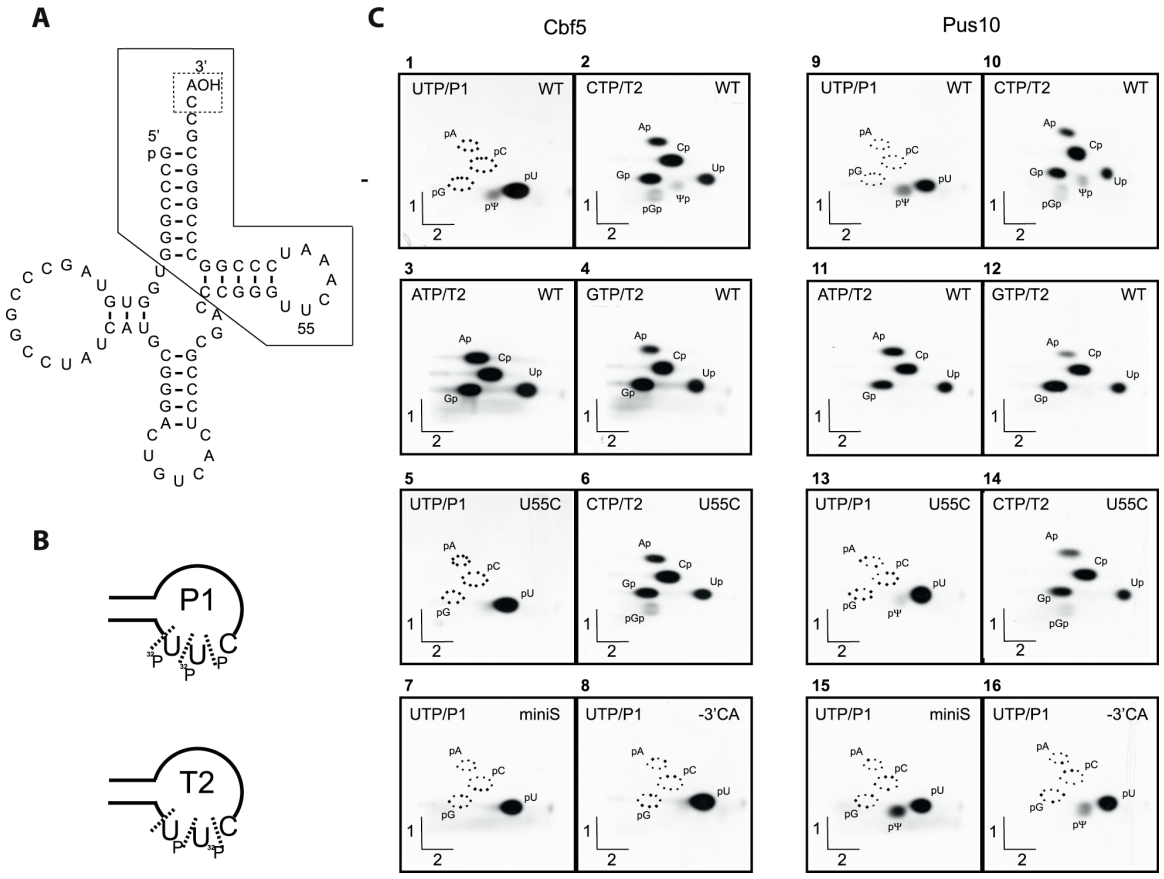

A
3'
AOH
5'
55
B
P1
T2
C
Cbf5
Pus10
1 UTP/P1 WT
2 CTP/T2 WT
9 UTP/P1 WT
10 CTP/T2 WT
3 ATP/T2 WT
4 GTP/T2 WT
11 ATP/T2 WT
12 GTP/T2 WT
5 UTP/P1 U55C
6 CTP/T2 U55C
13 UTP/P1 U55C
14 CTP/T2 U55C
7 UTP/P1 miniS
8 UTP/P1 -3'CA
15 UTP/P1 miniS
16 UTP/P1 -3'CA

Figure A3: $_{pfu}$Cbf5 modification of tRNA in the absence and presence of accessory proteins. [^{32}P] UTP-labeled wildtype (wt) or U55 mutant (U55C) tRNA (left and middle panels) was incubated with no protein (-), $_{pfu}$Cbf5 alone (C), or $_{pfu}$Cbf5 plus *Pyrococcus furiosus* accessory proteins Gar1, Nop10, and L7Ae (C+). After incubation, the tRNA was digested with RNase T1 and the fragment corresponding to the TΨ-loop (nts 54-65) was excised and purified from a 20% denaturing gel. This fragment was digested with nuclease P1, and Ψ and U were separated and analyzed by TLC and phosphoimaging. In the right panel, site-specifically radiolabeled target rRNA was incubated with the same combinations of proteins in the presence of Pf9 guide rRNA. The resulting RNA was digested with nuclease P1 and analyzed as for tRNA. The number of moles of Y incorporated per mole of RNA substrate (taking into account the number of uridines in the region analyzed) is indicated below each lane.

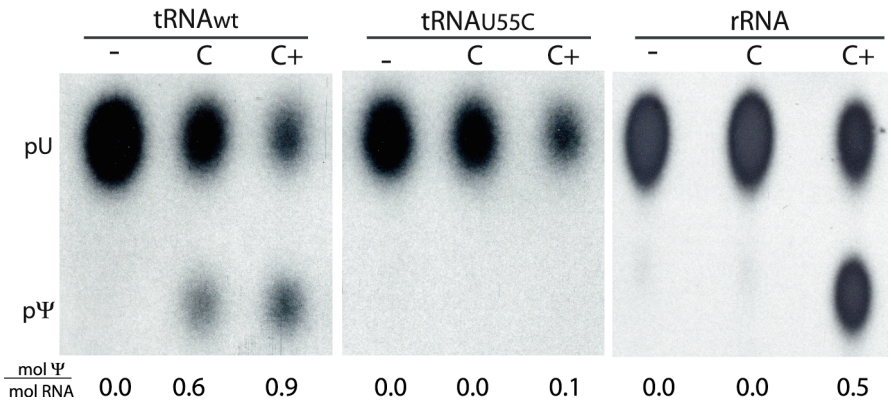

tRNAwt
tRNAU55C
rRNA
-
C
C+
-
C
C+
-
C
C+
pU
pΨ
mol Ψ / mol RNA
0.0
0.6
0.9
0.0
0.0
0.1
0.0
0.0
0.5

Figure A4: Detection of Ψ in *Escherichia coli* tRNA^{Cys} and tRNA^{Phe}. (A) CMC/RT analysis of tRNA^{Cys} of *Escherichia coli.* tRNA was isolated from wildtype (wt), *truB* (KO), or *truB* transformed with a plasmid containing the $_{pfu}$*pus10* gene (+Pus10) *Escherichia coli* and CMC modified. Primer extension was performed as described in Materials and Methods with a primer complementary to $_{eco}$tRNA^{Cys}, nucleotides 61-76. The arrow indicates a strong stop at Ψ55. (B) CMC/RT analysis of $_{eco}$tRNA^{Phe} as described above using a primer complementary to nucleotides 61-76 of $_{eco}$tRNA^{Phe}.

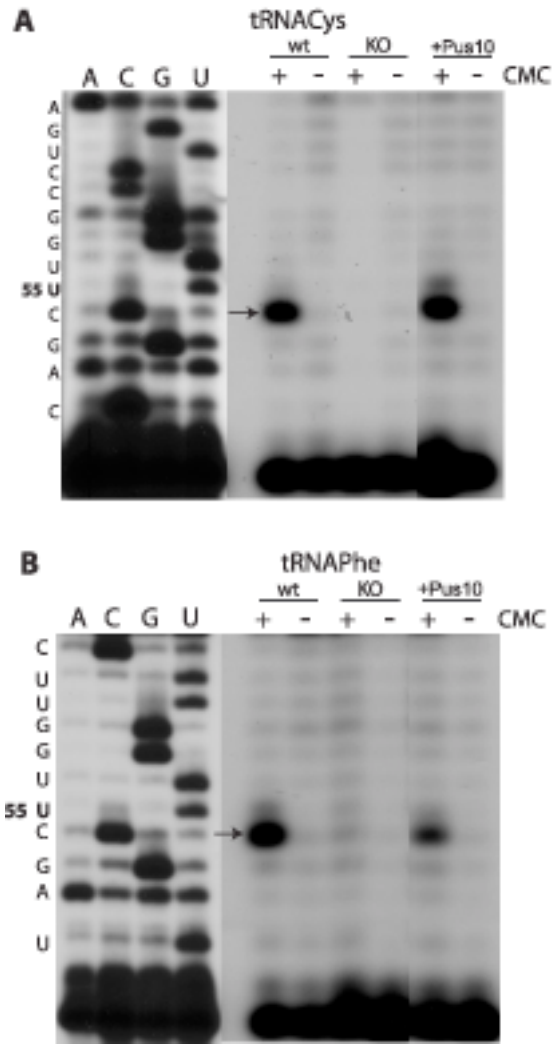
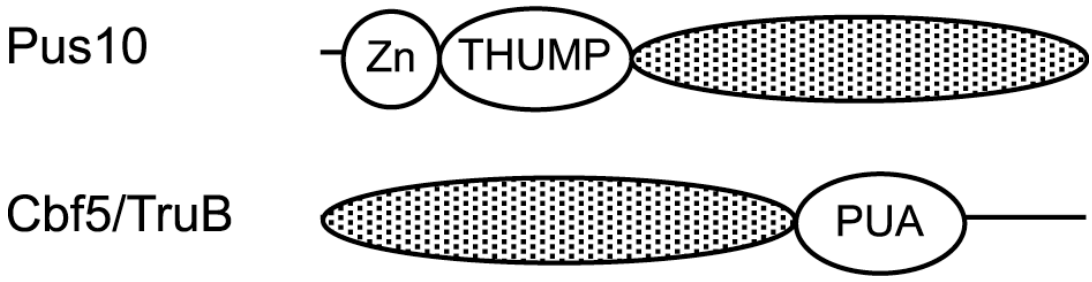

A
tRNACys
wt
KO
+Pus10
A C G U
+ - + - + - CMC
55
B
tRNAPhe
wt
KO
+Pus10
A C G U
+ - + - + - CMC
55

Figure A5: Domain organization of Cbf5 and Pus10. Schematic representation of predicted domains of Cbf5, TruB, Pus4 and Pus10 proteins. Conserved catalytic, PUA, Zn and THUMP domains are indicated. See text for discussion.


Pus10
Zn
THUMP
Cbf5/TruB
PUA