Small RNA-guided processes in the hyperthermophilic methanogen *Methanopyrus kandleri*

Philipps Universität Marburg

Dissertation

Zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem Fachbereich Biologie

der Philipps-Universität Marburg

vorgelegt von

Andreas A. H. Su

aus Konstanz

Marburg/Lahn

November 2014

Small RNA-guided processes in the hyperthermophilic methanogen *Methanopyrus kandleri*

Philipps Universität Marburg

Dissertation

Zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem Fachbereich Biologie

der Philipps-Universität Marburg

vorgelegt von

Andreas A. H. Su

aus Konstanz

Marburg/Lahn

November 2014

Die Untersuchungen zur vorliegenden Arbeit wurden von September 2011 bis September 2014 am Max-Planck-Institut für Terrestrische Mikrobiologie unter der Leitung von Herrn Dr. Lennart Randau durchgeführt.

Vom Fachbereich
der Philipps-Universität Marburg als Dissertation
angenommen am: 04.12.2014

Erstgutachter: Dr. Lennart Randau
Zweitgutachter: Prof. Dr. Rudolf K. Thauer
Drittgutachter: Prof. Dr. Renate Renkawitz-Pohl
Viertgutachter: Prof. Dr. Torsten Waldminghaus

Tag der mündlichen Prüfung: 12.12.2014

Teile dieser Arbeit sind in folgenden Artikeln veröffentlicht:

Hrle, A*., **Su, A. A. H***., Ebert, J., Benda, C., Randau, L*. and Conti, E*. (* joint first/corresponding authorship) *Structure and RNA-binding properties of the Type III-A CRISPR-associated protein Csm3.* RNA Biology, 2013, 10, 1670-1678.

Su, A. A. H., Tripp, V. and Randau, L. *RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri* Nucleic Acids Res, 2013, 41, 6250-6258.

Weitere Veröffentlichungen:

Su, A. A. H. and Randau, L. *A-to-I and C-to-U editing within transfer RNAs.* Biochemistry (Moscow), 2011, 76, 932-937.

“Imagine walking out in the countryside and not being able to tell a snake from a cow from a mouse from a blade of grass. That’s been the level of our ignorance.”

— *Carl R. Woese*

Table of Contents

Chapter V

Chapter VI

Appendix

Chapter 0

Summary

In this thesis, a combination of RNAseq, computational and biochemical methods was applied to analyze processes that use small RNAs (sRNAs) as guide molecules at extreme temperatures. Here, the hyperthermophilic archaeon *Methanopyrus kandleri*, which grows at temperatures of up to 110°C, was used as a model organism.

The genome of *M. kandleri* harbors two CRISPR-Cas systems that use CRISPR RNA (crRNA) as guide molecules to target foreign nucleic acids. RNAseq analysis revealed a high abundance and processing of crRNAs in *M. kandleri* that indicated that CRISPR-Cas systems are highly active at extreme temperatures. Furthermore, the crystal structure of the CRISPR-associated protein Csm3 was solved in collaboration with Prof. Dr. Elena Conti (MPI Martinsried). Csm3 was found to bind crRNAs and was shown to function as the crRNA-binding backbone protein in type III-A CRISPR-Cas interference complexes. A recently discovered nucleic acid-guided mechanism uses prokaryotic Argonaute (pAgo) proteins. In *M. kandleri*, a pAgo protein was found to be encoded within a potential operon of CRISPR-associated genes and the analysis of recombinant pAgo protein production revealed a high toxicity in *Escherichia coli* that might correlate with its potential defense function against plasmid DNA.

Methylation of rRNA is regulated by a different sRNA-guided mechanism that utilizes C/D box sRNAs to target a ribonucleoprotein complex to the rRNA methylation site. In *M. kandleri*, a record number of 126 C/D box sRNAs were detected by RNAseq analysis and indicate an increased potential for rRNA methylation reactions. Furthermore, most of the C/D box sRNAs were detected as circular molecules. Taken together, the circularization of C/D box sRNAs and the high requirement for rRNA methylation are suggested to be adaptations to the hyperthermophilic lifestyle of *M. kandleri*.

Finally, RNAseq analyses were used to identify tRNA precursors in *M. kandleri* that feature a unique C-to-U editing reaction of base 8. The occurrence of this editing event was used to deduce the order of tRNA processing steps in a non-compartmentalized cell, indicating that termini truncation precedes intron removal and editing.

Zusammenfassung

In der vorliegenden Arbeit wurden RNAseq, sowie bioinformatische und biochemische Methoden angewandt um zelluläre Prozesse zu analysieren die von kleinen RNAs (sRNA) gesteuert werden. Um zu untersuchen, wie diese Prozesse bei extremen Temperaturen funktionieren, wurde das hyperthermophile Archaeon *Methanopyrus kandleri*, welches bei Temperaturen von bis zu 110°C lebt, als Modellorganismus verwendet.

Das Genom von *M. kandleri* enthält zwei CRISPR-Cas Systeme, welche crRNAs (CRISPR RNAs) als Zielerkennungsmoleküle verwenden, um sich gegen Viren zur Wehr zu setzen. Durch RNAseq Analysen wurden große Mengen an prozessierten crRNAs detektiert. Dies deutet darauf hin, dass CRISPR-Cas Systeme bei extremer Temperatur sehr aktiv sind. Zudem wurde die Kristallstruktur des CRISPR-assoziierten Proteins Csm3 in Zusammenarbeit mit Prof. Dr. Elena Conti (MPI Martinsried) aufgeklärt. Anhand von RNA Bindestudien wurde Csm3 als crRNA-bindendes Protein identifiziert, welches als Gerüstprotein des Typ III-A CRISPR-Cas Interferenzkomplexes fungiert. Erst kürzlich konnte gezeigt werden, dass pAgo (prokaryotic Argonaute) Proteine kurze Nukleinsäuresequenzen verwenden, um Zielmoleküle zu erkennen. In *M. kandleri*, ist ein solches pAgo Protein in einem potentiellen Operon zusammen mit Proteinen des CRISPR-Cas Systems kodiert. Versuche pAgo rekombinant herzustellen wiesen auf eine hohe Toxizität des Proteins in *Escherichia coli* hin, was mit einer möglichen Abwehrfunktion gegen Plasmid DNA in Zusammenhang stehen könnte.

Die Methylierung von RNA Molekülen wird durch einen anderen Prozess gesteuert, welcher C/D Box sRNAs als Zielerkennungssequenzen verwendet, um Ribonukleoproteinkomplexe an Methylierungsstellen ribosomaler RNAs zu führen. In *M. kandleri* wurde eine Rekordzahl von 126 C/D Box sRNAs durch RNAseq Analysen detektiert, welche auf eine erhöhte Methylierung von rRNAs hinweist. Zudem wurde ein Großteil der C/D box sRNAs als zirkuläre Moleküle detektiert. Zusammenfassend können die Zirkularisierung von C/D Box sRNAs und der erhöhte Bedarf an RNA Methylierungen als Adaptionen an die hyperthermophile Lebensweise von *M. kandleri* angesehen werden.

Abschließend wurden durch RNAseq Analysen tRNA-Vorläufer Moleküle identifiziert, welche die nur in *M. kandleri* vorkommende C-zu-U Modifikation der Base 8 aufwiesen. Das Auftreten dieser Modifikation wurde genutzt, um daraus Rückschlüsse auf die Reihenfolge von tRNA Prozessierungsschritten in nicht kompartimentierten Zellen zu schließen. Unsere Analysen ergaben, dass das Beschneiden der Enden dem Entfernen von Introns, sowie der Editierungsreaktion vorrausgeht.

Chapter I

Introduction

Methanopyrus kandleri, a hyperthermophilic archaeon

Methanopyrus kandleri strain AV19 is a methanogenic archaeon living at the upper temperature limit of life (Fig. I 1). It was isolated from the base of a hydrothermal vent in the Gulf of California at 2000 m depth where it lives as an obligate chemolithoautotroph using H_2 and CO_2 as energy and carbon sources (Huber *et al*., 1989). Remarkably, *M. kandleri* needs extreme temperatures between 84°C and 110°C to proliferate and is, together with *Pyrolobus fumarii* living at 113°C (Blöchl *et al*., 1997) and strain 121 growing at 121°C (Kashefi *et al*., 2003), considered to be one of the most hyperthermophilic organisms known so far. This was confirmed more recently, when the newly isolated *M. kandleri* strain 116 was grown at 122°C under elevated hydrostatic pressure which extended the maximum temperature limit for life to 122°C (Takai *et al*., 2008). A search for features that allow life at these extreme temperatures revealed high intracellular concentrations of cyclic 2,3-diphosphoglycerate molecules that are essential for thermostabilization of proteins in *M. kandleri* (Shima *et al*., 1998). Moreover, the genome of this organism exhibits an unusually

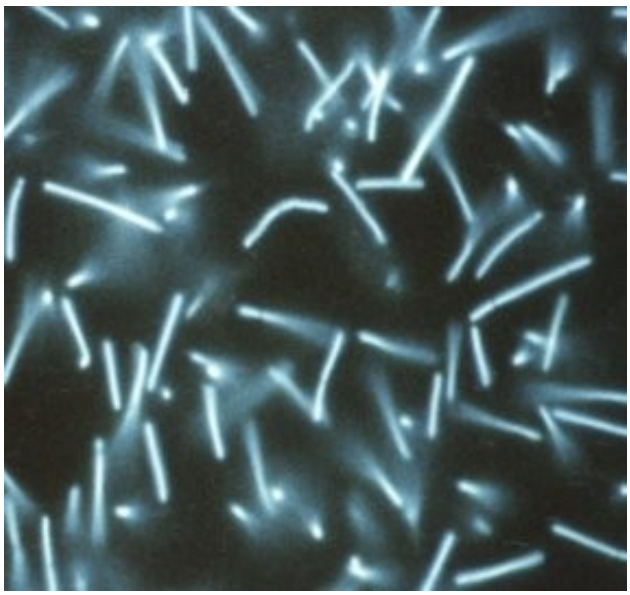

Figure I 1. Methanopyrus kandleri cells visualized by fluorescence microscopy
(K. O. Stetter & R. Rachel, University of Regensburg)

high G/C content of 61.2 % that might contribute to the stabilization of DNA. However, other features facilitating *M. kandleri* to withstand extreme temperatures, for example mechanisms for RNA stabilization, still remain elusive. In this thesis, RNAseq methodology was applied in *M. kandleri* to globally analyze the synthesis, maturation and modifications of RNA molecules at extreme temperatures.

The phylogenetic positioning of M. kandleri

The phylogenetic position of *M. kandleri* amongst other archaea is debated. Early analyses of 16S rRNA sequences placed *M. kandleri* near the root of the archaeal phylogenetical tree at a position distinct from other known methanogens (Burggraf *et al*., 1991). In contrast, alignments of ribosomal proteins and trees based on gene content grouped *M. kandleri* together with other archaeal methanogens (Slesarev *et al*., 2002). Although the latter theory is supported by the observation of peculiar changes of transcriptional proteins as a result of an accelerated evolution (Brochier *et al*., 2005), *M. kandleri* is still considered by some researchers to be an organism that is phylogenetically close to the last universal common ancestor (LUCA) (Mat *et al*., 2008).

C-to-U editing of tRNAs in M. kandleri

M. kandleri contains the largest fraction of orphan genes found in any prokaryotic genome so far (Jensen *et al*., 2003). One example for this is the cytidine deaminase acting on tRNA base 8 (CDAT8) that catalyzes a unique C-to-U editing reaction of tRNAs at position 8 (Randau *et al*., 2008). The U at position 8 is highly conserved in all three domains of life and is required for maintaining the L-shaped tertiary structure of tRNAs by interacting with the A at position 14 (Westhof *et al*., 1985). *M. kandleri* is the only known organism that encodes a C at position 8 of most tRNA genes and post-transcriptionally modifies the C to a U by a CDAT8 enzyme (Fig. I 2). Possible benefits of this unusual mechanism might be the evasion of viruses attacking tRNA genes or the proper folding of tRNA precursors at elevated temperatures (Randau *et al*., 2008).

In the present study, the unique C-to-U editing mechanism in *M. kandleri* was analyzed using the RNAseq technology. The analysis of sequencing reads with and without a U at position 8 enabled us use the C-to-U editing event as a marker to deduce the order of tRNA processing events in a non-compartimentalized prokaryotic cell.

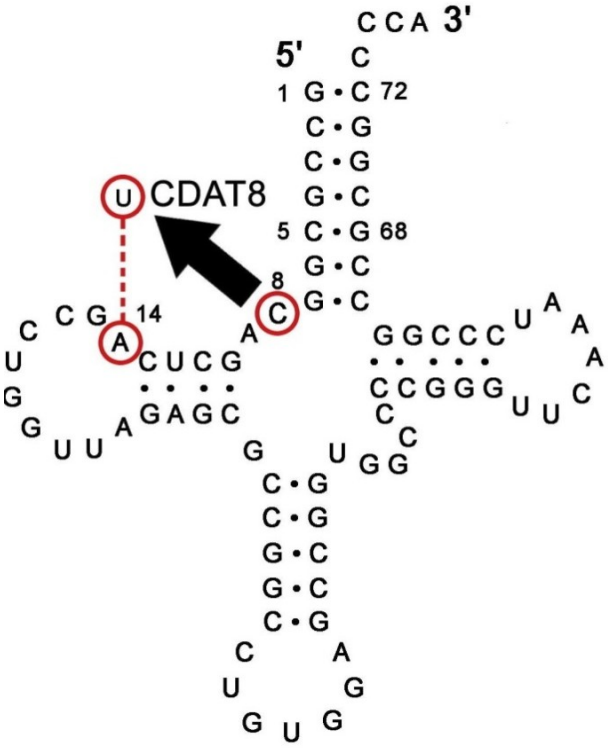

Figure I 2. C-to-U editing in *M. kandleri*
30 out of 34 tRNA genes in *M. kandleri* contain a C at position 8 as editing substrate for CDAT8. This editing event ensures the formation of the important tertiary base interaction between U8 and A14 as indicated in the secondary structure of *M. kandleri* tRNA-His (Randau et al., 2009; Heinemann et al., 2010).

M. kandleri harbors a CRISPR-Cas system with a prokaryotic Argonaute homologue

The genome of *M. kandleri* harbors five clustered regularly interspaced short palindromic repeat (CRISPR) clusters which are essential components of a prokaryotic adaptive immune system. A CRISPR is defined as an array of identical repetitive DNA sequences, interspaced by similarly sized unique spacer sequences that are derived from viruses or plasmids (Jansen *et al*., 2002, Moijca *et al*., 2005; Pourcel *et al*., 2005). The CRISPR array is transcribed and processed into small CRISPR RNAs (crRNAs), that together with a set of CRSIPR associated (Cas) genes confer immunity to mobile genetic elements (Barrangou *et al*., 2007). Bioinformatical analyses showed a huge diversity of *cas* genes in different CRISPR-Cas containing organisms. This led to a classification of the CRISPR-Cas system into the three major types I, II and III, depending on the presence of several marker genes. Furthermore, the three major types are divided into different subtypes (Makarova *et al*., 2011a). The only genes that are conserved in all three major types are *cas1* and *cas2* (Haft *et al*., 2005). These two genes are required for the acquisition of new spacers (Yosef *et al*., 2012; Nunez *et al*., 2014). In most CRISPR-Cas systems, the protein Cas6 is involved in the processing of long precursor crRNA (pre-crRNA) molecules into smaller, mature crRNAs. The remaining Cas proteins are proposed to form a multi protein interference complex that utilizes the processed crRNA as a guide to target and degrade the foreign nucleic acid (Van der Oost *et al*.,

2014). The best studied example of an interference complex is the CRISPR associated complex for antiviral defense (CASCADE) of *Escherichia coli*. This complex exhibits a seahorse like structure composed of five different proteins. The most abundant one is Cas7 that builds the backbone of the structure composed of six Cas7 copies. The protein Cas3 is located at the base of this structure and is responsible for the cleavage of the targeted DNA (Jore *et al*., 2011). In contrast to type I CRISPR-Cas systems, the type III CRISPR-Cas systems have been studied to a lesser extent. They can be divided into the two subtypes, III-A and III-B. The type III-A system specific proteins are Csm proteins (Csm1 - Csm6) whereas type III-B specific proteins are Cmr proteins (Cmr1 - Cmr6) (Van der Oost *et al*., 2014). For type III-A systems it has been shown that they target foreign DNA (Marraffini and Sontheimer, 2008). In contrast to that, type III-B systems can target RNA molecules (Hale *et al*., 2009). Electron microscopy studies of type III interference complexes revealed striking similarities with the structure of the well studied type I CASCADE. Type III interference complexes are also built around a crRNA binding, multiprotein backbone that consist for type III-A of Csm3 and for type III-B of Cmr4 proteins, similar to Cas7 in type I systems (Spilman *et al*., 2013; Staals *et al*., 2013; Rouillon *et al*., 2013; Van der Oost *et al*., 2014). Furthermore, bioinformatical analysis of Cas proteins revealed that Csm3 and Cmr4 share a number of conserved sequence motifs with Cas7, including a ferredoxin-like RNA binding fold (also referred to as RAMP-domain) (Makarova *et al*., 2011b). In this thesis, the functions and structural properties of the proteins Csm3 and Cmr4 have been analyzed in the hyperthermophilic methanogen *M. kandleri* to obtain insights into the CRISPR-Cas system of this organism.

The CRISPR-Cas system of *M. kandleri* exhibits two *cas* gene sets of the CRISPR-Cas subtypes III-A and III-B (Fig. I 3). Additionally single copies of the genes *cas1* and *cas2* are available that could be shared by both systems for integration of new spacer sequences. An interesting and unique feature of the CRISPR-Cas system in *M. kandleri* is the presence of a gene, encoding a homologue of the Argonaute (Ago) protein, which is located between the genes *cas1* and *cas2* in a predicted operon. The Ago protein is the key effector in the eukaryotic RNA interference (RNAi) mechanism and is required for small RNA-guided mRNA cleavage (Joshua-Tor and Hannon, 2011). Ago proteins can also be found in some bacteria and archaea and are referred to as prokaryotic Argonaute (pAgo) proteins. So far, 80 different pAgo proteins have been identified but their functions in prokaryotes are still not fully understood. The analysis of the genomic context of *pAgo* genes revealed that the proteins might be involved in a novel defense mechanism against mobile genetic elements (Makarova *et al*., 2009). Recently, pAgo associated nucleic acids were analyzed in *Rhodobacter sphaeroides* using an DNA/RNA deep sequencing approach, suggesting an RNA dependent DNA cleavage that prefers plasmid DNA (Olovnikov *et al*., 2013). In contrast, studies in *Thermus thermophilus*, *Methanocaldococcus janaschii* and *Thermus thermophilus*

indicate a DNA-dependent DNA cleavage function of pAgo (Sheng *et al*., 2013; Zander *et al*., 2014; Swarts *et al*., 2014). The DNA cleavage of *T. thermophilus* pAgo has been shown to follow an RNase H-type mechanism involving a pair of catalytic Mg^{2+} ions that are coordinated by a tetrad of Asp-Glu-Asp-Asp (Sheng *et al*., 2013).

One of the goals of this thesis was to analyze the function of the pAgo in *M. kandleri*. In this organism pAgo is encoded in a CRISPR-Cas-gene cluster and it is tempting to speculate that this might represent a rare case where a pAgo interacts with proteins of the CRISPR-Cas system that could potentially provide it with DNA guide molecules (Makarova *et al*., 2006).

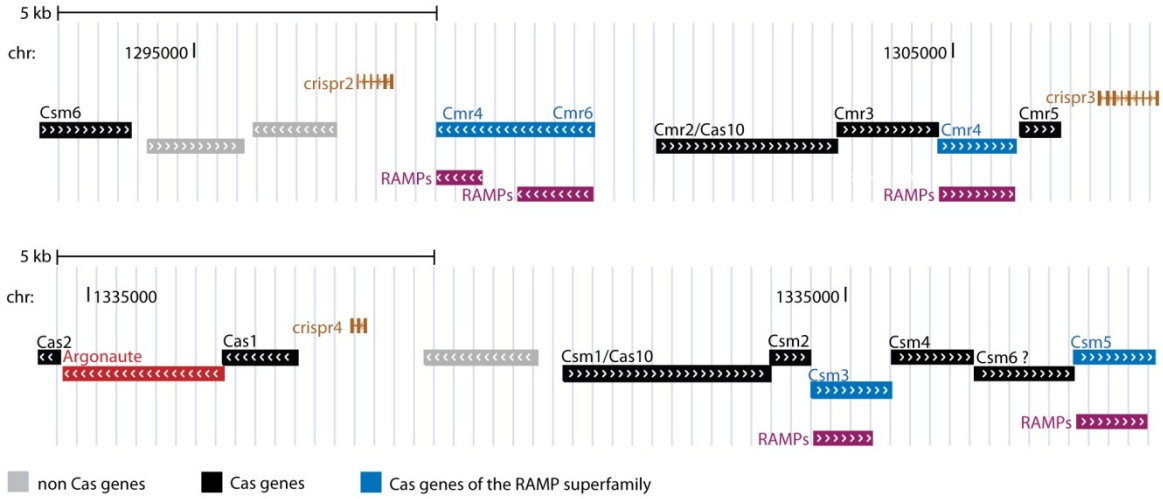

Figure I 3. Arrangement of *cas* genes and CRISPR loci in *M. kandleri*
The CRISPR loci 2, 3 and 4 as well as the *cas* genes and their corresponding positions on the genome are shown. The *cas* genes are marked in black and pAgo which is in a predicted operon with *cas1* and *cas2* is marked in red. Genes encoding potential RNA binding proteins with a predicted ferredoxin-like fold (RAMP domain) are depicted in blue (Figure modified from http://archaea.ucsc.edu).

Chapter II

RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile *Methanopyrus kandleri*

Andreas A. H. Su[1], Vanessa Tripp[1,2] and Lennart Randau[1,2,*]

[1]Max-Planck-Institute for Terrestrial Microbiology, Max Planck Research Group: Prokaryotic Small RNA Biology, Karl-von-Frisch Strasse 10, 35037 Marburg, Germany; [2]LOEWE Center for Synthetic Microbiology (Synmikro), 35037 Marburg, Germany.

Nucleic Acids Research, 2013, Vol. 41, No. 12.

Abstract

The methanogenic archaeon *Methanopyrus kandleri* grows near the upper temperature limit for life. Genome analyses revealed strategies to adapt to these harsh conditions and elucidated a unique transfer RNA (tRNA) C-to-U editing mechanism at base 8 for 30 different tRNA species. Here, RNA-Seq deep sequencing methodology was combined with computational analyses to characterize the small RNome of this hyperthermophilic organism and to obtain insights into the RNA metabolism at extreme temperatures. A large number of 132 small RNAs were identified that guide RNA modifications, which are expected to stabilize structured RNA molecules. The C/D box guide RNAs were shown to exist as circular RNA molecules. In addition, clustered regularly interspaced short palindromic repeats RNA processing and potential regulatory RNAs were identified. Finally, the identification of tRNA precursors before and after the unique C8-to-U8 editing activity enabled the determination of the order of tRNA processing events with termini truncation preceding intron removal. This order of tRNA maturation follows the compartmentalized tRNA processing order found in Eukaryotes and suggests its conservation during evolution.

Introduction

Organisms that belong to the domain Archaea are often adjusted to the harshest environments present on earth. *Methanopyrus kandleri* is a methanogenic archaeon that can survive extreme heat and pressure conditions. One of the earliest archaeal genome sequences was obtained for *M. kandleri* AV19 (Slesarev *et al*., 2002). This organism was isolated near hydrothermal vents from the Gulf of California at a depth of 2000 m below sea level and at temperatures of 84-110°C (Huber *et al*., 1989). More recently, *M. kandleri* strain 116 was isolated and shown to proliferate at 20 MPa pressure and 122°C (Takai *et al*., 2008), which extended the known upper temperature limit for life. The search for features that allow growth at these temperatures revealed that *M. kandleri* sustains high intracellular concentrations of the trivalent anion cyclic 2,3-diphosphoglycerate, and enzymes isolated from this organism can require over 1 molar salt concentrations to ensure their stability and activity (Shima *et al*., 1998; Breitung *et al*., 1992). A large number of orphan genes were identified in the *M. kandleri* genome, which revealed several unusual enzymes including the topoisomerase V and a two-subunit reverse gyrase (Slesarev *et al*., 1993; Krah *et al*., 1996). One orphan gene encodes a unique cytidine deaminase that acts on transfer RNA (tRNA) base 8 (CDAT8) (Randau *et al*., 2009). All organisms from all three domains of life contain tRNA genes with a conserved T residue at position 8 and the folding of tRNA molecules involves tertiary interactions between U8 and the equally conserved base A14. The only known exception to this rule is the minimal set of 34 tRNA genes that is present in the *M. kandleri* genome and that contains 30 genes that have an unusual C base at position 8. The tRNA precursors that contain a C8 base are then edited by CDAT8 deamination to ensure that mature tRNA molecules contain the universal U8 required for proper folding. It has been suggested that the acquisition of the orphan genes was mediated by viruses (Forterre *et al*., 2006) and that these rare genes evolved in the viral community. The identification of five antiviral clustered regularly interspaced short palindromic repeats (CRISPR) systems in the genome of *M. kandleri* underlines the importance of virus-host interactions in extreme environments.

The phylogenetic positioning of *M. kandleri* among other Archaea is debated. *M. kandleri* is the only known member of the class Methanopyrales belonging to the phylum Euryarchaeota. The 16S ribosomal RNA (rRNA) gene-based phylogenetic studies placed the *M. kandleri* branch deep at the root of the archaeal tree (Burggraf *et al*., 1991), whereas whole genome trees include *M. kandleri* with other methanogens in a monophyletic group (Brochier *et al*., 2004). Although the latter scenario is supported by the observation of accelerated evolution of transcriptional proteins, *M. kandleri* is still considered to represent an organism that is phylogenetically close to the Last Universal Common Ancestor at the root of the tree of life (Mat *et al*., 2008).

The extreme growth conditions of *M. kandleri*, the divergent phylogenetic positioning and the unique tRNA editing events raise question about the RNA properties of this organism. In the present study, we applied RNA-Seq deep sequencing methodology to isolated enriched small RNA samples from *M. kandleri* to obtain an insight into the small RNA metabolism at extreme temperatures and to follow tRNA C-to-U editing activity. A small RNome (sRNome) set was characterized that highlights the importance of the modification of structured RNAs and the defence against viruses. Sequencing of tRNAs with and without C8 editing revealed tRNA processing intermediates and highlighted that termini truncation precedes intron removal.

Material and Methods

Cell cultivation and RNA isolation

Methanopyrus kandleri cells were a kind gift of D. Söll. The organism was grown in the Archaeenzentrum Regensburg (H. Huber, M. Thomm, K. Stetter) in a 300 l fermenter as described (Slesarev *et al.*, 2002). Total RNA was isolated by SDS-lysis of the cell pellet and phenol/chloroform extraction, and small RNAs were purified from total RNA using the MirVana RNA extraction kit (Ambion).

RNA-sequencing

Six different *M. kandleri* small RNA libraries were prepared for sequencing. The following RNA pre-treatment protocols were applied before adapter ligation: Sample 1: 10 mg of *M. kandleri* sRNA was incubated in 1x T4 polynucleotidekinase (T4 PNK) buffer (NEB) for 6 h at 36°C in a total volume of 50 ml with no additional treatment. Sample 2: 10mg of *M. kandleri* sRNA was incubated in a buffer containing 50 mM sodium phosphate, 1 mM EDTA and 0.36% H_2O_2 at 20°C for 3 h to facilitate dethiolation of RNA. Sample 3: 10 mg of *M. kandleri* sRNA was incubated in a buffer containing 170 mM Tris-HCl (pH 8.8) at 37°C for 3 h to facilitate deacetylation of tRNA molecules. Samples 1-3 resulted in highly similar RNA-Seq sequencing output. Samples 4-6 were treated with T4 PNK to ensure proper termini for adapter ligation to RNA molecules that contain either 5'-OH termini or 3'-phosphate termini. Sample 4: 10 mg of small RNA was incubated at 37°C for 6 h with 20 U T4 PNK and in 1x T4 PNK buffer in a total volume of 50ml. Subsequently, 2 mM adenosine triphosphate (ATP) and 10 U T4 PNK were added, and the reaction mixture was incubated for 1 h at 37°C to generate monophosphorylated 5'-termini. Sample 5: The protocol for Sample 4 was followed but the final addition of 2 mM ATP was omitted. Sample 6: The protocol for Sample 4 was followed, but 2 mM ATP was added from the beginning. Samples 4 to 6 resulted in highly similar RNA-Seq sequencing output. RNA libraries were prepared with an Illumina TruSeq RNA Sample Prep Kit (Ambion), and sequencing on an Illumina HiSeq2000 sequencer was performed at the Max-Planck Genomecentre, Cologne (Max Planck Institute for Plant Breeding Research, Köln, Germany).

Identification of small RNA species

Sequencing reads were trimmed by (i) removal of Illumina TruSeq linkers and poly-A tails and (ii) removal of sequences using a quality score limit of 0.05. A total of 83 338 855 reads with an average length of 61 nt were mapped to the *M. kandleri* reference genome (GenBank: NC_003551) with CLC Genomics Workbench 5.5 (CLCBio, Aarhus, Denmark). The following mapping parameters were used: mismatch cost, 2; insertion cost, 3; deletion cost, 3; length

fraction, 0.5; similarity, 0.8. This program was also used to determine the coverage of individual RNA molecules. All predicted RNA molecules and their termini were manually verified, and all intergenic regions were checked for the presence of RNA molecules with coverage of >500 reads. Target prediction of identified C/D box sRNAs was performed with the PLEXY tool using the default parameters (Kehr *et al*., 2011). Provided possible target RNA sequences were as follows: (i) all rRNAs and tRNAs sequences and (ii) all identified sRNAs (Supplementary Tab. II S1). The following algorithms were used for further computational analysis of the data: RNA folding [Mfold (Zuker, 2003)], tRNA gene prediction [tRNAScan-SE (Lowe and Eddy, 1997), genomic tRNA database (Chan and Lowe, 2009)], small nucleolar RNA (snoRNA) gene prediction [snoscan (Lowe and Eddy, 1999)], crRNA identification [crisprdb (Grissa *et al*., 2007)], RNA alignments [ClustalW2 (Larkin *et al*., 2007)] and RNA visualization [VARNA (Darty *et al*., 2009)]. Gene annotations were obtained from GenBank.

Inverse Reverse transcriptase-polymerase chain reaction

In all, 10 ng of an *M. kandleri* sRNA preparation were treated with Superscript III reverse transcriptase (Invitrogen) and primers against C/D box sRNAs 5, 11, 16, 17, 27 to generate complementary DNA (cDNA) for the detection of circular sRNA molecules. The RNA was denatured at 100°C for 5 min and cooled on ice for 5 min to facilitate reverse transcription at 55°C for 30 min. Inactivation of the reverse transcriptase was carried out at 70°C for 15 min. Subsequently, the cDNA products were polymerase chain reaction (PCR) amplified with Platinum Taq DNA polymerase (Invitrogen) using forward and reverse primers. The following primers were used:

C/D_5_For:

5'-GATCCCATCCTCATCCCAC-3',

C/D_5_Rev:

5'-GATCTGGGAGGCCGTTAC-3',

C/D_11_For:

5'-GCGTGGGGTAGCATCGTC-3',

C/D_11_Rev:

5'-ACCCCGATGAGGAGGAAC-3',

C/D_16_For:

5'-GTTCGTCGGCCTACCTCG-3',

C/D_16_Rev:

5'-GGGATGACGACCCCTGG-3',

C/D_17_For:

5'-CGATCCTGCGACCACTCC-3',

C/D_17_Rev:

5'-GCGGTTGTTCGCTTCTTCATC-3',

C/D_27_For:

5'-GCTCAATCTTCATCCACAGGATC-3',

C/D_27_Rev:

5'-AGCCCGGCACTGACTCG-3'.

The PCR amplificates were cloned into a pCR2.1 TOPO vector (Invitrogen) and sequenced (Eurofins MWG Operon).

Data availability

The RNA-Seq data are available at NCBI's Gene Expression Omnibus website as series GSE44979.

Results

The small RNA profile of Methanopyrus kandleri

RNA-Seq methodology was used to facilitate the genome-wide analysis of small RNA production at single nucleotide resolution in *M. kandleri*. RNA molecules were isolated from *M. kandleri* cells, and small RNA molecules (<200 nt) were selectively enriched. The RNA preparation was split into six fractions and subjected to different RNA modification procedures (detailed in the 'Materials and Methods' section) before adapter ligation. Three RNA samples were treated with T4 PNK to enable sequencing of RNAs with 5'-OH termini. Six independent library preparations were used for HiSeq2000 RNA-Seq sequencing. The obtained reads were mapped to the 1.69 bp long *M. kandleri* genome. In total, these mappings contain 83 338 855 reads with an average length of 61 nt. Clear differences for the obtained sRNome coverage were observed for the three RNA samples that were treated with T4 PNK in comparison with the three samples without this treatment. This allowed us to focus our small RNA analysis on two of these conditions and provided us with four further mappings that we used for the assessment of the reproducibility of our observations. The overview of the genome-wide RNA profile of *M. kandleri* reveals that most

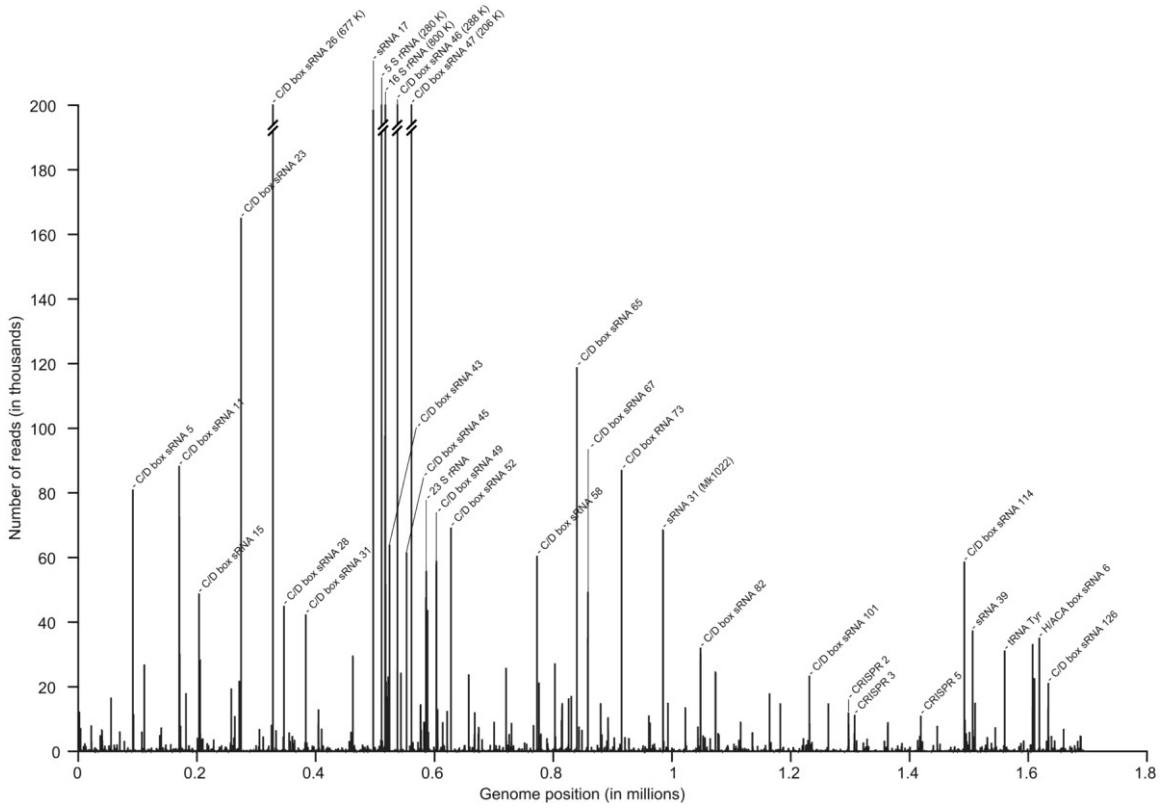

Figure II 1. The sRNome of *M. kandleri*
The overview graph illustrates the genome-wide coverage of Illumina HiSeq2000 reads mapped to the *M. kandler*i AV19 genome (Genbank: NC_003551, 1694969 bp). Prominent peaks were analyzed at single-nucleotide resolution, and abundant sRNAs (e.g. C/D box sRNAs) were identified. For peaks that cover over 200 000 reads, the number of reads is given in brackets.

sequence reads were obtained for fragments of the 5S and 16S rRNAs as well as for small C/D box sRNAs (Figure II 1). Members of this class of RNA modifying guide RNAs are abundant in *M. kandleri* and will be discussed in detail later in the text. In addition, several potential regulatory sRNAs were identified. Finally, mature crRNA were selectively enriched in the RNA samples treated with T4 PNK.

Abundance of tRNA molecules

Sequence reads that represent tRNAs are highly underrepresented in RNA-Seq studies. In the cell, tRNAs are usually found to represent the most abundant class of RNAs, which correlates with their high demand during protein biosynthesis. However, tRNA genes are often covered only by a few hundred reads in our RNA-Seq mappings. In addition, these sequence reads often span only a fraction of the tRNA genes. It has been noted before that the high level of tRNA modification and the stable secondary and tertiary structure of these molecules poses difficulties for the reverse transcriptase during the cDNA preparation step for RNA-Seq libraries. This is especially true for the tRNAs of *M. kandleri* that have to function at growth temperatures ~100°C, which is ensured by tRNA stems that nearly exclusively are formed by G-C base pairing and by an exceptionally diverse population of modified nucleosides (Sauerwald *et al.*, 2005). On the other hand, the obtained fragmented tRNA sequence reads offer insights into the maturation and processing stages as, for example, the insertion of bulky modifications can be pinpointed by reverse transcriptase stalling (Findeiss *et al.*, 2011). The tRNA genes with a significant number of mapped reads are tRNA Tyr (31353 reads) and tRNA Asp (23871 reads). Interestingly, in both cases, the majority of the reads represent tRNA precursors that still contain a 5' leader sequence (5'-GAGGGGTGCGGGA-3') for tRNA Tyr and 5'-GGAGGGATGAGA-3' for tRNA Asp). These precursors most likely do not contain modifications that could block reverse transcription. The tRNA Tyr species is the only tRNA that is required to start with a C as the first base to ensure proper aminoacylation by the archaeal tyrosyl-tRNA synthetase (Fechter *et al.*, 2000). As transcription initiation prefers purine residues, RNase P activity is required to mature the 50 end of functional tRNA Tyr. Most *M. kandleri* tRNAs contain short 5'-terminal leader sequences that contain multiple G residues commonly found at archaeal transcription initiation sites (Supplementary Tab. II S1). Three tRNA genes displayed problematic annotations. A tRNA Leu isoacceptor lacked three 5'-terminal residues, and the two tRNA Glu isoacceptors display an unusual second intron in the tRNAs' D-loop, which was previously predicted to be recognized by the splicing endonuclease (Marck and Grosjean, 2003). No sequence reads for tRNA His species were observed that contain a G-1 residue. This hallmark of tRNA His species is posttranscriptionally added by a special enzyme, the tRNA His guanylyltransferase, which is also present in *M. kandleri*. It is possible that this reaction happens late during tRNA His maturation after the incorporation of modifications

that abolishes reverse transcription efficiency or that the G-1 is further modified in a fashion that prevents 5'-terminal adapter ligation.

Maturation of tRNA molecules

Reads that mapped to the regions of tRNA genes were used to distinguish between different stages of tRNA processing. The following different tRNA maturation intermediates were identified: (i) tRNAs with 5'-terminal leader sequences, (ii) tRNA with intronic sequences, (iii) tRNAs with C8 to U8 editing by CDAT8 and (iv) tRNA fragments (possibly caused by bulky modifications). The first three intermediates can be inferred from sequencing reads at single nucleotide resolution that were mapped to the tRNA genes in the reference genome or tRNA genes in a modified reference genome that contained the C8-T8 exchange. The different precursors were quantified (Supplementary Tab. II S1). The occurrence of different tRNA maturation intermediates that did or did not contain features of other intermediates allowed for the generation of an order of tRNA processing events. Here, the C8-U8 editing serves as an ideal marker for one clearly distinguishable modification event. Ninety-nine percent of all tRNAs that contained the edited mature U8 base were identified by reads that started with the +1 base (Supplementary Tab. II S1 and Fig. II 2). Only tRNA precursors that still contain C8 or the four tRNAs that naturally have the U8 base are commonly found with 5'-terminal leader sequences. Thus, it can be deduced that 5' tRNA processing by RNase P precedes tRNA editing. In contrast, 89% of sequence reads for tRNAs with introns still contain this intron and also exhibit the C8-U8

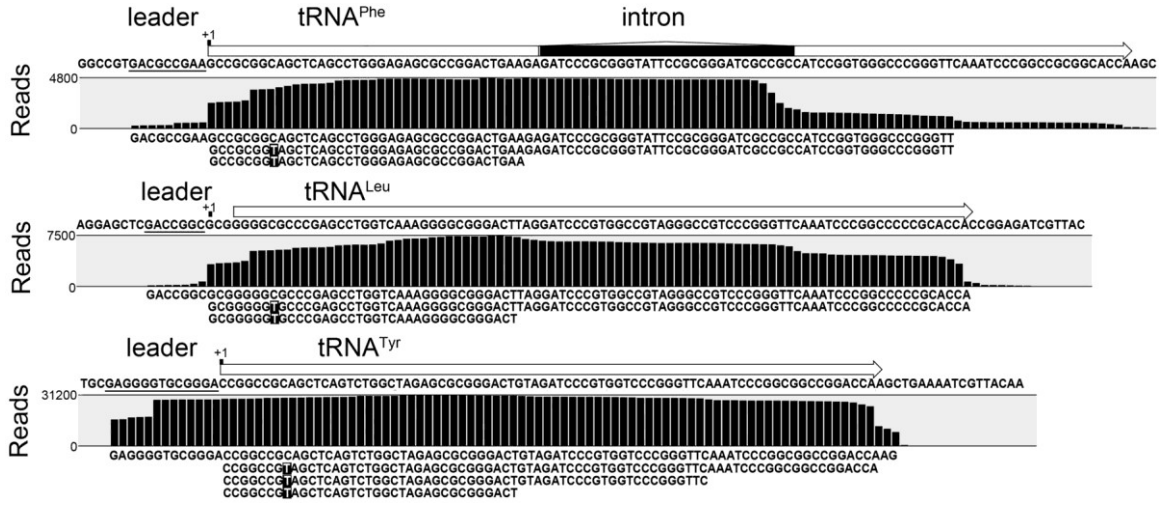

Figure II 2. Detection of tRNA processing events
Indicated is the coverage of Illumina Hiseq2000 reads at single nucleotide resolution for three exemplary tRNA genes. The Genbank annotation of the three tRNA genes and an intronic sequence is depicted above the coverage graphs and the identified mature 5' termini are indicated by '+1'. The 5' end of tRNA Leu was misannotated. Representative reads that allow the determination of the tRNA precursor processing state are given below the coverage graphs: (i) tRNAs with 5'-terminal leader sequences before RNase P processing, (ii) tRNA precursors with intronic sequences, (iii) tRNAs with C8-to-U8 editing (highlighted in black) and (iv) shortened tRNA reads that suggest the presence of tRNA modifications.

exchange. In addition, these tRNAs often terminate prematurely at positions with presumed modifications (Supplementary Table II S1 and Figure II 2). This indicates that certain modifications are inserted into the tRNA molecule before CDAT8 activity, and that the removal of tRNA introns occurs at the later stages of tRNA maturation. In addition, the two tRNA Glu isoacceptors are represented by reads with C8 to U8 editing (114 and 294 reads, respectively) that still contain the unusual intron in the D-loop of the tRNA. This indicates that, although CDAT8 requires a matured acceptor stem, it tolerates disruption and insertion of structured sequences in the D-loop and the anticodon-loop. This observation is in agreement with a model of CDAT8 interaction with the T-stem/loop and acceptor-stem portion of 30 different tRNA substrates (Randau *et al*., 2009). Removal of the D-loop intron appears to trail the removal of the canonical intron. In conclusion, the order of observed tRNA processing events is as follows: (i) 3' processing, (ii) 5' processing, (iii) modifications, C-to-U editing, (iv) intron removal (Fig. II 3).

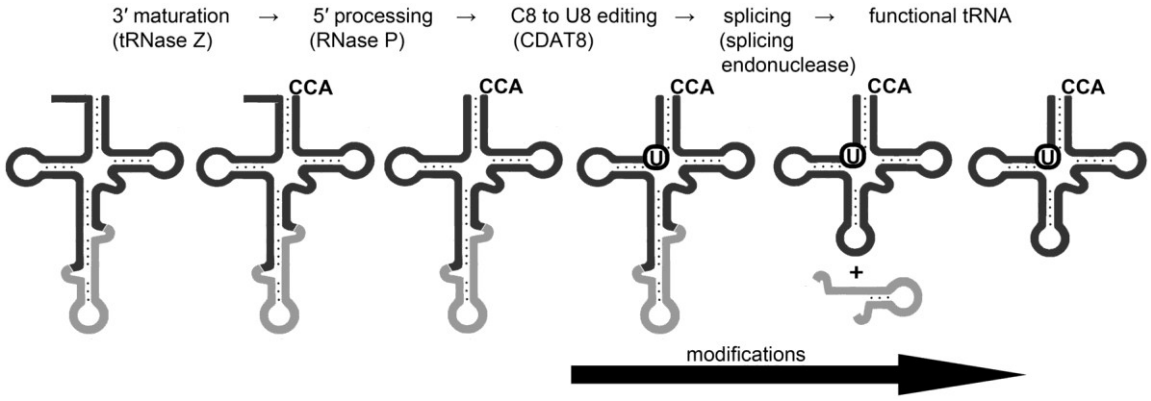

Figure II 3. Deduced order of tRNA processing events
Sequencing reads that represent different tRNA processing states were quantified (Supplementary Table II S1) and indicate the presented order of tRNA processing events.

Identification of C/D box and H/ACA box sRNAs

Most sequencing reads were obtained for C/D box sRNAs. These 55-97 nt small RNA molecules use a guide sequence located between the conserved C and D box RNA motifs to target RNA (most commonly rRNAs) for 2'-O-methylation (Omer *et al*., 2000). A different and less abundant class of guide RNAs, the H/ACA box sRNAs direct RNA pseudouridylation. A total of 126 C/D box sRNAs and 6 H/ACA box sRNAs were found in *M. kandleri* (Supplementary Tab. II S1). The abundance of individual C/D box sRNAs was found to be highly variable in the cell, despite conserved length, structure and functions. Some C/D box sRNAs were covered by few hundred reads, whereas others were among the most abundant sRNA molecules in the cell covered by hundred thousands of sequencing reads (Fig. II 1 and Supplementary Tab. II S1). We identified sequence coverage and manually annotated termini and potential C and D boxes for all detected

C/D box sRNAs (Supplementary Tab. II S1). Computational analyses using the algorithm 'snoscan' (Lowe and Eddy, 1999) identified 97 potential C/D box sRNAs of which 79 were verified experimentally and 17 potential C/D box sRNAs that were classified as 'questionable' of which only two were verified. Forty-five C/D box sRNAs were not predicted computationally. All predictions of C/D box sRNAs within protein coding genes were found to be erroneous. Potential targets for the guide sequence of these small RNAs were computationally predicted (Supplementary Tab. II S1). These targets include not only rRNAs but also tRNAs and other non-coding RNAs. Interestingly, the target with the overall best score is the 5' leader region of a newly identified H/ACA box sRNA (Supplementary Tab. II S1 and Supplementary Fig. II S1), which suggests potential cooperative activity of guide RNAs required for different RNA modifications. We observed the highest number of sequence reads that map to a protein coding gene for the gene MK0859 coding for GAR1, a member of the H/ACA small ribonucleoprotein complex. Analysis of the GAR1 mRNA sequence revealed a highly stable 3'-terminal hairpin structure with 10 consecutive G-C base pairs within the open reading frame that might influence mRNA stability (Supplementary Fig. II S2).

Permuted sequencing reads indicated that most C/D box sRNA exist as circular RNA molecules in the cell. In some cases, the circularized form was found to be more common than the linear RNA molecule (Supplementary Tab. II S1). Circular RNA molecules require internal cleavage to facilitate adapter ligation to the permuted RNA, which suggests that such circular molecules are likely underrepresented in RNA-Seq studies. To verify that circularization is not an artifact occurring during RNA library preparation, we performed inverse reverse transcriptase-polymerase chain reaction (RT-PCR) amplification with selected C/D box sRNA candidates. The amplification products were sequenced and confirmed the presence of the circularization sites (Fig. II 4).

Figure II 4. Inverse RT-PCR verifies circular C/D box sRNA formation
One hundred and twenty-six C/D box sRNAs were identified and permuted sequencing reads suggested circular RNA molecules (Supplementary Table II S1). Selected circular C/D box sRNA molecules were amplified by inverse RT-PCR with outward facing primers (arrows) and amplificates were sequenced. Sequencing information between primer sequences is underlined and circularization junctions are marked.

crRNAs

The genome of *M. kandleri* harbors five CRISPR clusters. The characteristic CRISPR repeat sequences are interspaced by sequences that can be derived from viral DNA (so-called spacers) and are relics of viral infections. Our RNA-Seq analyses revealed that all five CRISPR clusters are active, i.e. they are transcribed and processed into small crRNAs that have the potential to mediate immunity against viruses that are recognized via base complementarity between viral DNA and crRNA spacer sequence. Based on the presence of different sets of CRISPR-associated (Cas) genes and different repeat sequences and structures, CRISPR/Cas subtypes have been classified (Makarova *et al*., 2011). *M. kandleri* contains two sets of Cas proteins that belong to subtype III-A and III-B, respectively, and an Argonaute gene located between genes encoding Cas1 and Cas2. Mature crRNAs of other CRISPR subtypes were shown to harbor 5'-hydroxyl- and potential 2',3'- cyclic phosphate ends (Jore *et al*., 2011). These termini are incompatible with the adaptor ligation needed for RNA sequencing. To produce suitable ends for RNA sequencing of mature crRNAs and to identify whether these termini are present in *M. kandleri* crRNAs, a set of three RNA preparations was treated with T4 PNK. This enzyme adds 5' phosphates and removes potential 3' phosphate groups from mature crRNAs before library preparation and RNA-Seq. It is

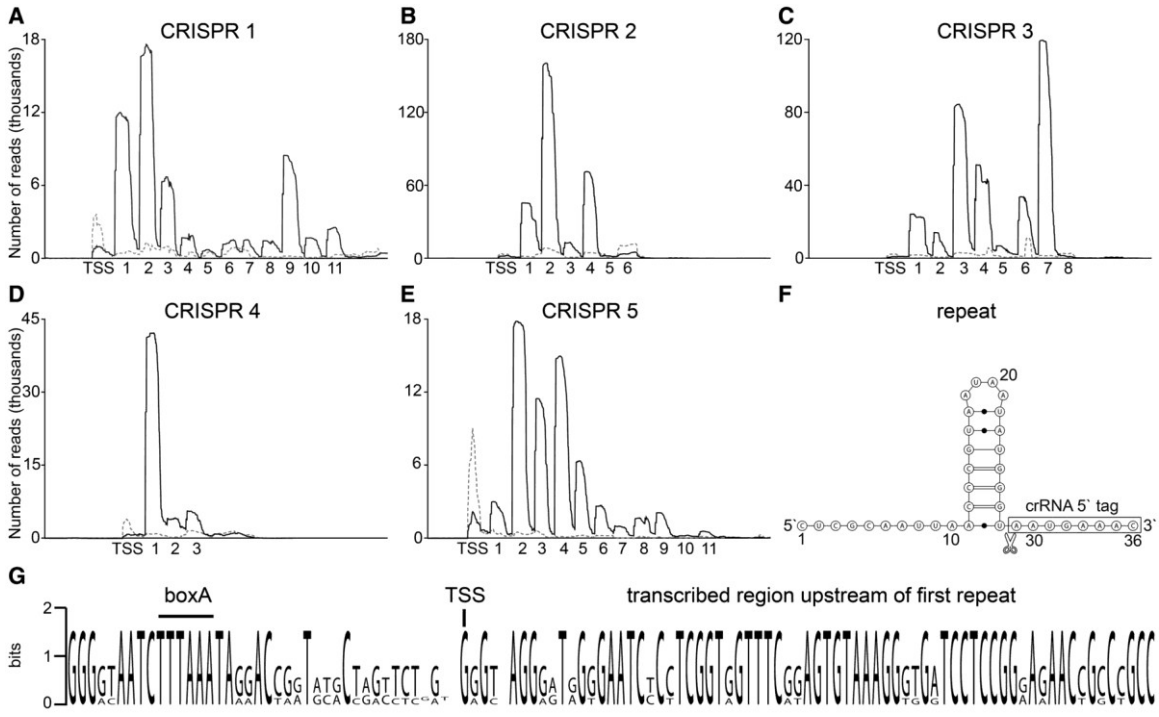

Figure II 5. Detection of crRNA abundance and processing
(A-E) Illumina Hiseq2000 reads of T4 PNK-treated (black line) and untreated (gray dashed line) RNA samples were mapped to CRISPR loci 1-5 in the *M. kandleri* genome. The number of identified processed crRNAs and the detected transcription start site (TSS) are indicated. (F) Shown is the consensus repeat sequence and proposed structure. The detected cleavage site is indicated, and mature crRNAs contain an 8 nt 5' tag. (G) A sequence logo of the aligned regions upstream of the first repeat in CRISPR loci 1-5 highlights a conserved transcription start site (TSS) and a boxA promoter sequence.

evident that T4 PNK treatment is required for the detection of all crRNAs of the five CRISPR clusters in *M. kandleri* (Fig. II 5A-E and Supplementary Tab. II S1). Most of the sequenced crRNAs harbor an 8 nt 5' tag derived from cleavage within the repeat sequence, whereas the 3' ends are gradually degraded and did not show a distinct cleavage site (Fig. II 5F). Our sequencing data revealed two crRNAs at the end of CRISPR cluster 4 and CRISPR cluster 5, respectively, which are not listed in the CRISPR database (Grissa *et al.*, 2007). These crRNAs are processed, even though the 3'-terminal repeat sequences show degeneration with several mutations. Sequencing of RNA samples without T4 PNK treatment revealed the transcription start site of the long CRISPR precursor RNA, whereas mature crRNAs were virtually absent. Mapping of the transcription start sites revealed that all five CRISPR precursors start with a G located 71-73 nt upstream of the first repeat and that the box A promoter element 5'-TTTAAA-3' was conserved (Fig. II 5G and Supplementary Tab. II S1).

Regulatory small RNAs and other sRNA classes

After identification of rRNAs, tRNAs, C/D box sRNAs, H/ACA box sRNAs, SRP RNA, RNase P RNA and crRNAs, the *M. kandleri* sRNome revealed a number of sRNAs that did not belong to any of these RNA families (Supplementary Tab. II S1). These RNAs include intergenic sRNAs and cis-antisense RNAs, which could provide means to regulate gene expression. An Lsm protein (MK0220) to facilitate interaction of regulatory sRNAs with their mRNA targets is present in *M. kandleri*. Unfortunately, the lack of genetic tools for *M. kandleri* does not permit function analyses for these sRNA candidates.

Discussion

Methanopyrus kandleri grows near the upper temperature limit for life. These extreme conditions provide challenges for the production and stability of all cellular macromolecules. Proper folding and the stable maintenance of the folded structure have to be guaranteed for structured non-coding RNA molecules. One strategy is the use of G-C base pairing over A-T base pairing in RNA structure elements (Haas *et al*., 1989). Additionally, posttranscriptional modifications play a crucial role in attenuating thermal RNA denaturation (Kowalak *et al*., 1994). Mass spectrometry analysis revealed an exceptionally diverse population of modified nucleosides in *M. kandleri* tRNAs (Sauerwald *et al*., 2005) and a high degree of 2'-O-methylated nucleosides (Palmer *et al*., 1992). These observations are in agreement with our detection of record numbers of C/D box sRNAs that are required to guide 2'-O-methylation of other RNA molecules and that underlines the importance of RNA modifications in hyperthermophilic organisms.

In addition, the stability of the guide RNAs that are involved in these RNA modifications needs to be guaranteed. C/D box sRNAs contain conserved RNA motifs (box C and box D sequences) that are required to form a kink-turn RNA secondary structure motif on binding of the L7Ae protein (Klein *et al*., 2001; Rozhdestvensky *et al*., 2003). Base pairing of the RNA termini ensures kink-turn formation in eukaryotic C/D box snoRNA (Kiss, 2001). However, in *M. kandleri*, reverse complementary sequences located at the C/D box sRNA termini are either absent or too short to ensure stable RNA stems at the organism's growth temperature (Supplementary Tab. II S1). Therefore, the observed circularization of C/D box sRNA might be used to stabilize these important guide RNAs at elevated temperatures.

Circular C/D box sRNAs have been found in three different archaeal organisms that all share a hyperthermophilic lifestyle (Danan *et al*., 2012; Starostina *et al*., 2004; Randau *et al*., 2012), which suggests that RNA circularization is a conserved feature in this extreme environment. RNA circularization was shown to increase RNA stability and to prevent nuclease degradation (Puttaraju *et al*., 1995), and it has recently been suggested that circular RNAs are common in all three domains of life (Salzman *et al*., 2012). However, the mechanism of archaeal C/D box sRNA circularization is currently still unknown.

A different RNA modification that is currently only known to occur in *M. kandleri* cells is cytidine deamination of tRNA at position 8, catalyzed by the orphan deaminase CDAT8. In all, 30 of 34 tRNA genes are transcribed as primary transcripts with a C at position 8 that is edited to U in the mature tRNA molecule. We took advantage of this unique editing marker to differentiate tRNA precursors in the RNA-Seq data and revealed that C-to-U editing occurs after termini truncation but before intron removal. This allows us to deduce the order of these processing events in a wild-type prokaryotic non-compartmentalized cell. In eukaryotic cells, compartmentalization can

be used to order tRNA precursor maturation events. The tRNA genes are transcribed in the nucleus and tRNA termini are trimmed and matured. In yeast, these tRNAs are exported into the cytoplasm, where subsequently modifications are introduced and introns are spliced (Yoshihisa *et al*., 2003). In prokaryotes, these processes cannot be physically separated. Our data show that C-to-U editing separates termini trimming and intron removal and that *M. kandleri* tRNA maturation follows the order found in some eukaryotic tRNAs.

The purpose of the 30 C-to-U editing events in *M. kandleri* tRNAs is not known. It is plausible CDAT8 activity ensures that mature tRNAs contain the U8 base required for proper folding and function of the tRNA in protein biosynthesis. Therefore, the evolutionary advantage that resulted in the maintenance of C8 bases in the tRNA gene has to be determined by tRNA maturation stages before the editing event. Prevention of virus integration into tRNA genes (Randau and Söll, 2008) or the coordination of tRNA maturation events at extreme temperatures are possible scenarios. It was shown that U8 is usually found as a conserved modified 4-thiouridine nucleotide in bacterial and archaeal tRNAs, and the pathway for the biosynthesis of this modification was recently elucidated for methanogenic archaea (Liu *et al*., 2012). The presence of C8 in tRNA molecules would prevent introduction of this modification and might also coordinate local flexibility of the tRNA before formation of a U8:A14 tertiary base pair is possible, which stabilizes a sharp turn from acceptor stem to the D-stem of the tRNA. This entails that the timing of the occurrence of the U8 base during tRNA maturation in *M. kandleri* is important.

In conclusion, we describe the diverse sRNome for an organism that lives near the upper temperature limit for life and identified exceptionally high numbers of RNA modification guide RNAs and RNAs used for the defence against viruses. Both RNA families function even at extreme temperatures on the basis of annealing with their target nucleic acids. Increased RNA methylation and RNA circularization are suggested adaptations to the hyperthermophilic life style. Finally, the unique tRNA C-to-U event might be used to coordinate tRNA maturation.

Supplementary Material

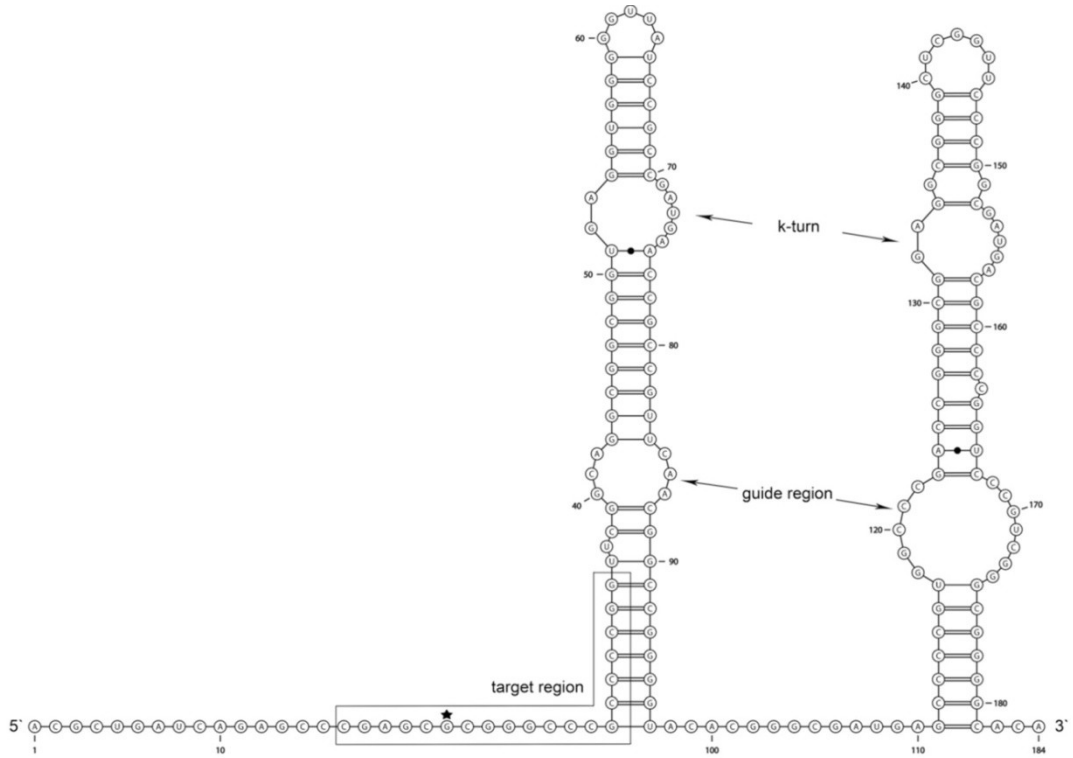

Figure II S1. Proposed structure of H/ACA sRNA 6 and potential mathylation target region
Two kink-turn (k-turn) and two pseudouridylation guide regions are indicated. Target prediction of C/D box sRNAs with PLEXY tool (Kehr et al., 2011) identified a region for 2'-O-methylation (star) in this RNA. This region is targeted by C/D box sRNA 117 with best overall quality score.

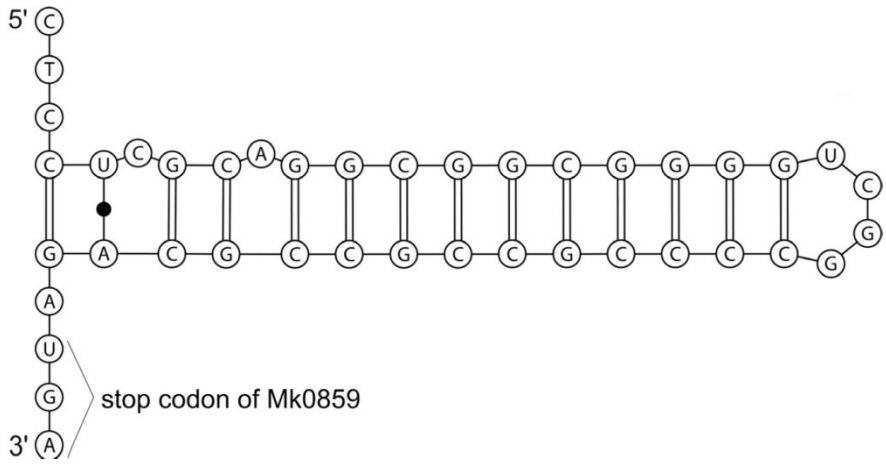

Figure II S2. 3' terminal hairpin of a highly abundant mRNA
The high number of sequencing reads mapping to MK0859 (GAR1, member of the H/ACA ribonucleoprotein complex) correlate with an unusually stable hairpin directly upstream of the stop codon of the MK0859 open reading frame.

Supplementary Table II S1 is available at NAR Online.

Chapter III

Structure and RNA-binding properties of the Type III-A CRISPR-associated protein Csm3

Ajla Hrle[1,†], Andreas AH Su[2,†], Judith Ebert[1], Christian Benda[1], Lennart Randau[2,*], and Elena Conti[1,*]

[1]Structural Cell Biology Department; Max Planck Institute of Biochemistry; Munich/Martinsried, Germany; [2]Max Planck Institute for Terrestrial Microbiology; Karl-von-Frisch-Straße 10, Marburg, Germany; †These authors contributed equally to this work.

RNA Biology, 2013, Vol. 10, No. 11.

Abstract

The prokaryotic adaptive immune system is based on the incorporation of genome fragments of invading viral genetic elements into clusters of regulatory interspaced short palindromic repeats (CRISPRs). The CRISPR loci are transcribed and processed into crRNAs, which are then used to target the invading nucleic acid for degradation. The large family of CRISPR-associated (Cas) proteins mediates this interference response. We have characterized *Methanopyrus kandleri* Csm3, a protein of the type III-A CRISPR-Cas complex. The 2.4 Å resolution crystal structure shows an elaborate four-domain fold organized around a core RRM-like domain. The overall architecture highlights the structural homology to Cas7, the Cas protein that forms the backbone of type I interference complexes. Csm3 binds unstructured RNAs in a sequence non-specific manner, suggesting that it interacts with the variable spacer sequence of the crRNA. The structural and biochemical data provide insights into the similarities and differences in this group of Cas proteins.

Introduction

For a long time, prokaryotic immune systems were believed to be restricted to "innate" immunity mechanisms (e.g., restriction modification systems)(Labrie *et al.*, 2010) and to defense mechanisms that result in cell death upon infection (e.g., toxin-antitoxin systems). (Makarova *et al.*, 2013) In the past decade, however, it has become clear that prokaryotes have evolved sophisticated and diverse adaptive immune systems that memorize previous attacks of foreign genetic elements. These systems consist of clusters of regulatory interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins(Mojica *et al.*, 2000; Mojica *et al.*, 2005; Barrangou *et al.*, 2007). CRISPR-Cas is a nucleic acid-based defense system against mobile genetic elements such as viruses (Barrangou *et al.*, 2007). The CRISPR-Cas machinery distinguishes foreign (non-self) target DNA from (self) targets that are, for example, provided by a host CRISPR locus (Marraffini and Sontheimer, 2010; Wiedenheft *et al.*, 2012).

The central element of CRISPR arrays is the arrangement of DNA sequences of variable length (spacers) derived from foreign genetic elements and separated by short 24-48 nt repeat sequences (Barrangou *et al.*, 2007). Upon infection, these clusters are transcribed into precursor crRNAs (pre-crRNA), which then are processed into mature CRISPR RNAs (crRNA) (Carte *et al.*, 2008; Haurwitz *et al.*, 2010; Richter *et al.*, 2012). The common features of mature crRNAs are the spacer, which identifies the matching target (protospacer) via base pairing, and the 5' -terminal 8 nt repeat tag (psi-tag), which is complementary to the self DNA but not to 2-4 nt short protospacer adjacent motif (PAM) sequences (Marraffini and Sontheimer, 2008). Adjacent to this array are the cas genes (Grissa *et al.*, 2007; Jansen *et al.*, 2002) These encode proteins that are responsible for mediating the CRISPR response and that have a variety of functions, including nucleic acid binding and cleavage (Makarova *et al.*, 2006).

CRISPR-Cas systems have been classified into three main types (I, II, and III) and 10 subtypes by bioinformatic analyses based on their *cas* gene organization, on the sequence and the structure (known or predicted) of the corresponding proteins (Makarova *et al.*, 2011). The three CRISPR types also differ in the composition and mechanisms of their effector complexes (Bhaya *et al.*, 2011). Type I effector complexes are termed Cascade (CRISPR-associated complex for antiviral defense), type II effector complexes consist of a single Cas protein and two RNA molecules, and type III interference complexes are further divided into type III-A (Csm complex targeting DNA) and type III-B (Cmr complex targeting RNA) (Marraffini and Sontheimer, 2008; Hale *et al.*, 2009). In recent years, structural information on Cas proteins has started to provide insights into the molecular mechanisms of crRNA binding and target recognition. The combination of X-ray crystallography (Carte *et al.*, 2008; Garside *et al.*, 2012; Nam *et al.*, 2012; Mulepati *et al.*, 2011; Wang *et al.*, 2011; Reeks *et al.*, 2013) and electron microscopic studies of the type I Cascade

(Wiedenheft *et al*., 2011; Wiedenheft *et al*., 2009; Zhu *et al*., 2012) and of the Type III-B Cmr-complex (Osawa *et al*., 2013; Cocozaki *et al*., 2012; Zhang *et al*., 2012) has shown how some of the Cas proteins interact and bind crRNA. Type I effector complexes are built around a central backbone composed of proteins of the Cas7 family (Wiedenheft *et al*., 2011; Lintner *et al*., 2011). The crystal structure of a Cas7 type I protein has revealed the presence of a central RRM/ferredoxin-like domain with several insertions and a C-terminal extension (Lintner *et al*., 2011). In type I systems, Cas7 oligomerizes upon crRNA binding. In the best-characterized effector complex so far, the *Escherichia coli* Cascade complex, the crRNA binds within a super-helical grove formed by six copies of Cas7 (Wiedenheft *et al*., 2011; van Duijn *et al*., 2012). This helical arrangement has also been observed within other type I systems (Lintner *et al*., 2011; Nam *et al*., 2012; Wiedenheft *et al*., 2011). Despite the absence of significant sequence similarity, bioinformatic analysis has predicted that Cas7-like proteins also exist in type III systems (Koonin *et al*., 2013). Recently, it was shown that a Csm3 (CRISPR-Cas Subtype Mtube, protein 3) from *Staphylococcus epidermidis* binds RNA molecules at multiple sites (Hatoum-Aslan *et al*., 2013). Here, we present the crystal structure and RNA-binding properties of *Methanopyrus kandleri* Csm3. The structural and biochemical analysis of this type III-A Cas protein indicates that Csm3 is a Cas7-like protein capable of binding crRNA, suggesting it forms the backbone of the CRISPR-Cas Type III-A system effector complex.

Results and Discussion

Structure determination of Csm3

We expressed full-length *Methanopyrus kandleri* (*Mk*) Csm3 (351 residues) in *E. coli* and purified it to homogeneity (Fig. III S1A). *Mk* Csm3 yielded crystals in an orthorhombic space group (*C*222) containing two molecules per asymmetric unit and diffracting beyond 2.4 Å resolution. An X-ray fluorescence scan on the crystals showed an unexpected peak at the Zinc excitation energy, suggesting the presence of intrinsically bound Zinc ion in the crystallized protein. We exploited the presence of this anomalous scatterer to solve the structure by single-wavelength anomalous dispersion method (SAD). The phases (obtained from a single bound Zinc ion) were of sufficient quality to build the polypeptide chain. The structure was refined at 2.37 Å resolution to an R_{free} of 21.0%/R_{work} of 18.0% and good stereochemistry (Tab. III S1). The final model includes most of the protein, with the exception of a disordered region between residues 200 and 214. The two independent molecules in the asymmetric unit are very similar, superposing with a root mean square deviation (rmsd) of 0.22 Å for more than 95% of the Cα atoms. Static light scattering experiments of Csm3 in solution showed a mass of 33.3 kDa (Fig. III S1B), consistent with the presence of a monomeric species. Thus, the interaction of the two molecules in the asymmetric unit reflects crystal packing contacts and not a physiological oligomer.

Csm3 is built of four domains organized around a central RRM-like fold

The crystal structure of *Mk* Csm3 reveals a compact architecture that can be described as composed of four domains: the core, the lid, the helical, and the C-terminal domains (Fig. III 1A, in green, blue, red, and yellow, respectively). The core domain has a β1-α1-β2-β3-α2-β4 arrangement of secondary structure elements (Fig. III 1B) with a topology typical of RRM-like and ferredoxin-like folds. Accordingly, the *Mk* Csm3 core domain folds into an antiparallel β-sheet, with two α-helices packed against the concave (back) surface. However, several features set the *Mk* Csm3 core domain apart from canonical RRM-like folds. In the β-sheet, strand β1 is long and highly bent, with a glycine residue (Gly12) at the bending point effectively dividing it into two separate structural elements (strands β1A and β1B, Fig. III 1B). Strands β3 and β4, which sandwich β1, are also elongated (~12 residues), while strand β2 is very short (three residues). The secondary structure elements of the core are connected by loop regions ranging from 2-10 amino-acid residues (between β3-α2 and between α2-β4, respectively) or by larger insertions (between β1-α1, β2-β3, and α1-β2) (Fig. III 1A and B). The 35-residue long β1-α1 insertion contains a short β-hairpin and a one-turn α-helix (αA). On one side, it packs against the 45-residue long β2-β3 insertion, which also contains an α-helix (αG). On the other side, it packs against the α2-β4 loop. Overall, these interactions form the lid domain, which is positioned at the top of the

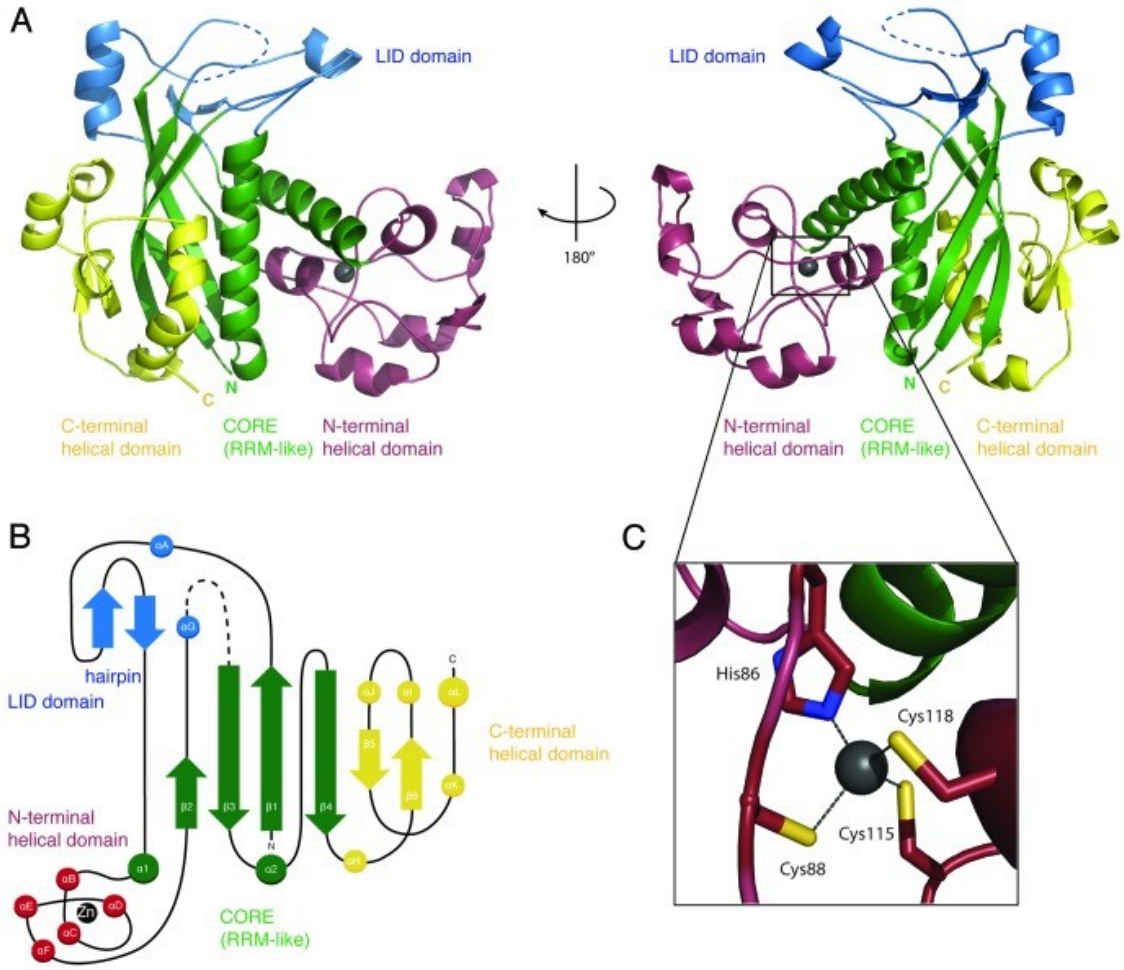

Figure III 1. Structure of *Methanopyrus kandleri* Csm3
(A) The structure of *Mk* Csm3 can be divided into four distinct elements: the core (green) and lid domain (blue), a helical N-terminal (red), and a C-terminal domain (yellow). The structural elements of the core adopt a ferredoxin-like fold with β-α-β-β-α-β arrangement. The core is topologically interrupted by multiple insertions forming the lid and the helical N-terminal domain. The C-terminal domain packs against the core and is of mixed structural composition. The dashed blue line represents the missing disordered region between residues 200 and 214. The two views are related by a 180° rotation as indicated. (B) Topology diagram of Mk Csm3. Helices are represented as circles and β-strands as arrows. The secondary structure elements have been labeled numerically maintaining the nomenclature of RRM domains. The β-strands of the C-terminal domain extending the RRM β-sheet have also been labeled numerically. The additional α-helices have been labeled with letters (αA to αL). (C) A structural zinc ion present in the helical N-terminal domain is shown as a gray sphere, together with the coordinating residues (a cysteine and three histidine residues).

β-sheet and is partially disordered (at a glycine-containing loop in the β2-β3 insertion). The 100-residue long α1-β2 insertion contains five short α-helices (αB to αF) connected by extended segments (Fig. III 1A and B). This insertion forms the α-helical domain and wedges between the two helices (α1 and α2) of the core domain, near the short edge of the β-sheet (i.e., near β2). The helical domain binds a Zinc ion that is buried and is likely to have a structural role in stabilizing the fold of this domain (Fig. III 1C). It connects helices (αD to αE) and is coordinated by His86, Cys88, Cys115, and Cys118. Only the latter two residues are well conserved among Csm3 orthologs (Fig. III S4A). However, other cysteine and histidine residues are present in the α1-β2 insertion of Csm3 from other species (Fig. III S4A). It is thus possible that other Csm3 proteins might have a Zinc-

binding domain in the corresponding region of the structure, albeit with a different topology. Finally, the RRM-like domain is followed by a C-terminal domain (αH-β6-αI-αJ-β5-αK-αL) (Fig. III 1A and B). The C-terminal domain extends the core β-sheet by two antiparallel β-strands (β5 and β6), which flank the long edge of the core domain β-sheet (at β4). It also contains three short α-helices and a long C-terminal α-helix. The C-terminal α-helix packs against α2, at the convex surface of the β-sheet. The short helices (αI and αJ) contact the lid and partly occlude the front surface of the β-sheet. In canonical RRM domains, this front surface features hydrophobic residues that are part of the so-called RNP1 and RNP2 motifs and that bind RNA (Maris *et al*., 2005). However, *Mk* Csm3 lacks the typical solvent-exposed hydrophobic residues that bind RNA in canonical RRM domains. Thus, the *Mk* Csm3 RRM seems to fulfill a structural purpose similar to other previously reported examples (Fribourg *et al*., 2003; Kadlec *et al*., 2004).

Structural comparison of Csm3 with the Cas proteins of the RAMP superfamily

We compared the structure of *Mk* Csm3 with those of Cas5, Cas6, and Cas7, which represent the three major groups of evolutionary distinct RRM-containing proteins in the RAMP (repeat-associated mysterious protein) superfamily (Makarova *et al*., 2011) (Fig. III S2A). *Bacillus halodurans* (*Bh*) Cas5d (PDB ID: 4F3M) has two RRM-like domains adjacent to each other and functions as an endoribonuclease in the pre-processing of crRNA transcripts. The N-terminal RRM-like domain contains the putative endoribonuclease site, which is centered at a histidine residue (Garside *et al*., 2012; Nam *et al*., 2012). *Pyrococcus furiosus* (*Pf*) Cas6 (PDB ID: 3UFC) has an N-terminal RRM-like domain that packs against a twisted β-sheet domain (Wang *et al*., 2011). This RRM-like domain contains an endoribonuclease site that is also centered at a histidine residue, although the exact position differs from that of *Bh* Cas5d. The similarity of *Mk* Csm3 with *Bh* Cas5d and *Pf* Cas6 is limited to the RRM-like domain (Fig. III S2A). Using the structural alignment program SSM as implemented in Coot (Emsley *et al*., 2004) *Mk* Csm3 superposed with *Bh* Cas5d with an rmsd of 3.9 Å over 80 Cα atoms and with *Pf* Cas6 with an rmsd of 4.1 Å over 125 Cα atoms. No prominent histidine residue or possible catalytic triad is however apparent from these structural alignments of Csm3. Consistently, *Mk* Csm3 did not exhibit any prominent endonucleolytic activity with repeat RNA or precursor RNA substrates (data not shown). Bioinformatic analyses have predicted that Csm3 belongs to the Cas7 family of RAMP proteins (Makarova *et al*., 2011b). Superposition of *Mk* Csm3 with the *Sulfolobus solfataricus* (*Sso*) Cas7 (PDB ID: 3PS0) structure results in an rmsd of 4.2 Å over 110 Cα atoms. As with Cas5 and Cas6, the structural similarity with Cas7 is primarily at the RRM-like domain. However, *Mk* Csm3 shares significant overall architectural analogy with Cas7 (Fig. III 2A). In particular, the two proteins have a similar arrangement of domains around the RRM-like fold. Cas7 contains a lid domain, a (mostly) helical domain, and a C-terminal domain at equivalent structural positions as described

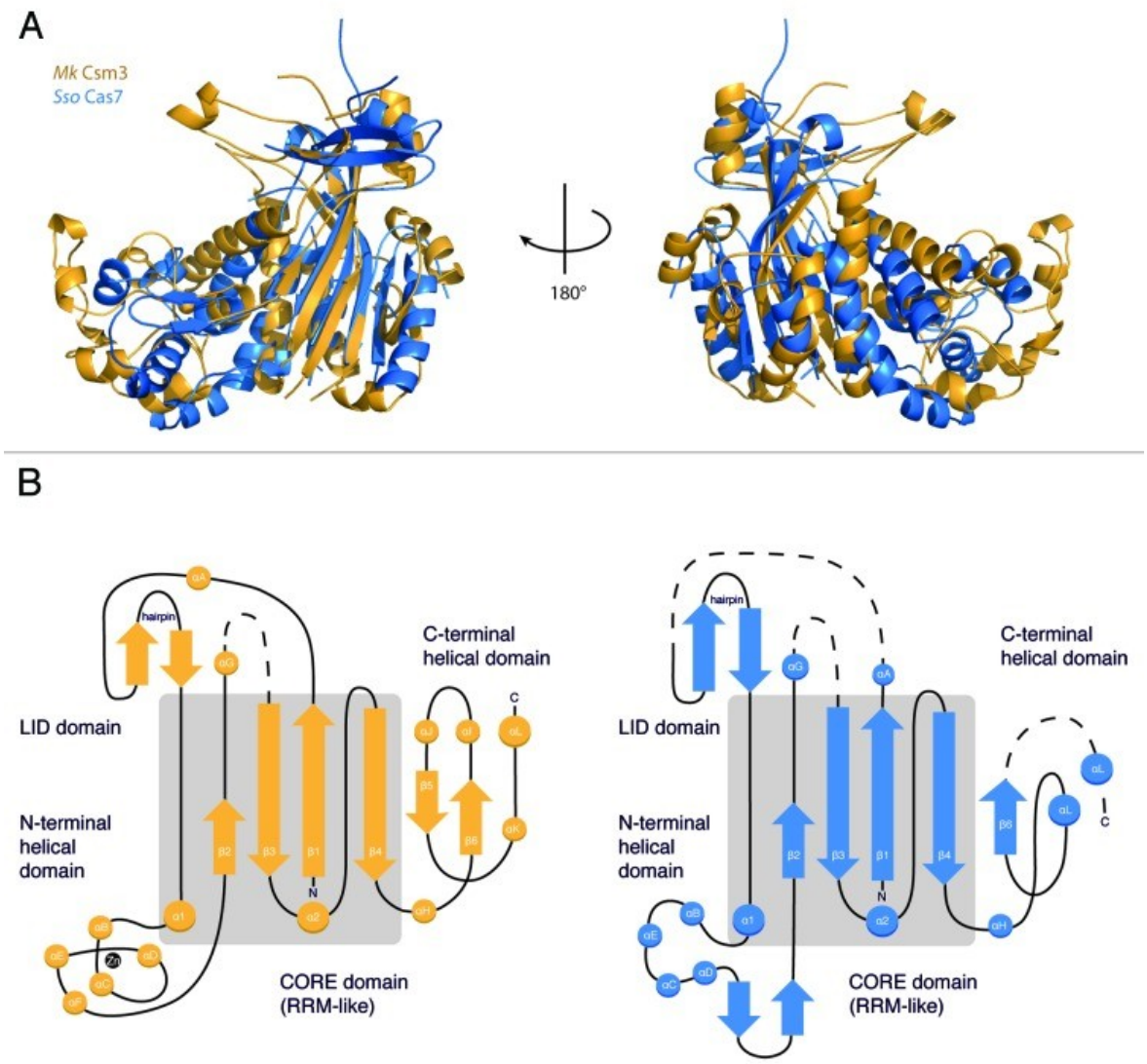

Figure III 2. Structural similarity between Csm3 and Cas7
(A) *Sso* Cas7 (PDB ID: 3PS0, rmsd: 4.2Å, blue) shares the highest structural homology with *Mk* Csm3 (gold) beyond the core domain (gray). Both proteins have a similar arrangement of auxiliary domains surrounding the RRM-like fold, as well as a conserved architecture of the C-terminal domain. (B) Topology diagram of *Mk* Csm3 and *Sso* Cas7 showing the connectivity of the RRM fold relative to the other domains. The topological arrangement of the insertions is similar in both proteins. Similarities in secondary structure elements are highest within the core and low in the auxiliary domains.

above for Csm3 (Fig. III 2B). Although Cas7 is not a Zinc-binding protein and although the exact topological arrangement of secondary structure elements differs from *Mk* Csm3, the overall dimensions and shape of the two proteins is remarkably similar (Fig. III 2A). As Cas7 is a scaffold RNA-binding protein, we assessed whether *Mk* Csm3 might have similar RNA-binding properties.

Csm3 binds single-stranded RNAs in a sequence non-specific manner

We performed electrophoretic mobility shift assays (EMSA) with crRNA substrates that were generated by in vitro transcription (Fig. III 3A). These assays indicated that *Mk* Csm3 binds crRNAs (Fig. III 3A). The *Mk* crRNAs contain a highly conserved repeat sequence of 36 nucleotides that includes a predicted stable stem-loop of 16-18 nucleotides (Fig. III 3B) and a highly conserved

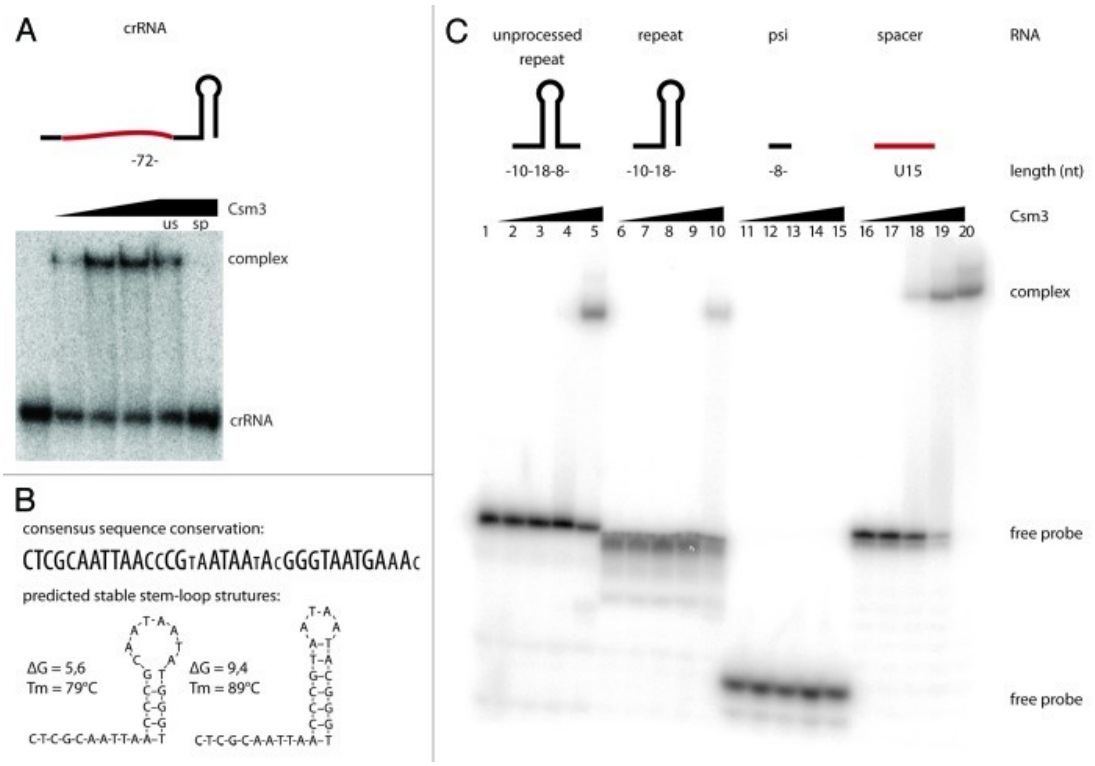

Figure III 3. RNA-binding properties of Csm3
(A) Mk Csm3 binds to a physiological crRNA substrate (left panel). 32P-labeled crRNA transcripts were incubated in the absence or presence of 5 µM, 10 µM, and 20 µM Mk Csm3. (B) Electrophoretic mobility shift assays were performed with the respective [32P]-5'-end labeled RNAs and increasing concentrations of Mk Csm3 (0 µM, 1 µM, 30 µM, 100 µM). Mk Csm3 binds to single-stranded RNA substrates (lane 16-20) but not significantly to the repeat sequences (lanes 1-5 and 6-10). Binding to single-stranded RNA is dependent on length but not sequence (compare lanes 16-20 and 11-15). Weak binding of Mk Csm3 to processed and unprocessed repeat sequences (lanes 1-5 and 6-10, respectively) is likely attributed to the ssRNA overhangs. (C) Methanopyrus kandleri repeat sequence conservation and predicted RNA folding.

eight nucleotide AATGAAA(C/G) motif at the 5' end (psi-tag). They also contain variable spacer sequences ranging from 40-50 nucleotides (Su *et al*., 2013). We dissected which parts of the crRNA are recognized by *Mk* Csm3. In gel-shift assays, Csm3 showed weak binding to processed and unprocessed repeat sequences (Fig. III 3C), but not to its stem-loop structure alone (Fig. III S3A). We tested whether *Mk* Csm3 binds single-stranded RNA, which is present in part of the repeat sequence as well as in the variable spacer. In gel-shift assays, Csm3 bound a 15-mer polyU RNA or 15-mer polyA about 10 times stronger than the repeat sequence (Fig. III 3A). Thus, the length of the single-stranded RNA might affect the strength of the interaction with *Mk* Csm3. *Mk* Csm3 did not exhibit detectable RNA binding toward the 8 nt psi-tag in the gel-shift assays (Fig. III 3C; Fig. III S3A). We conclude that *Mk* Csm3 binds single-stranded RNAs from 15 nucleotides onwards in an apparently sequence non-specific manner and that RNA structures impair binding. This suggests that the variable sequence of the crRNA is bound by *Mk* Csm3, rather than the structured and conserved repeat. To identify the RNA-binding interface, we examined the surface features of *Mk* Csm3 in terms of charge distribution (Fig. III 4A) and evolutionary conservation

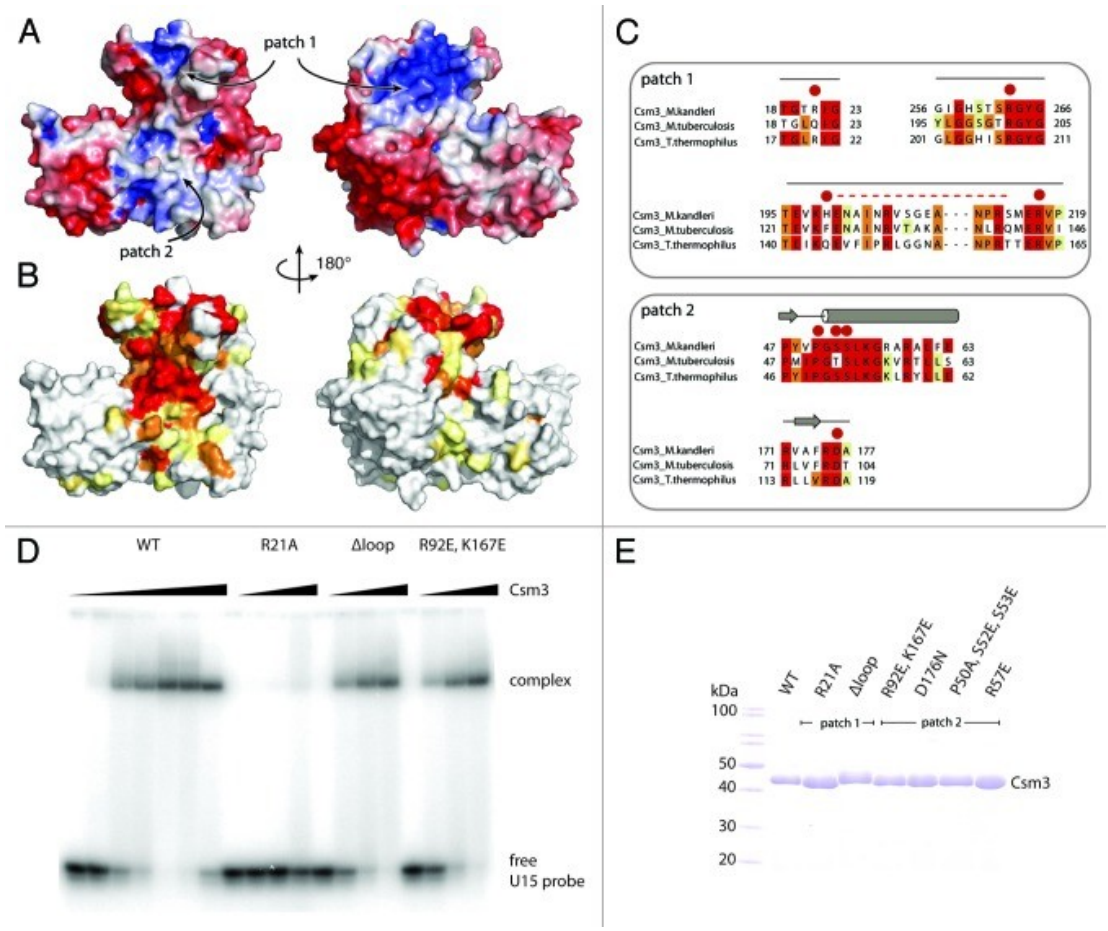

Figure III 4. Identification of Csm3 RNA-binding residues
(A) The structure of Csm3 is shown in surface representations, in the same orientations as in Figure III 1A, colored according to electrostatic potential. Charged patches (blue) are present at the back of the lid domain as well as at the interface between the core and N-terminal helical domain. Negatively charged surfaces (red) are located along the front of the N-terminal insertion and cover the C-terminal domain. Two surface patches discussed in the text (patch 1 and 2) are indicated. (B) Corresponding surface representations of Csm3 colored according to conservation with the Csm3 family. The conservation is based on a comprehensive alignment (Fig. III S4B). Increase in conservation is shown in increasingly darker shades (from to red). No or low conservation (white and yellow) is found in the N-terminal insertion and the C-terminal domain. Highly conserved residues (orange and red) are located within the lid (patch 1) and core domains (patch 2) and coincide with positively charged surfaces (A). (C) Sequence alignments of Csm3 orthologs in regions corresponding to surface patches 1 and 2 (A and B). Residues selected for mutation analysis are highlighted with red dots. The unstructured loop (H199-S214) replaced by a (GS)3 linker is represented as a dashed red line. (D) RNA binding of Csm3 mutants to a single-stranded RNA substrate U15. Wild-type (WT) protein and the double mutation within the core domain (patch2) bind with comparable affinity. Replacement of the unstructured loop (H199-S214) by a -(GS)3- linker does not impair binding, while the single mutation R21A has completely lost RNA binding ability at this condition. (E) Coomassie-stained 12% SDS-PAGE gel of the purified protein samples used in the assays.

(Fig. III 4B and C). The lid domain contains a striking patch (1) of conserved and surface exposed positively charged residues including Arg217, Arg263, and Arg267 (Fig. III S4A). Another positively charged residue, Arg21, is located at the center of this patch and approaches the position of *Sso* Cas7 His160, a residue that has been shown to be important for RNA binding (Lintner *et al*., 2011) A single mutation of *Mk* Csm3 Arg21 to Ala abolished RNA binding in EMSA assays (Fig. III 4D). In the lid domain, the positively charged surface patch is near the disordered glycine-containing loop (Fig. 4C). This loop is conserved (Fig. III 4C) but does not appear to be involved in RNA binding, as

its deletion did not show a significant change in the EMSA assay as compared with the wild-type (WT) protein (Fig. III 4D). Another striking surface patch (2) of *Mk* Csm3 is located at the interface between the lid domain and the helical domain. In particular, helix α1 exposes several conserved residues, including Pro50, Ser52, Ser53, and Arg57 (Fig. III S4B). Mutation of this conserved surface patch (2), however, did not significantly impair RNA binding (Fig. III S4C). We concluded that Csm3 uses the lid domain to bind single-stranded RNA. It is possible to envisage that the other conserved surface patches on Csm3 mediate other types of macromolecular interactions, including protein-protein interactions that form in Csm3-containing effector complexes.

Conclusions

CASCADE/Cmr/Csm complexes share common functionalities, as reflected in their similar composition of proteins. Proteins of the Cas7 family are the backbone of the Type I effector complexes and are involved in interactions with both crRNA and other Cas proteins (Nam *et al*., 2012; Lintner *et al*., 2011; Wiedenheft *et al*., 2011). Computational analyses predicted that Csm3 might fulfill the role of the backbone protein Cas7 in type III interference assemblies.39 Here, we show that *Mk* Csm3 has indeed a remarkably similar architecture as compared with *Sso* Cas7. We found that the structural similarity involves not only the central RRM-like domain, but also insertions at equivalent structural positions in the RRM fold. At the sequence level, however, the two proteins have almost completely diverged. In line with the structural similarity to Cas7, *Mk* Csm3 recognizes crRNA. We found that Csm3 binds to a variable sequence of ssRNA via the flexible insertion that forms a lid on top of the RRM domain. The overall affinity toward RNA is significant yet not strong. It is abolished through mutation of an arginine residue (Arg21Ala) yet hardly reduced when mutating other conserved residues within the positively charged surfaces. It is possible however that this region contributes to RNA binding when in the context of a fully assembled Csm complex. Type III systems further process premature crRNA to mature crRNA; Csm3, together with Csm2 and Csm5, were reported to be required for crRNA 3' termini maturation (Hatoum-Aslan *et al*., 2011). However, in our studies, we could not identify potential catalytic residues nor could we observe nucleolytic activity in biochemical assays. This is in agreement with *S. epidermidis* Csm3 studies indicating that crRNA maturation cleavage events are not performed by the Cas10/Csm complex (Hatoum-Aslan *et al*., 2013). Cas7 proteins oligomerize with a helical arrangement around the crRNA and interact with other Cas proteins of the effector complex, such as the Cas5 in Type I-A (Jore *et al*., 2011). The *S. solfataricus* Cas7 protein was shown to be monomeric and is thought to require Cas5 and crRNA for nucleation and stabilization of its assembly (Lintner *et al*., 2011). In agreement, *Mk* Csm3 also behaves as a monomer in solution and might only oligomerize in the context of the Csm complex. It is possible that the insertion domains that surround the RRM and/or the RRM itself might provide interfaces for protein-protein interactions (Maris *et al*., 2005). We note, however, in contrast to observations with bacterial *S. epidermidis* Csm3, we did not observe binding of RNA molecules in six nucleotide increments for *Mk* Csm3 (Hatoum-Aslan *et al*., 2013). Our structural observations provide a first step toward the structural elucidation of the Csm proteins and their respective role in the surveillance complex. Additionally, the structure will contribute to characterizing the evolutionary relationship within the Cas7 protein family. Further tentative type III members of this family (Cmr1, Cmr4, Cmr6, Csm5) (Makarova *et al*., 2011) remain to be analyzed and classified.

Experimental Procedures

Protein expression and purification

Mk Csm3 wild-type and mutant proteins were expressed as recombinant His- and His-SUMO-tagged fusion protein using BL21-Gold (DE3) Star pRARE (Stratagene) in TB medium and induced overnight at 18 °C. The cells were lysed in buffer A (50 mM Tris pH 7.5, 200 mM NaCl, 10% Glycerol) supplemented with 10 mM Imidazole, DNase, protease inhibitors (Roche) by sonication. Proteins (wild-type and mutants) were purified using Nickel-based affinity chromatography. The His-SUMO tag was cleaved by adding SUMO protease overnight. Proteins were further purified by size-exclusion chromatography (Superdex 75, GE Healthcare) in gel-filtration buffer (buffer A supplemented with 2 mM DTT). Point mutations were introduced by Quick Change site directed mutagenesis according to the manufacturer's instruction (Stratagene).

Crystallization, data collection, structure determination,and analysis

Crystallization was performed at room temperature using hanging drop vapor diffusion method and equal volumes of the protein at 20 mg/ml (gel-filtration buffer) and of crystallization buffer (25% MPD and 50 mM MES 6.0). Crystals were both flash-frozen directly from the crystallization drop as well as subjected to further dehydration (increasing amounts of MPD up to 60%) and diffracted beyond 2.4 Å. All diffraction data was collected at 100 K at the beamline PXII of the Swiss Light Source (SLS) synchrotron and processed using XDS (Kabsch, 2010). The structures were determined using the native data and Zn-SAD phases to build an initial model. This was then used as a search model for molecular replacement of higher resolution data using Phaser (McCoy *et al*., 2010). Model building was performed manually with the program Coot38 and refined with PHENIX (Adams *et al*., 2007). The data collection and refinement statistics are summarized in Table III S1. Figures were prepared using PyMOL (http://www.pymol.org).

Biochemical assays

The RNA molecules U15, A10, A15, A20, A40 were synthesized (Purimex). The crRNA (locus 5, spacer 5) was produced by in vitro run-off transcription and purified by elution of the crRNA transcript from a polyacrylamide gel as described (Richter *et al*., 2012). The RNA molecules were 5'-labeled with T4 polynucleotide kinase (New England Biolabs) and γ-[32P] ATP (Perkin- Elmer). For the gel-shift assays, 0.5 pmol labeled RNA was mixed with 1 µM, 10 µM, 30 µM, 100 µM protein in a 10 µL reaction containing 20 mM Hepes at pH 7.5, 100 mM KOAc, 4 mM Mg(OAc)2, 0.1% (vol/vol) NP-40, and 2 mM DTT. Fifteen ng/µL (500 fmol/µl = 500x molar excess) yeast tRNA mix (Amicon) were used as non-specific and 15 ng/µl unlabeled crRNA transcripts were used as specific competitor molecules. The mixtures were incubated for 20 min at 55 °C before adding 2

µL 50% (vol/vol) glycerol containing 0.25% (wt/vol) xylene cyanole. Samples were run on a 8% (wt/vol) polyacrylamide gel at 4 °C and visualized by phosphorimaging (GE Healthcare).

Supplementary Materials

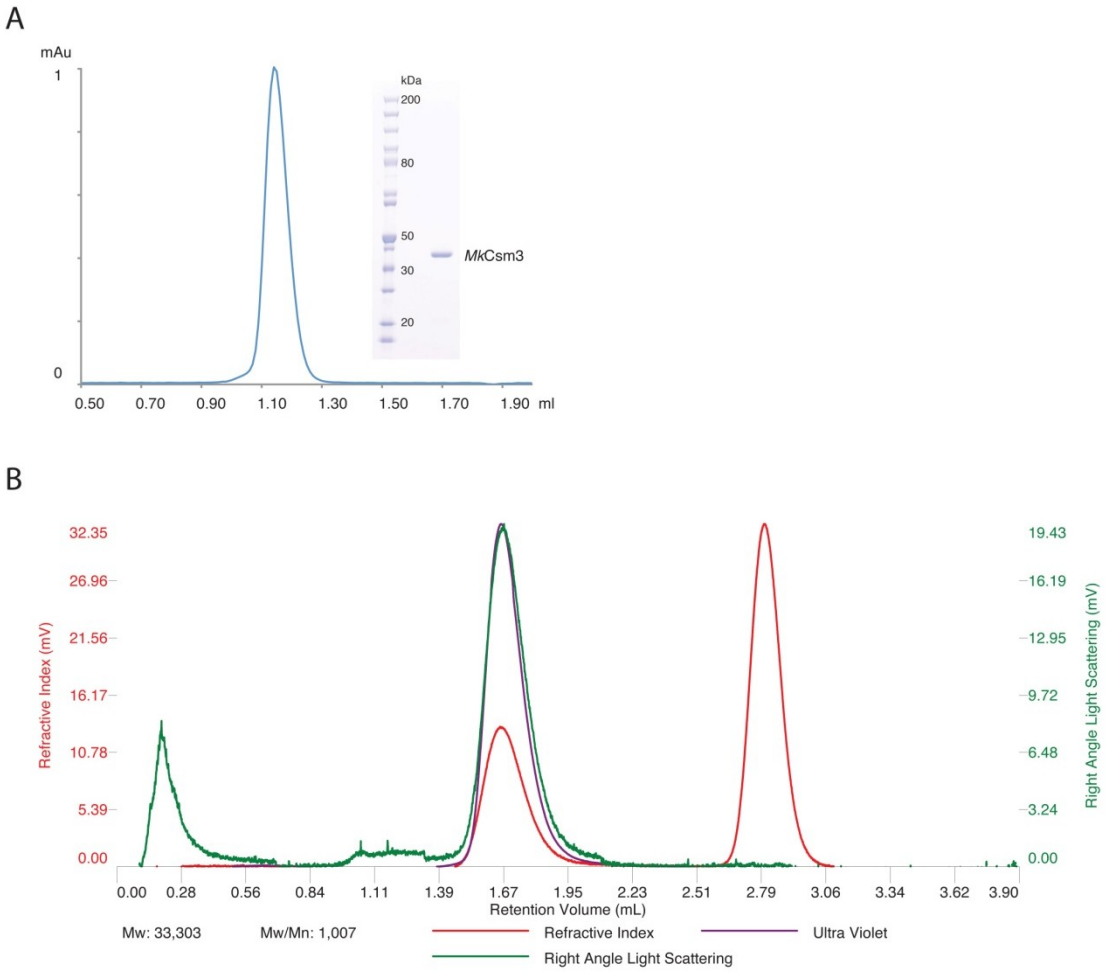

Figure III S1. Analysis of the protein used in biochemical assays
(A) Coomassie-stained SDS-PAGE and size exclusion chromatography elution profile of purified *Mk* Csm3. (B) Static light scattering (SLS) chromatogram and values underline the monomeric behaviour in solution.

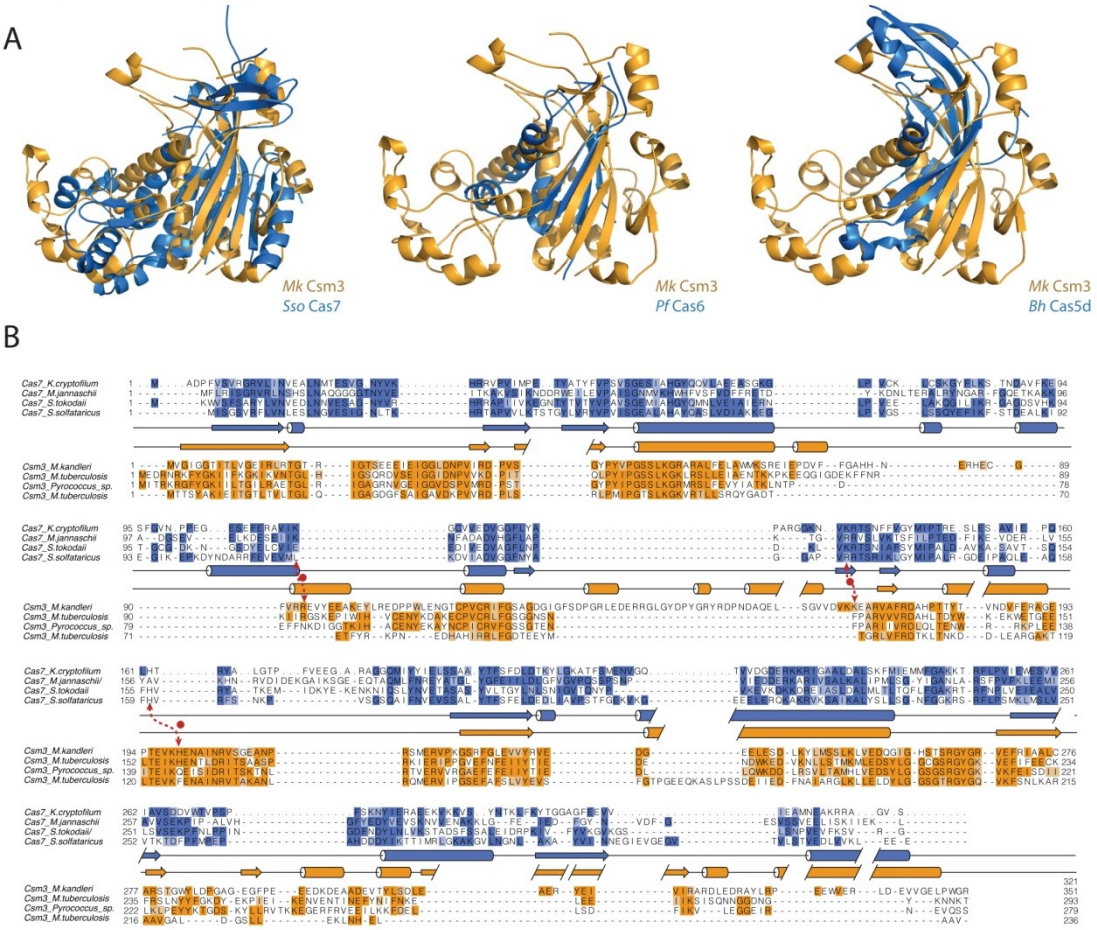

Figure III S2. Structure and sequence-based similarities of Csm3 with key Cas proteins
(A) Structure-based homology modeling of Mk Csm3 (gold) with known structures of CRISPR proteins (blue) structurally manually aligned based on the DALI server output and their common feature, the ferredoxin fold. The protein is shown in the same orientation to that used in Figure III 1. The ferredoxin-like fold is the region of highest conservation. The homology of the N-terminal ferredoxin fold domains of Bh Cas5d (PDB ID: 3KG4, rmsd: 3.9) and *Pf* Cas6 (PDB ID: 3UFC, rmsd: 4.1Å) is restricted to the core fold of *Mk* Csm3. *Sso* Cas7 (PDB ID: 3PS0, rmsd: 4.2Å) shares the highest structural homology with *Mk* Csm3 beyond the core domain. Both proteins have a similar arrangement of insertions within the ferredoxin-like fold domain as well as the overall architecture of the C-terminal domain. (B) Structure-based sequence alignment of *Mk* Csm3 (gold) and *Sso* Cas7 (blue). The alignment includes sequences from representative species of both families. Lighter letters denote residues identical in more than one third of the species considered; darker letters identify residues identical in more than two thirds of the species. The colors are based on the comprehensive alignment in Supplementary Figure III 4A. Secondary structure elements are shown above the sequences with cylinders for α-helices and arrows for β-strands. Sequence conservation between *Mk* Csm3 and *Sso* Cas7 is mostly restricted to structural residues that define the core domain and is low within adjacent regions.

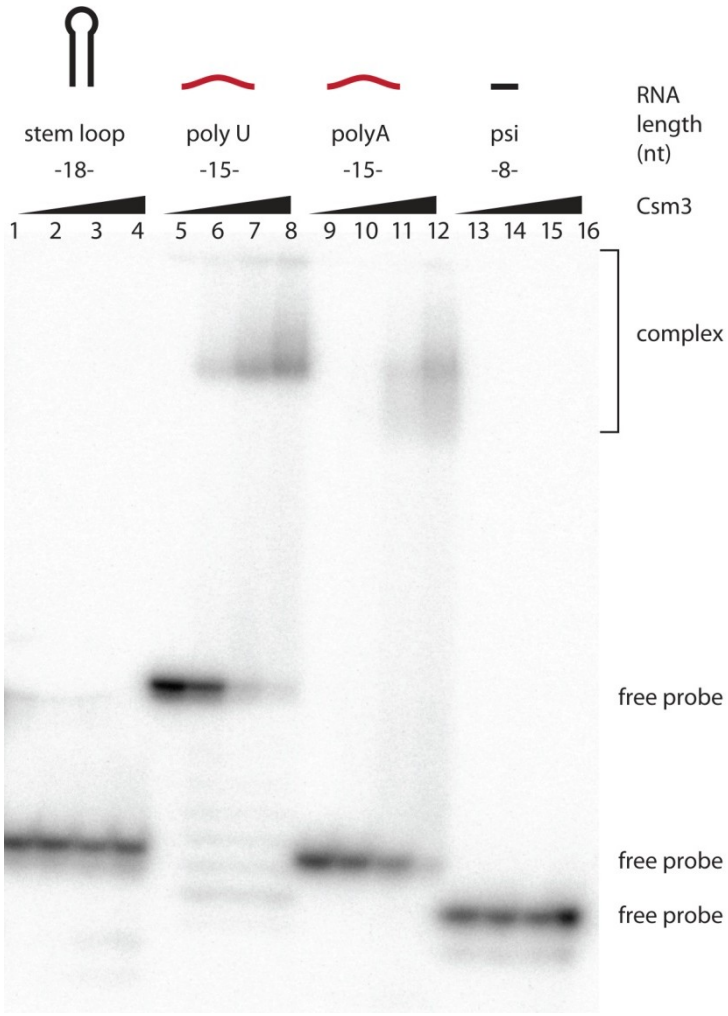

Figure III S3. Dissecting RNA-binding properties of Csm3
Electrophoretic mobility shift assays (EMSA) were carried out with the respective [32P]-5' end labeled RNAs and increasing concentrations of Mk Csm3. (A) Mk Csm3 does not specifically recognize the stem-loop (lanes 1-4) or psi-tag (lanes 13-16) of the conserved repeat sequence, yet shows sequence unspecific binding to U15 (lanes 5-8) and A15 (lanes 9-12).

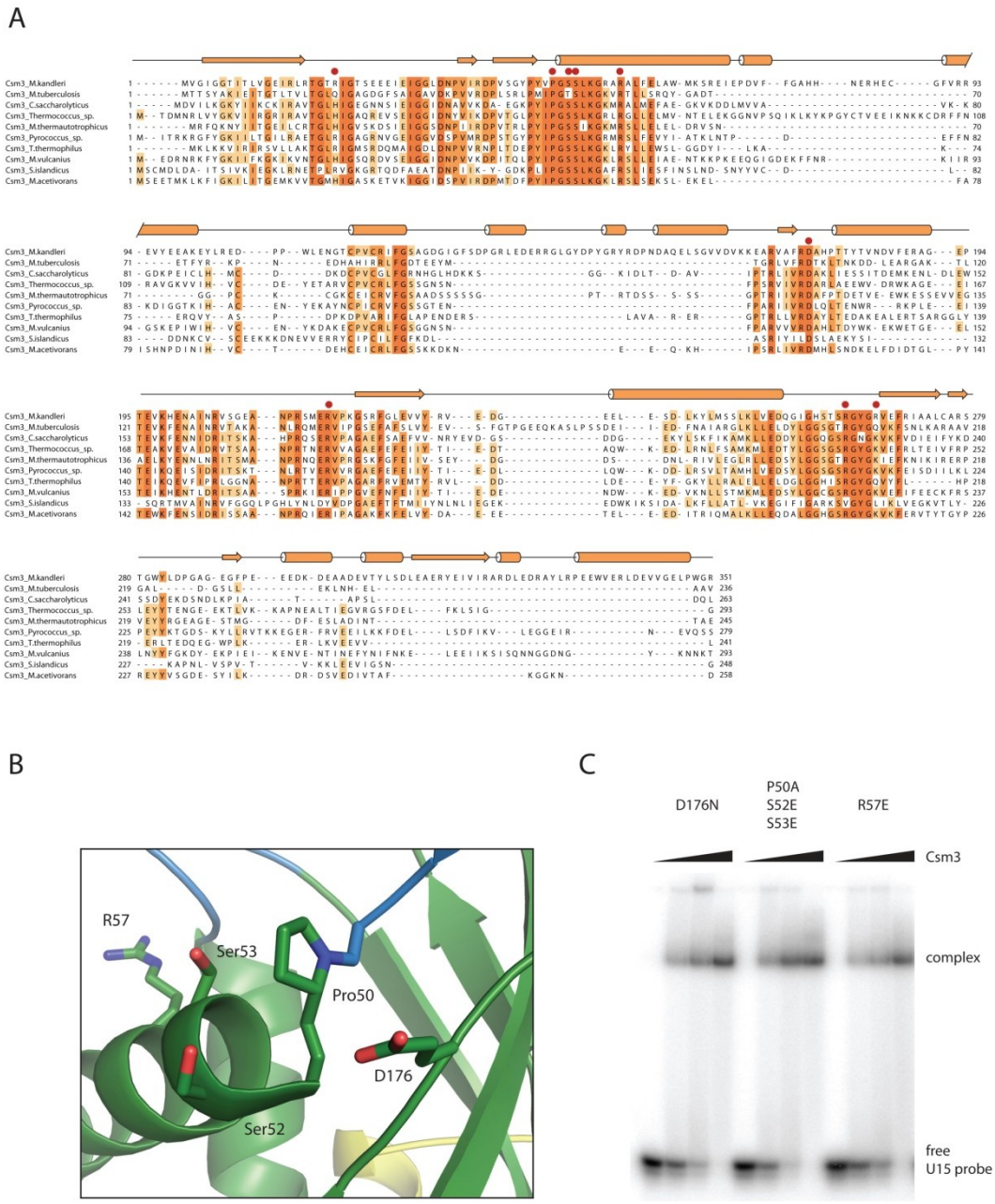

Figure III S4. Mapping of Csm3 RNA-binding residues

(A) Comprehensive sequence alignment of *Mk* Csm3 within the Csm-family. The alignment includes sequences from representative species. Lighter letters denote residues identical in more than one third of the species considered; darker letters identify residues identical in more than two thirds of the species. (B) A close-up view of conserved residues (Pro50, Ser52, Ser53, Arg57 and D176) of the core domain, which were selected as representative mutations within the core (patch2). (C) Electrophoretic mobility shift assays (EMSA) were carried out with the respective [32P]-5' end labeled RNAs and increasing concentrations of *Mk* Csm3. Mutations located within the core (patch2) bind U15 ssRNA with comparable affinity to WT protein.

Table III S1. Data collection and structure refinement statstics of Csm3

	Data collection	
	Native	**Native, Zn-SAD**
Resolution range (Å)	69.49 - 2.37 (2.49 - 2.37)	49.86 - 3.20 (3.00-3.08)
Unit cell(Å)[a]	a = 95.75	a = 70.51
	b = 101.02	b = 70.51
	c = 174.17	c = 193.36
Total reflections	308429	672581
Unique reflections	34685	38958
Multiplicity[a]	8.9 (7.1)	56.3 (52.4)
Completeness (%)[a]	99.5 (95.2)	99.9 (99.7)
Mean I/sigma(I) [a]	17.10 (2.72)	32.72(1.76)
	Refinement	
B-factor	42.57	
R-factor(%)	18.09	
R-free(%)	21.34	
RMS(bonds) (Å)	0.005	
RMS(angles) (Å)	0.85	
Ramachandran favored (%)	95	
Ramachandran outliers(%)	0.15	

[a]Values in parentheses correspond to the highest resolution shell.

Chapter IV

Analysis of Cas7 homologue proteins in *Methanopyrus kandleri*

Andreas A. H. Su[1], Ajla Hrle[2], Elena Conti[2] and Lennart Randau[1]

[1]Prokaryotic Small RNA Biology Group; Max Planck Institute for Terrestrial Microbiology; Karl-von-Frisch-Straße 10, Marburg, Germany

[2]Structural Cell Biology Department; Max Planck Institute of Biochemistry; Am Klopferspitz 8, Munich/Martinsried, Germany

Unpublished manuscript

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins are essential components of a prokaryotic, adaptive immune system that targets mobile genetic elements. CRISPR loci encode CRISPR RNAs (crRNAs) that are used as guide sequences for Cas protein interference complexes to target foreign nucleic acids. Multiple copies of Cas7 proteins are involved in crRNA binding and build up the helical backbone of interference complexes in CRISPR-Cas type I systems.

Here, we report the analysis of the Cas7 homologue proteins Csm3 and Cmr4 of the type III-A and III-B system from *Methanopyrus kandleri*. In nuclease assays, Csm3 exhibited a potential crRNA processing activity that might be independent of RNA binding. Furthermore, crystallization of the Csm3 homologue Cmr4 was attempted in order to analyze structural differences in the backbone proteins of type III-A and III-B systems. Finally, EMSAs and footprinting analysis indicate a non-specific recognition of crRNAs by Csm3 via the lid- and core domain and provide new insights into the crRNA-Csm3 interaction.

Introduction

Viruses are the most abundant biological agents on earth and outnumber prokaryotes by approximately 10-fold (Suttle, 2007). This leads to a constant battle between prokaryotes and viruses that provides a selective pressure for the evolution of various prokaryotic immune systems. Systems like abortive infection, restriction-modification and adsorption inhibition provide innate immunity (Samson *et al.*, 2013). However, prokaryotes have also evolved an adaptive immune system based on clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins, referred to as the CRISPR-Cas system (Jansen *et al.*, 2002; Barrangou *et al.*, 2007). CRISPR-Cas systems can be found in approximately 45% of bacteria and nearly all archaea and provide immunity against viruses and plasmids (Makarova *et al.*, 2013; van der Oost *et al.*, 2014).

CRISPR loci consist of variable sequences derived from mobile genetic elements (spacers), interspaced by repetitive sequences (repeats) and serve as the immunological memory of the CRISPR-Cas system (Mojica *et al.*, 2005; Blotin *et al.*, 2005; Barrangou *et al.*, 2007). Cas genes are often located in proximity of CRISPR loci and encode for Cas proteins that play an essential role in the CRISPR-Cas immunity pathway. According to the composition of Cas marker genes and differences in the immunity mechanism, CRISPR-Cas systems can be classified into three major types (type I-III) and various subtypes (Makarova *et al.*, 2011). In general, the CRISPR immunity pathway involves the three steps adaptation, expression and interference. In the adaptation step, the foreign DNA is recognized and a small DNA fragment is integrated into the CRISPR locus as a new spacer. This process is conserved among the three major CRISPR-Cas types and is mediated by a complex of the proteins Cas1 and Cas2 (Nunez *et al.*, 2014). The expression stage involves the transcription of the CRISPR locus, resulting in a long precursor CRISPR RNA (pre-crRNA). Processing of the pre-crRNA within the repeat sequences is mediated by either Cas6 (in type I and III systems) or RNaseIII (in type II systems) and results in mature CRISPR RNAs (crRNA) (Carte *et al.*, 2008; Haurwitz *et al.*, 2010; Deltcheva *et al.*, 2011). In the final interference step, the crRNAs associate with Cas proteins to build interference complexes that use the crRNAs as guides to target and degrade foreign nucleic acids (Brouns *et al.*, 2008; Sapranauskas *et al.*, 2011; Rouillon *et al.*, 2013).

The best studied interference complex is the CRISPR associated complex of antiviral defense (CASCADE) of the type I system (Brouns *et al.*, 2008). Cryo-electron microscopy (cryo-EM) structures of type I complexes revealed a seahorse-like structure and a helical backbone composed of six Cas7 proteins (Wiedenheft *et al.*, 2011; Hochstrasser *et al.*, 2014). The Cas7 protein exhibits a crescent-like structure with an RNA-recognition motif (RRM) and is responsible for the binding of the crRNA guides (Lintner *et al.*, 2011; Wiedenheft *et al.*, 2011; Hochstrasser *et*

al., 2014). Type III interference complexes are divided into DNA targeting Csm complexes (type III-A) and RNA targeting Cmr complexes (type III-B) (Marraffini and Sontheimer, 2008; Hale *et al*., 2009). Cryo-EM structures of the Csm- and Cmr complexes revealed a similar overall architecture like the CASCADE with helical, multicopy backbone proteins that are represented by Csm3 and Cmr4 respectively (Rouillon *et al*., 2013; Staals *et al*., 2013; Spilman *et al*., 2013). The crystal structure of Csm3 has been solved and revealed a structural homology to Cas7 with a four domain architecture composed of a core, a lid, a helical and a C-terminal domain with a core RRM-like motif. Furthermore, electrophoretic mobility shift assays (EMSA) revealed a non-specific RNA binding activity of *Methanopyrus kandleri* (MkCsm3) that is in line with its function as the Csm complex backbone protein (Hrle *et al*., 2013).

In type III systems, a secondary processing site of crRNAs has been detected yielding two mature crRNAs species that differ by 6 nucleotides (Hale *et al*., 2012; Hatoum-Aslan *et al*., 2011). Csm3 has been shown to be required for this processing step and functions as a ruler protein by binding the crRNA in a six nucleotide periodicity (Hatoum-Aslan *et al*., 2011; Hatoum-Aslan *et al*., 2013). However, the RNase responsible for secondary crRNA processing in type III CRISPR-Cas systems still remains to be determined.

Here, we analyze the interaction between crRNAs and the MkCsm3 protein. EMSA analysis of MkCsm3 mutants revealed an involvement of the core and lid domain in crRNA binding and footprinting analysis indicates that no specific region of the crRNA is recognized. The potential of Csm3 to be involved in secondary processing of crRNA was analyzed by nuclease assays and is discussed. Finally the crystallization of the *Methanopyrus kandleri* Cmr4 (MkCmr4) protein was attempted to analyze structural differences in the backbone proteins of type III-A and III-B systems.

Material and Methods

Protein expression and purification

MkCsm3 and MkCmr4 were amplified by PCR of genomic DNA and ligated into pET43-1b (Novagen) and pTrcHis (Life Technologies), respectively. The proteins were produced in the *Escherichia coli strain* Rosetta2 (Strategene) for 3 hours with 1 mM IPTG at 37°C, resulting in MkCsm3-6xHis and 6xHis-Xpress-MkCmr4 fusion proteins. Cells were disrupted in buffer 1 (50 mM TrisHCl pH 7, 200 mM NaCl, 10 mM beta-mercaptoethanol and 4 mM $MgCl_2$) by sonication and cleared by centrifugation (45 000x g, 30 min., 4°C). Expressed proteins were purified from the cleared supernatant via HisTrap HP nickel affinity column (GE Healthcare) on a fast protein liquid chromatography system (Äkta) with a linear imidazole gradient from 0 - 500 mM in buffer 1. Protein containing fractions were further purified by gel filtration chromatography using a HiLoad 16/60 Superdex 75 column (GE Healthcare). Untagged MkCsm3 was obtained from our collaboration partner Ayla Hrle from the laboratory of Elena Conti (MPI Martinsried, Germany) and expressed as described previously (Hrle *et al*., 2013).

Production and labeling of crRNA substrates

For generation of crRNA substrates, two complementary oligonucleotides were synthesized (MWG Eurofins) that contained the respective sequence of crRNA 2.1 or 5.6 (CRISPR locus 2, spacer 1 or CRISPR locus 5, spacer 6) and the sequence of the T7 RNA polymerase promoter. The sequence of crRNA 2.1 was additionally fused to a minimal *cis*-acting hammerhead ribozyme sequence at the 5'-end (Carbonell *et al*., 2011). 1 µmol of each oligonucleotide was phosphorylated with 50 U of T4 PNK (NEB) in the provided buffer, supplemented with 10 mM ATP for 1 h at 37°C. The phosphorylated oligonucleotides were hybridized by incubation for 5 min at 95°C, followed by a slow cooling to room temperature for 2 h. The hybridized oligonucleotides were cloned via EcoRI and HindIII restriction sites into a pUC19 vector. For *in vitro* transcription, 20 ng/mL of a HindIII cleaved plasmid was used in a reaction containing 40 mM HEPES/KOH, pH 8.0, 22 mM $MgCl_2$, 5 mM dithiothreitol, 1 mM spermidine, 4 mM of ATP, CTP, GTP and UTP and 1 mg T7 RNA polymerase for 3 h at 37°C. The hammerhead self cleavage reaction was induced by dilution with 4 volumes 30 mM $MgCl_2$ in DEPC H_2O and incubation for 1 h at 60°C. The produced crRNA was purified from exessive pyrophosphate by centrifugation (5 min, 15 000x g) and further purified by phenol/chloroform extraction (pH 5.2) followed by ethanol precipitation. Subsequently the crRNAs were mixed with 2x formamide buffer (95% formamide, 5 mM EDTA, pH 8.0, 2.5 mg bromophenol blue, 2.5 mg xylene cyanol), heated at 95°C for 5 min and separated by denaturing PAGE (8 M urea, 1x TBE, 12% polyacrylamide). The gel was stained with toluidine staining solution (40% methanol, 1% acetic acid, 0.1% toluidine blue). The corresponding bands

were cut out of the gel, eluted on ice over night in 500 µl elution buffer (20 mM TrisHCl, pH 7.5, 250 mM sodium acetate, 1 mM EDTA, pH 8.0, 0.25% SDS) and EtOH precipitated. For nuclease, footprinting and binding assays, 5 pmol of the purified crRNA substrates were 5'-end-labeled with [γ-^{32}P]-ATP (5000 ci/mmol, Hartmann Analytic) using T4 PNK (Ambion) for 2 h at 37°C. The labeled crRNAs were separated in denaturing PAGE and visualized by phosphorimaging. Finally, the detected RNA bands were cut out, gel eluted and EtOH precipitated as described before.

Electrophoretic mobility shift assays

In electrophoretic mobility shift assays (EMSA), 2000 cpm of labeled crRNA 2.1 substrates were incubated with 5 µM or 10 µM MkCsm3 in binding buffer (50 mM TrisHCl pH 7, 200 mM NaCl, 10 mM Beta-Mercaptoethanol) for 20 min at 55°C. The reactions were stopped on ice and mixed with 5x DNA loading dye (Quiagen). Subsequently, the RNA was separated by non-denaturing PAGE (10% polyacrylamide, 1x TBE) and visualized by phosphorimaging.

RNA footprinting assays

For RNA footprinting assays, 3500 cpm of the 5'-end-labeled crRNA 5.6 substrate was incubated with RNase T1, RNase V1 or nuclease S1 (Ambion) according to the manufacturer's instruction. For footprinting assays without $MgCl_2$, the reactions were incubated in binding buffer (50 mM TrisHCl pH 7, 200 mM NaCl, 10 mM beta-mercaptoethanol) instead of the provided structure buffer (Ambion). Prior to the nuclease reactions the crRNA was pre-incubated with up to 60 µM of MkCsm3-6xHis in binding buffer or 1x structure buffer (Ambion) at 65°C for 5 min. The resulting cleavage products were separated on 30 cm x 40 cm sequencing gels (8 M urea, 1x TBE, 12% polyacrylamide) by denaturing PAGE (Gibco BRL) and visualized by phosphorimaging.

Nuclease activity assays

For initial detection of RNA cleavage activity (Fig. IV 2A and B), 3500 cpm of 5'-end-labeled crRNA substrate 5.6 were incubated with up to 60 µM MkCsm3-6xHis in binding buffer (50 mM TrisHCl pH 7, 200 mM NaCl, 10 mM Beta-Mercaptoethanol) supplemented with 4 mM $MgCl_2$, $MnCl_2$, $CuCl_2$, $NiCl_2$, $CoCl_2$, $ZnCl_2$ or $CaCl_2$ for 5 min at 65°C. For subsequent assays (Fig. IV 2C), 3500 cpm of 5'-end-labeled crRNA substrate 2.1 (hammerhead cleaved) were incubated with 2,5 µM or 10 µM of untagged MkCsm3 in binding buffer supplemented with 4 mM $MgCl_2$ for 20 min at 50°C. The reactions were stopped on ice, phenol/chloroform extracted (pH 5.2) and ethanol precipitated. The precipitated RNA was resuspended in 10 µl formamide buffer (95% formamide, 5 mM EDTA, pH 8.0, 2.5 mg bromophenol blue, 2.5 mg xylene cyanol), heated at 95°C for 5 min. Subsequently, the products were separated on 30 cm x 40 cm sequencing gels (8 M urea, 1x TBE, 12% polyacrylamide) by denaturing PAGE (Gibco BRL) and visualized by phosphorimaging.

Results and Discussion

Analysis of the MkCsm3 - crRNA interaction

Electron microscopy (EM) analyses of a CRISPR-Cas type III-A complex in *Sulfolobus solfataricus*, combined with mass spectrometry (MS) indicated that multiple copies of Csm3 build the crRNA-binding backbone of the complex (Rouillon *et al*., 2013). This was confirmed by a crystal structure of Csm3 from *M. kandleri* (MkCsm3) that reveals a structural homology to the CRISPR-Cas type I backbone protein Cas7 and showed non-specific binding of RNA molecules. The MkCsm3 structure exhibits four different domains including a core-, lid-, C-terminal- and zinc binding N-terminal domain. The core domain contains the typical β1-α1-β2-β3-α2-β4 secondary structure elements of RRMs. The lid domain includes an unstructured, positively charged loop and has been shown to contribute to the binding of non-specific poly U RNAs (Hrle *et al*., 2013).

To analyze which domains of MkCsm3 are involved in the binding of a more physiological crRNA substrate, we performed electrophoretic mobility shift assays (EMSAs) with *in vitro* transcribed crRNAs. Therefore, the crRNA substrates were incubated with MkCsm3 mutants that were designed to disrupt RNA binding activity. The locations of mutations were chosen based on alignments of Csm3 proteins from different organisms and structural alignments of Csm3 to Cas7 proteins (Fig. III 4C). A mutant with a complete deletion of the positively charged loop and a set of single, double or triple amino acid substitutions were designed and produced by the lab of Prof. Dr. Elena Conti and sent to us (Δloop, H199A, R92E/K167E, D176N, S52/53E, P50A/S52E/S53E, and R57L). EMSAs revealed a strongly reduced crRNA binding affinity of the mutants Δloop, and the two mutants with amino acid changes in the core domain R92E/K167E and D176N (Fig. IV 1A). This indicates that the unstructured, positively charged loop and the core domain, including the RRM, might be involved in the binding of crRNAs. The reduced crRNA binding of the mutants R92E/K167E and D176N is in contrast a previous study, where full binding of a 15-mer polyU RNA was detected with these mutants under slightly different buffer conditions (Hrle *et al*., 2014). This might indicate that MkCsm3 interacts more specifically with a crRNA than with a polyU RNA and therefore a mutation has a stronger effect on the binding of crRNAs than on the binding of non-specific polyU RNAs.

To analyze whether MkCsm3 specifically recognizes crRNA substrates at a certain region (e.g. the conserved repeat sequence), we performed RNA footprinting analyses. Therefore, an *in vitro* transcribed crRNA substrate was incubated with rising concentrations of wild type MkCsm3 and RNase T1. Subsequently the protection of certain RNA regions against RNase T1 cleavage was analyzed. In the absence of $MgCl_2$, footprinting analyses revealed no significant protection against RNase T1 cleavage. This indicates that MkCsm3 has no specific or strong binding site on the crRNA

(Fig. IV 1B). Unexpectedly, the same footprinting analysis in the presence of $MgCl_2$ revealed additional bands upon addition of MkCsm3 indicating a potential RNA cleavage ability (Fig. IV 1 C). Further experiments with RNase V1 and nuclease S1 did not reveal a protection against cleavage upon addition of MkCsm3.

These results are in line with the published non-specific binding of Csm3 with RNA molecules (Hatoum-Aslan *et al*., 2013; Hrle *et al*., 2013) and indicate that Csm3 is does not provide the specificity towards crRNAs in Csm complexes. However, there is still the possibility that MkCsm3 needs its multimeric state and other Csm complex proteins to bind specifically. A reconstitution of the whole Csm complex proteins (Csm1 - Csm6) followed by biochemical binding assays would be needed to test this possibility. As the Csm complexes of *S. epidermidis* and *S. solfataricus* have been purified recently, we decided to skip attempts to accomplish a reconstitution of the *M. kandleri* Csm complex (Hatoum-Aslan *et al*., 2013; Rouillion *et al*., 2013). Proteins that are candidates for specific recognition of crRNAs in type III-A systems are Csm1 and Csm4. EM and MS analyses showed that Csm1 and Csm4 build a sub-complex with a cleft that might be able to recognize a crRNA at its 5'-end. Subsequently, Csm3 could be recruited and bind the crRNA non-specifically to serve as a scaffold for the crRNA (Rouillion *et al*., 2013).

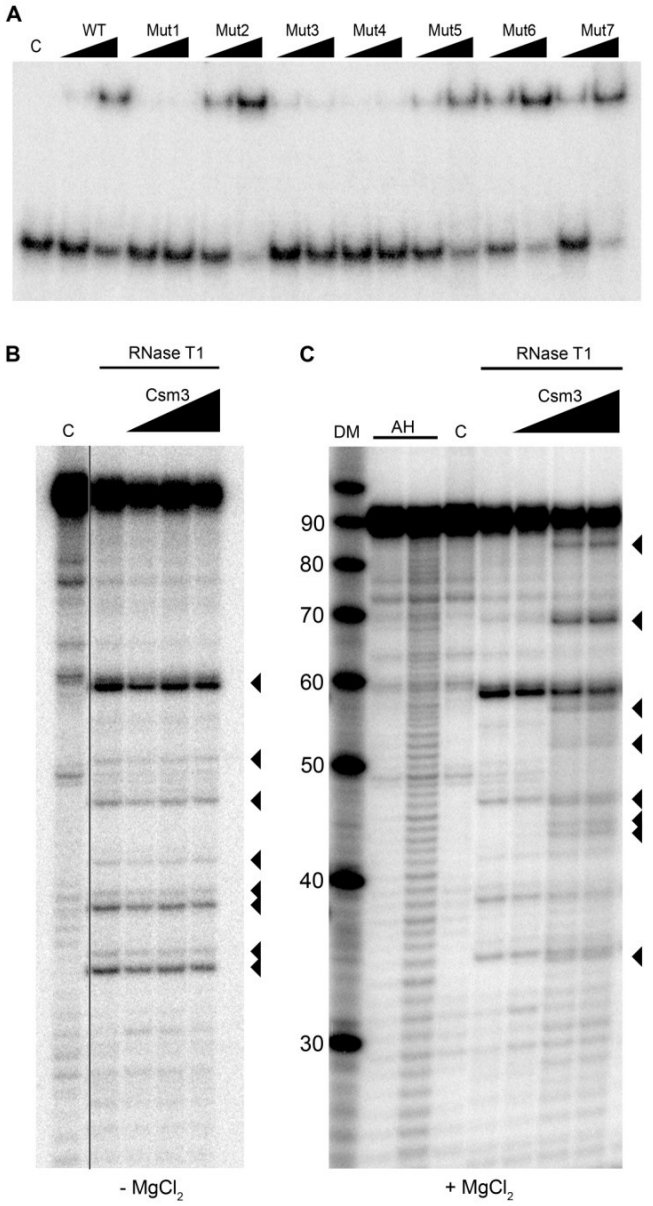

Figure IV 1. Analysis of the MkCsm3 - crRNA interaction

(A) Radiolabeled crRNA (CRISPR locus 2, spacer 1) was incubated in the absence (C) or presence of wild type (WT) or mutant MkCsm3 (Mut1: Δloop; Mut2: H199A; Mut3: R92E/K167E; Mut4: D176N; Mut5: S52/53E;Mut6: P50A/S52E/S53E; Mut7: R57L) in concentrations of 2.5 µM and 10 µM. Formation of stable complexes was determined by EMSA.

(B) The radiolabeled crRNA (CRISPR locus 5, spacer 6) was incubated without $MgCl_2$ and in the absence (C) or presence of 15 µM, 30 µM and 60 µM of MkCsm3 and RNase T1. The reactions were separated on a 12% denaturing gel. Arrows indicate prominent bands that result from RNase T1 digest and do not show protection by MkCsm3.

(C) Radiolabeled crRNA (CRISPR locus 5, spacer 6) was incubated in buffer with 4 mM $MgCl_2$ and in the absence (C) or presence of 3 µM, 16 µM or 24 µM of MkCsm3 and RNase T1. The reactions were separated on a 12% denaturing gel. For size determination a decade marker (Life Technologies) (DM) and alkaline hydrolysed crRNA (AH) after 2 minutes and 5 minutes incubation time were used. Arrows indicate additional products that derive from a potential MkCsm3 processing activity.

Analysis of a potential MkCsm3 nuclease activity

The primary processing of crRNAs by Cas6 is followed by a secondary processing event in CRISPR-Cas type III systems. In *Staphylococcus epidermidis* this results in two mature crRNA species with a difference in length of six nucleotides. The deletion of the genes *csm2*, *csm5* or *csm3* resulted in the absence of the two mature crRNA species and indicated that these gene products are involved in the secondary processing (Hatoum-Aslan *et al*., 2011). For MkCsm3 a potential Mg^{2+} dependent crRNA processing activity was detected during our footprinting analysis (Fig. IV 1B), indicating that this protein might be the RNase responsible for the secondary processing event.

To test if MkCsm3 has a crRNA processing activity, we analyzed the degradation of an *in vitro* transcribed crRNA substrate in the absence and presence of MkCsm3 and $MgCl_2$. The crRNA 2.1 (CRISPR locus 2, spacer 1) was chosen as RNAseq analyses in *M. kandleri* revealed sharply processed 3'-ends of this crRNA and indicated an especially specific processing with this substrate (Su et al., 2013). EMSAs revealed nine cleavage products in the presence of $MgCl_2$ and MkCsm3 that were cleaved non-specifically at one cytosine and several guanine or uracil bases. A cleavage pattern in six nucleotide steps, like it has been shown in *S. epidermidis* (Hatoum-Aslan *et al*., 2013), could not be detected in our *in vitro* assays. Furthermore, the high amount of MkCsm3 required to detect cleavage products indicates a low activity of the protein. To test the influence of other divalent metal ions on the activity, RNA degradation was analyzed in the presence of $MgCl_2$, $MnCl_2$, $CuCl_2$, $NiCl_2$, $CoCl_2$, $ZnCl_2$ and $CaCl_2$. Cleavage products were only detected in the presence of $MgCl_2$ and $MnCl_2$. However, in the MkCsm3 crystal structure, a Zn^{2+} ion is coordinated and Mg^{2+} or Mn^{2+} ions were not detected (Hrle *et al*., 2013).

Taken together, the non-specifc processing, the low activity, and the dependence on Mg^{2+} or Mn^{2+} instead of Zn^{2+} ions hint that the observed cleavage products might result from a contaminating *E. coli* RNase that has been co-purified with MkCsm3. To rule out this possibility the cleavage activity of MkCsm3 mutants was tested. Cleavage of *in vitro* transcribed crRNA was detected with all available mutants (Fig. IV 2C). The strongest activity was detected with the mutant H199A (Mut2) that is mutated in the unstructured loop, however the reason for this still remains elusive. Even with mutants that showed a decreased binding activity (Mut1: Δloop; Mut3: R92E/K167E and Mut4: D176N) cleavage products were detected (Fig. IV 1A). This indicates that the RNA binding activity does not correlate with the observed cleavage activity and the effect of a potential contaminating RNase can still not be ruled out.

Recently, a complete Csm complex consisting of Csm1-Csm5 was expressed and purified from *S. epidermidis*. Even within this complex a crRNA processing could not be detected and the protein responsible for the maturation of crRNAs still remains elusive. However, EMSAs of the *S. epidermids* Csm3 protein revealed a binding of crRNAs in a 6 nucleotide steps and indicate that

Csm3 acts as a ruler protein that measures the extent of 3'-end processing (Hatoum-Aslan *et al.*, 2013). A possible scenario is that multiple copies of Csm3 bind the crRNA in the Csm complex and protect the crRNA in 6 nucleotide steps from the 5'-end on. Subsequently, excess parts of the crRNA (due to variable sizes of spacer sequences) are then accessible for degradation by an unknown RNase or chemical degradation (Rouillion *et al.*, 2013). This is in agreement with RNAseq analyses that confirmed defined 5'-ends but gradually processed 3'-ends of crRNAs in *M. kandleri* (Su *et al.*, 2013).

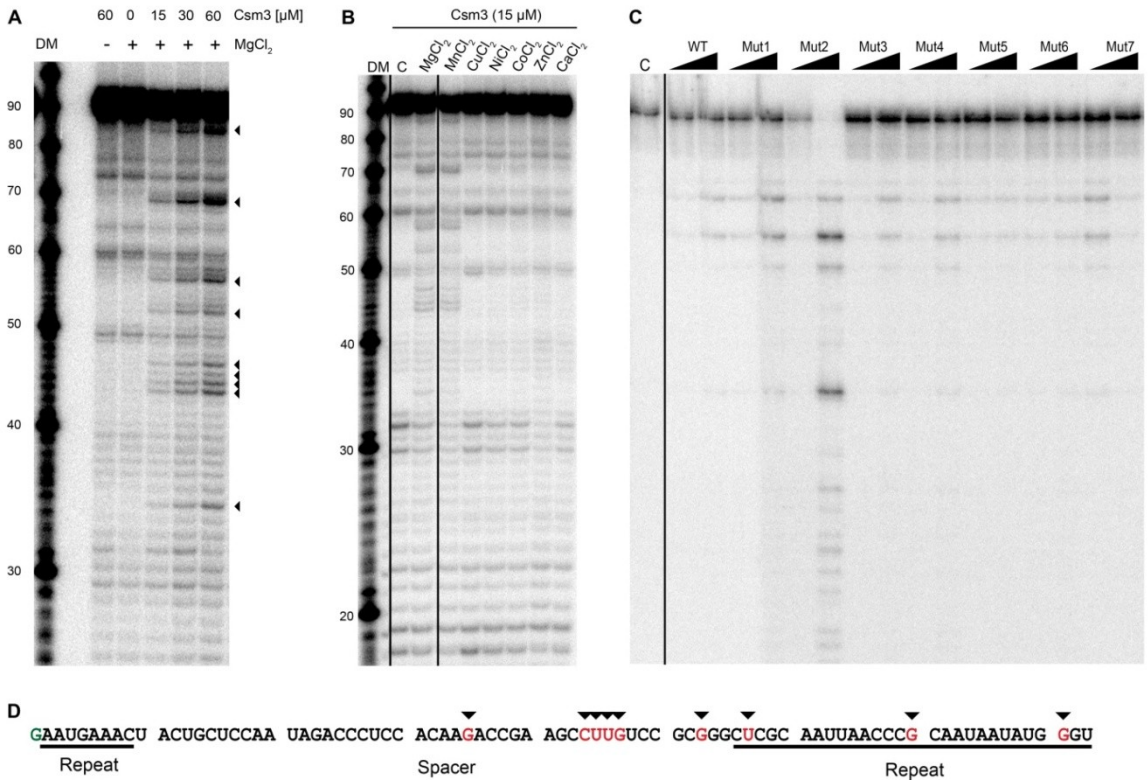

Figure IV 2. Potential RNA processing activity of MkCsm3
(A) Radiolabeled crRNA (CRISPR locus 5, spacer 6) was incubated in the absence or presence of 4 µM MgCl2 with the indicated amount of MkCsm3 and analyzed on a 12% denaturing gel. Potential cleavage products are indicated by arrows and the Decade Marker (Life Technologies) (DM) was used for size determination.
(B) Radiolabeled crRNA (CRISPR locus 5, spacer 6) was incubated in the absence (C) or presence of 4 µM MgCl2, MnCl2, CuCl2, NiCl2, CoCl2, ZnCl2 or CaCl2 on a 12% denaturing gel. The Decade Marker (Life Technologies) (DM) was used for size determination.
(C) Radiolabeled crRNA (CRISPR locus 2, spacer1) was incubated in the absence (C) or presence of wild type (WT) or mutant MkCsm3 (Mut1: Δloop; Mut2: H199A; Mut3: R92E/K167E; Mut4: D176N; Mut5: S52/53E;Mut6: P50A/S52E/S53E; Mut7: R57L) in concentrations of 2.5 µM and 10 µM and loaded on a 12% denaturing gel.
(D) The sequence of the used crRNA substrate (CRISPR locus 5, spacer 6) is shown. Spacer and repeat sequences are indicated and potential cleavage sites are marked by arrows. The first G residue (marked in green) does not belong to the endogenous sequence and is the result of the T7 RNA polymerase promoter used for in vitro transcription.

Cmr4 expression and crystallization attempts

The genome of *M. kandleri* harbors a second CRISPR-Cas system of the type III-B that is composed of the genes *cmr1-cmr6*. These genes encode for a type III-B interference complex (Cmr complex) in which the protein Cmr4 has been shown to function as the backbone subunit, analogous to Csm3 for the type III-A system (Spilman *et al*., 2013; Staals *et al*., 2013; Zhang *et al*., 2013). Due to its functional and sequence similarity to Csm3 proteins and the general high rigidity of thermophilic proteins, MkCmr4 is a promising protein for crystallization. Therefore, MkCmr4 was produced in *E. coli* and purified via Ni-NTA and gel filtration chromatography. The gelfiltration chromatogram revealed that a large fraction of the protein is present in a high oligomeric state as well as in a dimeric state and only a small fraction is available as a monomer (Fig. IV 3). A high oligomeric state seems to be a common feature of heterologously expressed Cas7 proteins that might be formed due to sequence unspecific, helical binding of contaminating *E. coli* RNAs (Plagens *et al*., 2014). However, nucleic acid binding of MkCmr4 was tested on an ethidium bromide stained agarose gel and revealed no detectable co-purified nucleic acids. Finally, the dimer fraction of MkCmr4 was sent to the collaborating laboratory of Prof. Dr. Elena Conti in Martinsried, Germany for crystallization. Once MkCmr4 is crystallized, its structure will be compared with MkCsm3 to analyze structural differences between CRISPR-Cas type III-A and III-B system backbone proteins.

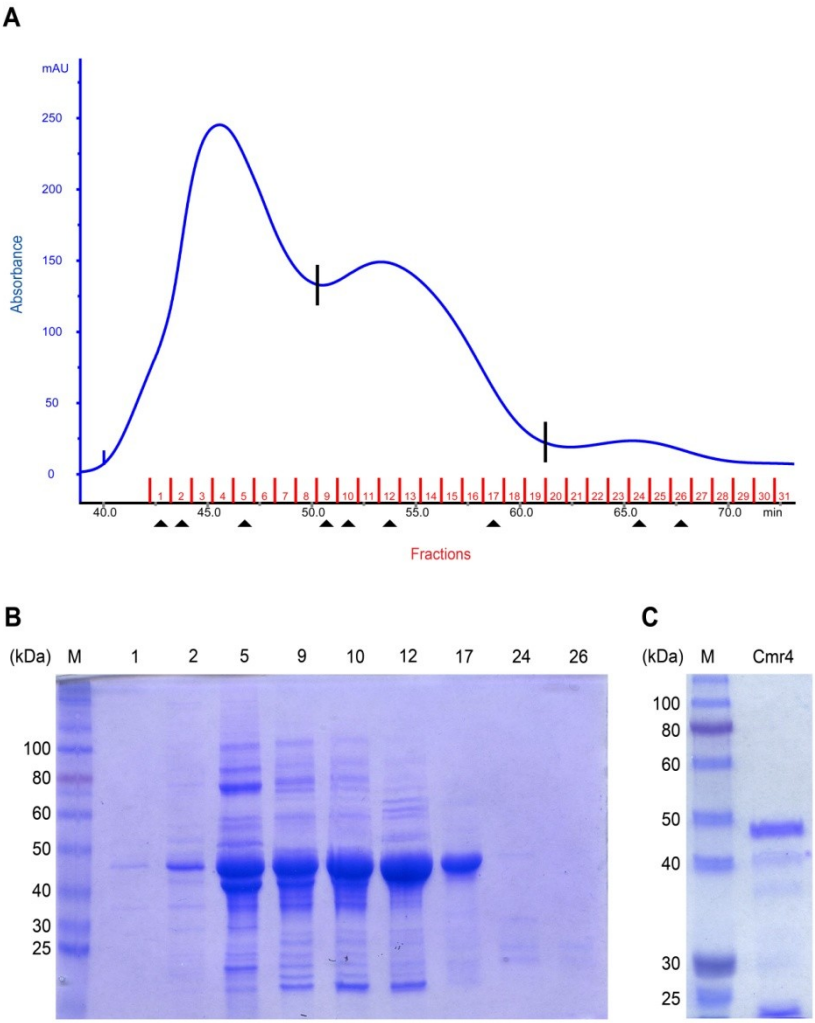

Figure IV 3. MkCmr4 purification

(A) The gel filtration chromatogram of the MkCmr4 purification is shown. The dimer fractions (bordered by two black lines) were pooled and used for crystallization. Fractions that were analyzed in (B) are indicated by arrows.

(B) The indicated fractions of the MkCmr4 gel filtration purification were analyzed by 12% SDS-PAGE with the ColorPlus Prestained Protein Ladder (NEB) (M).

(C) 0,5 µl of 10 mg/ml concentrated MkCmr4 that was sent for crystallization was analyzed on a 12% SDS-PAGE with the ColorPlus Prestained Protein Ladder (NEB) (M).

Chapter V

Analysis of a prokaryotic Argonaute encoded in a CRISPR-Cas gene cluster in *Methanopyrus kandleri*

Andreas A. H. Su[1] and Lennart Randau[1]

[1]Prokaryotic Small RNA Biology Group; Max Planck Institute for Terrestrial Microbiology; Karl-von-Frisch-Straße 10, Marburg, Germany

Unpublished manuscript

Abstract

Argonaute (Ago) proteins are universally distributed in all three domains of life. In eukaryotes, Ago proteins are the core components of small RNA-guided silencing mechanisms that involve the regulation of gene expression like RNA interference (RNAi) or the repression of foreign genetic elements. Recently, a defense function against foreign plasmid DNA has been discovered for prokaryotic Argonaute (pAgo) proteins.

Here, we analyzed the role of the pAgo protein of *Methanopyrus kandleri* (MKAgo) that is encoded in an operon together with genes of the clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated (Cas) immune system. The alignment of different pAgo sequences revealed potential active site residues in MkAgo that might be important for protein stability. Furthermore, cloning experiments indicate a toxic effect of heterologously expressed MkAgo in *E. coli* that might correlate with its putative function of repressing foreign plasmid DNA.

Introduction

Prokaryotes and eukaryotes have developed elaborate defense mechanisms that use small RNAs (sRNA) for specific detection and silencing of foreign nucleic acids. In prokaryotes, various systems based on clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins have been identified. Analogous to these systems, eukaryotes have evolved RNA interference (RNAi) and related sRNA guided gene silencing mechanisms (Ketting *et al*., 2011).

The key components of all known sRNA-guided gene silencing mechanisms in eukaryotes are Argonaute (Ago) proteins. Characteristically, Ago proteins interact with small guide RNAs like short interfering RNAs (siRNAs), micro RNAs (miRNAs) or PIWI interacting RNAs (piRNAs) to induce post-transcriptional gene silencing (Meister, 2013; Hutvagner and Simard, 2008). The canonical pathways of sRNA-guided gene silencing mechanisms involve double stranded siRNA or miRNA that are derived from exogenous or endogenous sources. These sRNA classes are processed by the RNase III enzyme Dicer into duplexes of 21-25 nucleotides length (Jinek and Doudna 2009). Subsequently, one of the sRNA strands is loaded into Ago proteins that, often together with other proteins, build the RNA induced silencing complex (RISC) (Rivas *et al*., 2005). RISC uses bound sRNAs as guide sequences to target mRNAs that either can be degraded by a slicer activity of the Ago protein or translationally silenced (Jinek and Doudna 2009). Other sRNA-guided gene silencing mechanisms use the interaction of Ago with siRNAs to target heterochromatin and induce epigenetic silencing or with piRNAs to cleave transposable elements (Verdel *et al*., 2004; Siomi *et al*., 2011).

The architecture of Ago proteins exhibits two lobes composed of four domains. The N-terminal and PIWI Argonaute Zwille (PAZ) domains form one lobe and the middle (MID) and PIWI domains form the second lobe (Song *et al*., 2004; Rashid *et al*., 2007; Elkayam *et al*., 2012). The PAZ domain is involved in the binding of the 3'-end of the guide RNA (Jinek and Doudna, 2009). The MID domain builds a tight binding pocket that extensively interacts with the 5'-end terminal phosphate of the guide RNA (Rivas *et al*., 2005; Elkayam *et al*., 2012). Ago proteins contain a PIWI domain that shows structural homology to an RNase H fold and contains an RNA cleavage activity (Liu *et al*., 2004; Rivas *et al*., 2005). The catalytic center is formed by a DEDX (X is generally an Asp or His) tetrad that coordinates two Mg^{2+} ions (Nakanishi *et al*., 2012; Sheng *et al*., 2014). However, not all Ago proteins contain a cleavage activity. For example, in mammals only AGO2 is the catalytically active 'slicer' amongst four different Ago family members (Liu *et al*., 2004, Meister *et al*., 2013).

Interestingly, Ago proteins have also been identified in bacteria and archaea but in contrast to their eukaryotic counterparts, the physiological functions of prokaryotic Argonaute (pAgo)

proteins are only poorly understood. Recent studies in *Rhodobacter sphaeroides* and *Thermus thermophilus* found pAgo proteins associated with short DNA or RNA molecules. Furthermore, in both organisms pAgo proteins have been shown to function against plasmid DNA by decreasing the plasmid yield or plasmid transformation efficiency (Olovnikov *et al*., 2013; Swarts *et al*., 2014). This indicates that in contrast to eukaryotic Ago proteins that target RNA, pAgo proteins seem to have a direct effect on DNA (Hur *et al*., 2014). However, there are also striking differences between the two analyzed pAgo proteins. While the *R. sphaeroides* pAgo (RsAgo) copurifies with RNA-DNA duplexes, the *T. thermophilus* pAgo (TtAgo) interacts with DNA-DNA duplexes (Olovnikov *et al*., 2013; Swarts *et al*., 2014). The DNA dependent DNA cleavage activity of pAgo proteins is further supported by biochemical analysis of the *Methanocaldococcus jannaschii* pAgo that revealed an exclusive cleavage of DNA targets out of a DNA/DNA duplex (Zander *et al*., 2014). In contrast to that, in RsAgo the catalytic DEDX tetrad required for a slicer function is not conserved and no cleavage activity was observed *in vitro*. Accordingly, a cleavage mechanism was proposed where RNA loaded RsAgo mediates DNA cleavage by recruitment of an unknown nuclease (Olovnikov *et al*., 2013).

A unique genetic context is found for the pAgo from the hyperthermophilic archaeon *Methanopyrus kandleri* (MkAgo). In this organism, the gene encoding MkAgo is located in an operon coding for the Cas proteins Cas1 and Cas2 (Makarova *et al*., 2006). These two genes are essential parts of the prokaryotic adaptive immune system CRISPR-Cas and are involved in the integration of foreign DNA into CRISPR clusters (Nunez *et al*., 2014). Therefore, MkAgo is the only known putative link between the CRISPR-Cas systems and the eukaryotic RNAi mechanism.

Here, we analyzed the function of the MkAgo protein in *M. kandleri*. We aimed to express MkAgo heterologously in *E. coli* to analyze potential co-purified protein and nucleic acid binding partners via RNAseq, DNAseq and mass spectrometry (MS). However, cloning of the MkAgo coding sequences into constructs for protein expression in *E. coli* exclusively resulted in mutated gene versions. We found that only the cloning of mutated or fused gene versions was possible but resulted in insoluble expressed proteins. We speculate that the high rate of mutations indicates a high toxicity of MkAgo in *E. coil* and might be the result of MkAgo silencing its own expression plasmid. Finally, we identified a potential active site in MkAgo that seems to be required for protein stability.

Material and Methods

Microbial strains and growth conditions

E. coli strains Top10 (Life Technologies), Rosetta 2 pLysS (Stratagene) and Arctic Express (Agilent) were grown in LB medium at 37°C shaking at 200 rpm if not stated otherwise. *M. kandleri AV19* cells were a kind gift of D. Söll, Yale University, USA and were grown in the Archaeenzentrum Regensburg, Germany (H. Huber, M. Thomm, K. Stetter) in a 300 l fermenter as described (Slesarev *et al*., 2002).

Preparation of genomic M. kandleri DNA

Genomic DNA from *M. kandleri* was prepared from 0.5 g of cells resuspended in 5 ml buffer (50 mM TrisHCl pH 7, 300 mM NaCl and 10 mM Beta-Mercaptoethanol). Cells were lysed with mortar and pestle, followed by sonication. After centrifugation (13 000x g, 10 min., 4°C), the supernatant was phenol chloroform extracted (pH 7.5), EtOH precipitated and resuspended in 200 µl ddH_2O.

Cloning of MkAgo into E. coli expression constructs

The MkAgo coding gene (Mk1311) was PCR amplified from *M. kandleri* genomic DNA with primers containing the required restriction sites and cloned into various expression vectors (Tab. V 2). The vector pEC-A-Hi-SUMO was a kind gift from Elena Conti, Martinsried, Germany. Cloning of the MkAgo coding gene into pEC-A-Hi-SUMO via the ligation independent cloning (LIC) technique was performed as described (Scholz *et al*., 2013). Cloning of MkAgo into pTrcHis TOPO (Life Technologies) was performed according to the the manufacturer's instructions and sequenced using the provided primers Xpress forward and pTrcHis reverse. Other cloned constructs were analyzed by sequencing using custom made primers (MWG Eurofins).

Expression of MkAgo mutants in E. coli

A construct encoding the codon optimized, C-terminally 6xHis-tagged MkAgo D495N mutant protein was purchased from by GenScript, USA. The gene was obtained in the vector pUC57 and sub-cloned into pET20b via NdeI and XhoI. A vector encoding the double mutant MkAgo D495N D563N was created by using the QuikChange II Site-Directed Mutagenesis Kit (Strategene) according to the manufacturer's instructions. Protein production was induced with 0.5 mM IPTG for 1h at 37°C in Rosetta2 pLysS (Strategene) if not stated otherwise. After induction of protein expression, the cells were disrupted by sonication in a buffer containing 50 mM TrisHCl pH 7, 300 mM NaCl, 10 mM Beta-Mercaptoethanol and 4 mM $MgCl_2$ and cleared by centrifugation (45 000x g, 30 min., 4°C). Soluble expressed proteins were purified out of the cleared supernatant by immobilized metal ion affinity chromatography using Ni-NTA agarose (Quiagen). The presence of soluble protein was determined by SDS PAGE (12%).

Results and Discussion

Cloning of MkAgo coding sequences into E. coli expression vectors induces mutations and indicates a potential toxicity of MkAgo

The pAgo protein of *M. kandleri* is encoded in a potential operon with two genes coding for the Cas proteins Cas1 and Cas2 (Fig. V 1). This provides the only known putative link between the CRSPR-Cas systems and the eukaryotic RNAi mechanism and raises the question of the physiological function of MkAgo (Makarova *et al*., 2006). Due to the lack of genetic tools in *M. kandleri* the analysis of the MkAgo function is restricted to biochemical approaches. Therefore, we aimed to use heterologously expressed MkAgo for the co-purification of potential protein and nucleic acid binding partners present in a *M. kandleri* cell extract. Subsequently, we planned to identify the binding partners by MS or DNA/RNAseq to draw conclusions about the physiological function of MkAgo.

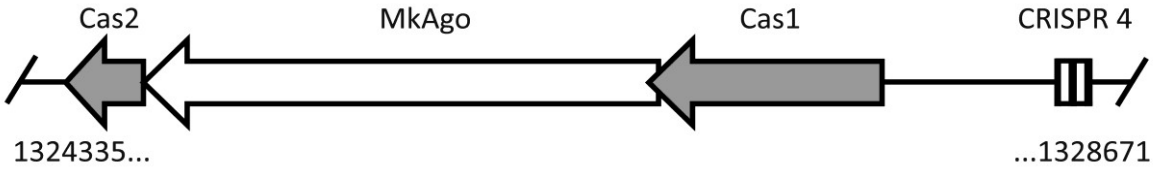

Figure V 1. Location of the gene encoding MkAgo within an operon encoding Cas genes
Depicted are the genes encoding the CRISPR associated proteins Cas1 and Cas2 interspaced by the gene encoding MkAgo. The operon is in close proximity to the CRISPR locus 4 that contains of three repeat sequences (black bars) and two spacer sequences (white bars).

To faciliate the expression of 6xHis-tagged MkAgo in *E. coli*, its coding sequence was cloned into the pTrcHis TOPO expression vector. Positive clones were sequenced and revealed that each construct contains at least one point mutation in the pAgo coding gene that alters the amino acid sequence (Tab. V 1). In most cases low point accepted mutation (PAM) matrix scores indicate that the mutations can have drastic effects on the protein structure or function. A possible explanation for the induction of mutations during cloning can be a high toxicity of the gene product (Giacalone *et al*., 2006). Therefore, we conclude that low amounts of MKAgo resulting from leaky expression are toxic to *E. coli* cells and in response to that the detected mutations were induced to affect MkAgo structure and function.

To overcome problems with potential toxic effects of MkAgo, we tested vectors encoding different fusion proteins like the maltose-binding protein (MBP), the N-utilization substance protein A (NusA), the small ubiquitin-related modifier (SUMO) or cystein protease domain (CPD) (Marblestone *et al*., 2006; Malakhov *et al*. 2004; Shen *et al*., 2009;). The constructs could either not be cloned without yielding mutations in the gene or expression resulted in insoluble or non-specifically degraded protein in *E. coli* (Tab. V 2). The observation that only constructs coding for

Table V 1. Detected mutations during cloning of MkAgo coding sequences into the pTrcHis TOPO expression vector

Clone #	Position	Mutation	Changed aminoacid	PAM1 score
3	608	CTC -> CCC	Leu -> Pro	2
10	1594	CCC -> TCC	Pro -> Ser	17
11	884	CTC -> CCC	Leu -> Pro	2
13	416	GGT -> GAT	Gly -> Asp	6
	493	CCC -> TCC	Pro -> Ser	17
	556	TGG -> CGG	Trp -> Arg	8
	599	GAC -> GGC	Asp -> Gly	11
	1573	GTG -> GTA	Val -> Val	9901
30*	422	GTC -> GCC	Val -> Ala	18
	590	GTC -> GCC	Val -> Ala	18
	835	GGG -> AGG	Gly -> Arg	0
31*	389	GTC -> GCC	Val -> Ala	18
32	1502	GAG -> GGG	Glu -> Gly	7
	2077	AGG -> GGG	Arg -> Gly	1
33*	73	CGA -> TGA	Arg -> Stop	/
	423	GTC -> GTT	Val -> Val	9901
38	1286	GAC -> GGC	Asp -> Gly	11
39*	389	GTC -> GCC	Val -> Ala	18

*only sequenced with the Xpress Forward primer (first ~1000 bp sequenced)

Table V 2. Overview of tested MkAgo expression constructs

Vector	Cloning sites	Tag	Mutations	Produced protein
pTrcHis TOPO	TOPO cloning	N-6xHis N-express epitope N-Ek	Yes (Tab. V 1)	/
pET-20b	EcoRI HindIII	N-pelB C-6xHis	Yes	/
pET-21a CPD	NdeI StuI	C-CPD C-6xHis	Yes	/
pETM-43	EcoRI HindIII	N-MBP C-6xHis	Yes	/
pET-43.1a	EcoRI HindIII	N-NusA C-8xHis	Yes	/
pET-43.1a	SpeI EcoRI	N-NusA N-6xHis N-S tag N-Th N-Ek C-8xHis	No	Degraded in *E. coli*
pEC-A-Hi-SUMO	LIC	N-6xHis N-SUMO	No	Insoluble

Ek, enterokinase cleavage site; Th, Thrombin cleavage site; LIC, ligation independant cloning

insoluble or cleaved proteins can be cloned without detecting mutations in the gene provides another line of evidence for a toxic effect of MkAgo in *E. coli*. However, this is an unexpected strong effect for a hyperthermophilic protein that is *in vivo* active at temperatures over 84°C.

Our findings are in line with the observation of reduced plasmid yield and transformation efficiencies that have been observed for vectors expressing *T. thermophilus* and *R. sphaeroides* pAgo (TtAgo and RsAgo) proteins (Swarts *et al.*, 2014; Olovnikov *et al.*, 2013). In both cases pAgo proteins have been shown to target their own expression vector and a similar mechanism could be the reason for the toxicity of MkAgo in *E. coli*.

MkAgo contains a potential active site that is essential for proper protein folding

For the high toxicity of MkAgo in *E. coli*, a potential nuclease activity might be the reason. Therefore, we analyzed similarities between MkAgo and the well studied TtAgo protein. A crystal structure of TtAgo reveals that its nuclease activity is conferred by an RNase H like DEDD tetrad that coordinates two catalytically active Mg^{2+} ions (Sheng *et al.*, 2013) (Fig. V 2). Furthermore, structural studies of other RNase H enzymes showed that single exchanges of Asp to Asn residues in the catalytic tetrad abolish the nuclease activity (Nowotny *et al.*, 2005). To overcome cloning problems of MkAgo coding sequences in *E. coli* we planned to attempt the production of a mutated MkAgo protein that exhibits a reduced nuclease activity while retaining its ability to bind potential nucleic acid or protein interaction partners. In *R. sphaeroides* it has been shown that even the potentially nuclease inactive RsAgo, lacking the catalytic DEDD tetrad, is still able to interact with nucleic acids and to repress expression of plasmid DNA (Olovnikov *et al.*, 2013).

To detect the potential active site in MkAgo, we performed sequence alignments with two homologous pAgo sequences. We chose the TtAgo sequence as a detailed crystal structure with an identified catalytic pocket is available (Sheng *et al.*, 2014). The homologous sequence of *A.*

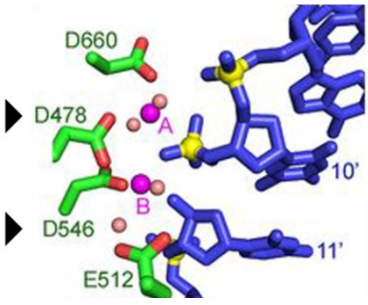

Figure V 2. Detailed view on the catalytic pocket of the TtAgo protein in a ternary complex with a guide and target DNA molecule

The catalytic DEDD tetrad residues (green) of TtAgo are in close proximity to the cleaved target DNA molecule (blue ribbon). The active site residues coordinate two Mg2+ ions (magenta balls A and B) with the help of bridging water molecules (pink balls). One of the water molecules is positioned for an in-line nucleophilic attack on the cleavable phosphate. The two central Asp478 and Asp546 residues (indicated by arrows) correspond to Asp495 and Asp563 that have been identified in MkAgo in this study (Figure adapted from Sheng et al., 2014).

aeolicus Ago was used as an additional control. The alignment identified the residues Asp495 and Asp563 in MkAgo that correspond to Asp478 and Asp546 residues of the DEDD catalytic tetrad in TtAgo (Fig. V 3). Both residues have been shown in the TtAgo crystal structure to interact with both catalytic Mg^{2+} ions in the active site (Fig. V 2). To inactivate MkAgo, the Asp495 was chosen to be substituted by an Asn (MkAgo D495N mutant) due to its central position in the catalytic pocket and its direct involvement in the coordination of both Mg^{2+} ions according to the TtAgo structure. A codon optimized MkAgo D495N mutant coding gene was synthesized (GenScript, USA) and cloned into a pET20b expression vector. The expression of the MkAgo D495N mutant resulted in insoluble protein with an insignificant fraction of soluble protein that was susceptible to precipitation within a few days (Fig. V 4). A larger scale eight liter expression did not yield enough soluble protein for subsequent experiments. Further approaches to increase the ratio of soluble to insoluble proteins like the use of designed *E. coli* strains for the expression of insoluble proteins (Arctic Express, Agilent Technologies), the reduction of expression temperature to 12°C, 18°C and 30°C or the reduction of IPTG concentration down to 1 µM were not successful. The mutation of the second active site residue Asp563 by site directed mutagenesis created the double mutant D495N D563N that still resulted in insoluble protein. However, we were not successful in reverting the D495N back to D495 to test the single mutant D563N for solubility.

We conclude that the positions Asp495 and Asp563 are active site residues of a DEDD active site in MkAgo and that Asp495 is a crucial residue for the stability and function of MkAgo. Furthermore, the detected low amount of soluble MkAgo D495N provides weak evidence that the mutant protein might not be active anymore in *E. coli*. Taken together, our results indicate that the single expression of MkAgo in *E. coli* is not a suitable strategy to obtain sufficient amounts of protein. Potential approaches to express MkAgo in the future could be an expression of a codon optimized MkAgo in the archaeon *Sulfolobus acidocaldarius*. An established expression system is available for this organism (Berkner *et al*., 2010) and the heterologous expression in an organism

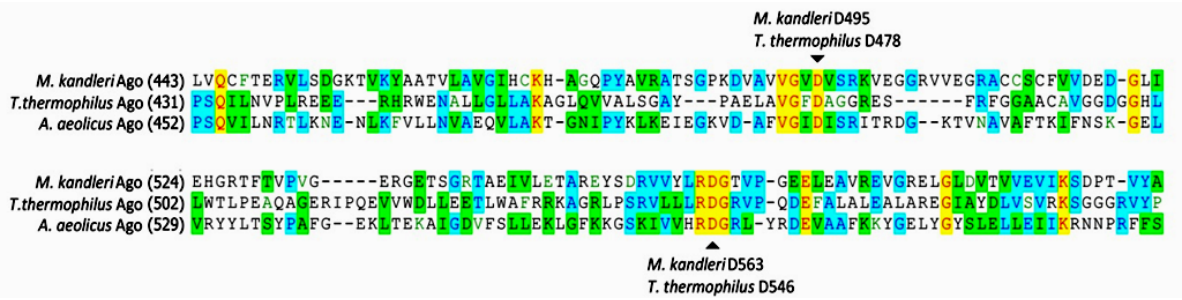

Figure V 3. Identified Asp495 and Asp563 residues of the potential active site in MkAgo
The alignment of pAgo sequences identified the residues Asp495 and Asp563 in MkAgo as residues that correspond to the residues Asp478 and Asp546 of TtAgo (indicated by arrows). Both residues have been shown in TtAgo to be parts of the catalytic RNAse H like DEDD tetrade. Identical residues (yellow), conserved residues (blue) and similar residues (green) are highlighted.

from the same domain of life could be an advantage. This has been shown for the human Argonaute2 that was not expressable in *E. coli* but an expression in insect cells yielded enough protein for subsequent crystallization (Elkayam *et al.*, 2012). Another approach might be a co-expression of MkAgo with Cas1 or Cas2 in *E. coli*. As MkAgo is encoded together with Cas1 and Cas2 in a potential operon, these proteins are potential interaction partners of MkAgo and could regulate the MkAgo function or even represent a toxin-antitoxin system (Yamaguchi *et al.*, 2011). However, extensive trials to clone Cas1 and Cas2 coding genes into a pRSF Duet co-expression vector did not yield any transformands but might be extended in the future to other co-expression vector systems.

An open question for the function of pAgo proteins in general is how guide DNA (or RNA) molecules are generated and acquired. In *M. kandleri* Cas1 and Cas2 could be potential candidates for this task as they are involved in the acquirement of foreign DNA sequences into the genome (Nunez *et al.*, 2014). A potential scenario could be that Cas1, Cas2 or a complex of both recognize and cleave foreign DNA molecules and provide small DNA molecules for the integration into CRISPR loci and as guide sequences for a pAgo mediated immune response. It will be interesting to analyze if such a mechanism exists in *M. kandleri* and which kind of mechanisms might have evolved in other prokaryotes to produce guide molecules for pAgo proteins.

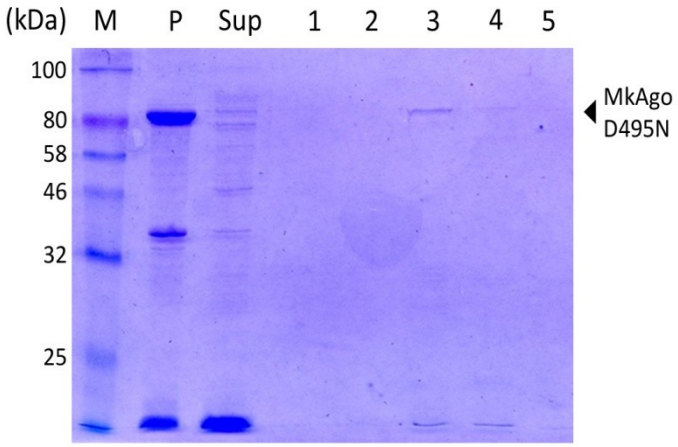

Figure V 4. Purification of MkAgo D495N mutant protein
The 6xHis-tagged MkAgoD495N protein was purified via Ni-NTA agarose affinity chromatography. Pellet (P), supernatant (Sup), and elution fractions (1-5) were analyzed together with the Color Protein standard (NEB) (M) by 12% SDS PAGE.

Chapter VI

Conclusions

Small RNA (sRNA) molecules play important roles in various cellular processes and can guide proteins to their target nucleic acids. Processes that use small guide RNAs include defense mechanisms against foreign nucleic acids like the eukaryotic RNA interference (RNAi) or prokaryotic CRISPR-Cas immune systems (Meister 2013; Van der Oost *et al.*, 2014). Recently, even prokaryotic Argonaute (pAgo) proteins have been shown to be involved in DNA- or RNA-guided defense against foreign plasmid DNA (Olovnikov *et al.*, 2013; Zander *et al.*, 2014; Swarts *et al.*, 2014). Furthermore, C/D box and H/ACA sRNAs are utilized to guide 2'-*O*-methylation and pseudouridylation modifications of target RNA molecules in Archaea and Eukarya (Yip *et al.*, 2013). These sRNA-guided mechanisms have been detected and are highly active in thermophilic archaea.

The archaeal methanogen *Methanopyrus kandleri* lives at extreme temperatures of 84-110°C and exhibits a unique C to U editing of tRNAs as well as a pAgo protein encoded in a CRISPR-Cas system. In this thesis, a combination of RNAseq analyses, bioinformatical and biochemical approaches was applied in *M. kandleri* to analyze adaptations of sRNA-guided mechanisms and the maturation of sRNA molecules at extreme temperatures.

Small RNA profiling in the hyperthermophilic Methanopyrus kandleri

In this thesis, RNAseq methodology, combined with bioinformatical methods, was applied to analyze the sRNome of the hyperthermophilic organism *M. kandleri* and to obtain insights into the RNA metabolism at extreme temperatures.

The unique C-to-U editing mechanism in *M. kandleri* might be a strategy to prevent viral integration into tRNA genes or to coordinate tRNA maturation at extreme temperatures (Randau *et al*., 2009). Here, we exploited this C-to-U editing as a marker to deduce the order of tRNA maturation events in an archaeal organism via sequencing of different tRNA precursor molecules. The observed order of 3'- and 5'-end processing preceding splicing resembles the order found in eukaryotes and indicates the conservation of the maturation processes during evolution (Yoshihisa *et al*., 2003). The RNAseq analysis of crRNAs revealed their constitutive production and processing in *M. kandleri* which is as a good indicator for the high activity of CRISPR-Cas systems at extreme temperatures. A polynucleotide kinase treatment was applied to modify the 5'- and 3'-ends of RNAs which allows an efficient sequencing of mature crRNAs. This method can be applied to various other bacterial or archaeal organisms to specifically enrich RNAseq libraries for crRNAs or globally map the 5' phosphorylation status of RNA molecules (Juranek *et al*., 2012; Richter *et al*., 2012, Zoephel *et al*., 2013, Plagens *et al*., 2014). Furthermore, the RNAseq analyses revealed a record number of C/D box sRNAs as well as the circularization of these molecules. Both aspects are suggested to represent adaptations to the hyperthermophilic lifestyle of *M. kandleri*. In this thesis, circular RNAs were detected as permuted reads that were mapped to the genome in a chiastic manner. A technique to enrich RNAseq libraries for circular transcripts is the use of RNaseR that exclusively degrades linear RNA, while sparing circular transcripts (Danan *et al*., 2011). However, in our analyses the detection of circular RNAs was possible without this treatment due to a high number of randomly linearized C/D box sRNAs.

In the future, the obtained insights into the sRNome of a hyperthermophilic archaeon presented in this thesis, especially the biogenesis and circularization mechanisms of C/D box sRNAs as well as their methylation sites on rRNAs will be further analyzed. Therefore, RNAseq, biochemical and genetic approaches will be applied to the hyperthermophilic archaeon *Sulfolobus acidocaldarius* which provides the advantage of having an established genetic tool set (Wagner *et al*., 2009; Berkner *et al*., 2010). The C/D box sRNAs identified in this study provide a large pool of guide sequences that will be used, in combination with data of other archaeal organisms, to globally map methylation sites of rRNAs *in silico.* In conclusion, the analyses of the *M. kandleri* sRNome provide first insights into the RNA metabolism at extreme temperatures. In the future it will be interesting to analyze how exactly the mechanisms of RNA methylation contribute to withstand extreme temperatures and if additional targets and functions of C/D box sRNAs will be revealed.

Analysis of a prokaryotic Argonaute in a CRISPR-Cas system

Argonaute (Ago) proteins are the core components of the eukaryotic RNA interference (RNAi) mechanism and utilize guide RNAs to identify their target nucleic acids. Recently, the analysis of prokaryotic Ago (pAgo) proteins revealed an interaction with guide molecules (DNA or RNA) to defend the organism against foreign plasmid DNA (Olovnikov *et al*., 2013; Zander *et al*., 2014; Swarts *et al*., 2014). In *M. kandleri*, a pAgo protein is encoded in a potential operon with CRISPR-associated (*cas*) genes. This unique location of a pAgo coding gene provides a putative link between the CRSPR-Cas system and the eukaryotic RNAi mechanism (Makarova *et al*., 2006).

In this thesis, the potential function of the *M. kandleri* Argonaute (MkAgo) protein was analyzed. Extensive efforts to integrate pAgo coding sequences into various expression vectors revealed a high toxicity of pAgo in *E. coli*. This could indicate that MkAgo interferes with its own expression plasmid in *E. coil* and would be in line with reported defense mechanism against plasmid DNA (Olovnikov *et al*., 2013; Swarts *et al*., 2014).

Future research could focus on approaches for a successful production of the MkAgo protein in insect cell- or *Sulfolobus acidocaldarius* expression systems to analyze the function of this protein. Additionally, we analyzed components of the *M. kandleri* CRISPR-Cas system to improve our understanding of the function and protein composition of CRISPR interference complexes and a putative involvement of the MkAgo protein.

Analysis of the CRISPR-Cas type III-A backbone protein

In this thesis, the protein Csm3 of the type III-A CRISPR-Cas system was analyzed in *M. kandleri* to obtain insights into the function of this protein within the interference complex. At the beginning of this project, only little was known about the function, structure and composition of type III-A interference complexes and no crystal structures were available. Recombinant Csm3 was produced in *E. coli* and crystallized in collaboration with Prof. Elena Conti (MPI Martinsried). The crystal structure of the Csm3 protein represents the first structure of a type III-A system protein and, together with biochemical analyses, revealed that Csm3 proteins are crRNA-binding proteins. Consistent with our studies, electron microscopy analyses in *Sulfolobus solfataricus* revealed that six Csm3-like proteins serve as a helical crRNA-binding backbone of the type III-A interference complex (Rouillon *et al*., 2013).

In general, studies of protein complexes in *M. kandleri* are restricted to biochemical approaches due to the lack of genetic tools for this organism. To elucidate the function and structure of the entire interference complex, recombinant Csm3 can be used as bait proteins to co-purify complexes from *M. kandleri* cell lysates. Purified complexes can then be used to analyze the composition of proteins, e.g. by mass spectrometry and electron microscopy.

Appendix

References

Adams, P. D., Afonine, P. V, Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica, 66*, 213-21.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315,* 1709-12.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*, 215-33.

Benda, C., Ebert, J., Scheltema, R. A., Schiller, H. B., Baumgärtner, M., Bonneau, F., Conti, E. (2014). Structural Model of a CRISPR RNA-Silencing Complex Reveals the RNA-Target Cleavage Activity in Cmr4. *Molecular Cell*, *56*, 43-54.

Berkner, S., Wlodkowski, A., Albers, S. V., Lipps, G. (2010). Inducible and constitutive promoters for genetic systems in Sulfolobus acidocaldarius. *Extremophiles*, *14*, 249-59.

Bhaya, D., Davison, M., Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual Review of Genetics*, *45*, 273-97.

Blin, G., Denise, A., Dulucq, S., Herrbach, C., Touzet, H. (2009). Alignments of RNA structures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*, 309-22.

Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M., Benning, C. (1998). AGO1 defines a novel locus of Arabidopsis controlling leaf development, *17*, 170-80.

Bolotin, A., Quinquis, B., Sorokin, A., Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, *151*, 2551-61.

Breitung, J., Borner, G., Scholz, S., Linder, D., Stetter, K., Thauer, R. K. (1992). Salt dependence, kinetic properties and catalytic mechanism of N-formy1methanofuran: tetrahydromethanopterin formyltransferase from the extreme thermophile Methanopyrus kandleri. *Eurpean Journal of Biochemistry, 981*, 971-81.

Brochier, C., Forterre, P., Gribaldo, S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries : tackling the Methanopyrus kandleri paradox. *Genome Biology*, *5*, 17.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science, 321*, 960-4.

Burggraf, S., Stetter, K. O., Rouviere, P., Woese, C. R. (1991). Methanopyrus kandleri: an archaeal methanogen unrelated to all other known methanogens. *Systematic and Applied Microbiology, 14*, 346-51.

Carbonell, A., Flores, R., Gago, S. (2011). Trans-cleaving hammerhead ribozymes with tertiary stabilizing motifs: in vitro and in vivo activity against a structured viroid RNA. *Nucleic Acids Research*, *39*, 2432-44.

Carte, J., Wang, R., Li, H., Terns, R. M., Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & Development*, *22*, 3489-96.

Chan, P. P., Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, *37*, 93-7.

Cocozaki, A. I., Ramia, N. F., Shao, Y., Hale, C. R., Terns, R. M., Terns, M. P., Li, H. (2012). Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure*, *20*, 545-53.

Danan, M., Schwartz, S., Edelheit, S., Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Research*, *40*, 3131-42.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A, Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*, 602-7.

Elkayam, E., Kuhn, C.-D., Tocilj, A., Haase, A. D., Greene, E. M., Hannon, G. J., Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. *Cell*, *150*, 100-10.

Emsley, P., Cowtan, K. (2002). Coot: model-building tools for molecular graphics. *Acta Crystallographica, 60*, 2126-32.

Fechter, P., The, A., Giege, R. (2000). Identity of tRNA for Yeast Tyrosyl-tRNA Synthetase : Tyrosylation Is More Sensitive to Identity Nucleotides than to Structural Features. *Biochemistry, 39*, 1725-33.

Findeiss, S., Langenberger, D., Stadler, P. F., Hoffmann, S. (2011). Traces of post-transcriptional RNA modifications in deep sequencing data. *The Journal of Biological Chemistry, 392*, 305-13.

Forterre, P. (2006). DNA topoisomerase V: a new fold of mysterious origin. *Trends in Biotechnology*, *24*, 245-7.

Fribourg, S., Gatfield, D., Izaurralde, E., Conti, E. (2003). A novel mode of RBD-protein recognition in the Y14-Mago complex. *Nature Structural Biology*, *10*, 433-9.

Garside, E. L., Schellenberg, M. J., Gesner, E. M., Bonanno, J. B., Sauder, J. M., Burley, S. K., MacMillan, A. M. (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*, *18*, 2020-8.

Giacalone, M., Gentile, A., Lovitt, B., Berkley, N., Gunderson, C., Surber, M. (2006). Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system. *BioTechniques*, *40*, 355-64.

Grissa, I., Vergnaud, G., Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*, 52-7.

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.

Haas, E. S., Daniels, C. J., Reeve, J. N. (1989). Genes encoding 5s rRNA and tRNAs in the extremely thermophilic thermus fewidus archaebacterium. *Gene, 77*, 253-63.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Terns, M. P. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell*, *45*, 292-302.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, *139*, 945-56.

Hatoum-Aslan, A., Maniv, I., Marraffini, L. A. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 21218-22.

Hatoum-Aslan, A., Maniv, I., Samai, P., Marraffini, L. A. (2014). Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. *Journal of Bacteriology*, *196*, 310-7.

Hatoum-Aslan, A., Samai, P., Maniv, I., Jiang, W., Marraffini, L. A. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *The Journal of Biological Chemistry*, *288*, 27888-97.

Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K., Doudna, J. A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, *329*, 1355-8.

Hochstrasser, M. L., Taylor, D. W., Bhat, P., Guegler, C. K., Sternberg, S. H., Nogales, E., Doudna, J. A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 6618-23.

Hrle, A., Su, A. A. H., Ebert, J., Benda, C., Randau, L., Conti, E. (2013). Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biology*, *10*, 1670-8.

Huber, R., Kurr, M., Jannasch, H. W., Stetter, K. O. (1989). A novel group of abyssal methanogenic archaebacteria (Methanopyrus) growing at 110°C. *Nature, 342,* 833-4.

Hur, J. K., Olovnikov, I., Aravin, A. A. (2014). Prokaryotic Argonautes defend genomes against invasive DNA. *Trends in Biochemical Sciences*, *39*, 257-9.

Hutvagner, G., Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. *Nature Reviews Molecular Cell Biology*, *9*, 22-32.

Jansen, R., Embden, J. D. A. Van, Gaastra, W., Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, *43*, 1565-75.

Jinek, M., Doudna, J. A. (2009). A three-dimensional view of the molecular machinery of RNA interference. *Nature*, *457*, 405-12.

Jore, M. M., Lundgren, M., van Duijn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., Brouns, S. J. J. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology*, *18*, 529-36.

Joshua-Tor, L., Hannon, G. J. (2011). Ancestral roles of small RNAs: an Ago-centric perspective. *Cold Spring Harbor Perspectives in Biology*, *3*, 3772.

Juranek, S., Eban, T., Altuvia, Y., Brown, M., Morozov, P., Tuschl, T., Margalit, H. (2012). A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs. *RNA*, *18*, 783-94.

Kabsch, W. (2010). XDS. *Acta Crystallographica, 66*, 125-32.

Kadlec, J., Izaurralde, E., Cusack, S. (2004). The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nature Structural & Molecular Biology*, *11*, 330-7.

Kehr, S., Bartschat, S., Stadler, P. F., Tafer, H. (2011). PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics*, *27*, 279-80.

Ketting, R. F. (2011). The many faces of RNAi. *Developmental Cell*, *20*, 148-61.

Klein, D. J., Schmeing, T. M., Moore, P. B., Steitz, T. A. (2001). The kink-turn : a new RNA secondary structure motif. *The EMBO Journal, 20*, 4214-21.

Koonin, E. V, Makarova, K. S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biology*, *10*, 679-86.

Kowalak, J. A., Dalluge, J. J., Mccloskey, J. A., Stetters, K. (1994). The Role of Posttranscriptional Modification in Stabilization of Transfer RNA from Hyperthermophiles. *Biochemistry, 33*, 7869-76.

Krah, R., Kozyavkin, S. A., Slesarevt, A. I., Gellert, M. (1996). A two-subunit type I DNA topoisomerase (reverse gyrase) from an extreme hyperthermophile, *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 106-10.

Labrie, S. J., Samson, J. E., Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews. Microbiology*, *8*, 317-27.

Larkin, M. a, Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A, McWilliam, H., Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947-8.

Lintner, N. G., Kerou, M., Brumfield, S. K., Graham, S., Liu, H., Naismith, J. H., Lawrence, C. M. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *The Journal of Biological Chemistry*, *286*, 21643-56.

Liu, J., Carmell, M. A, Rivas, F. V, Marsden, C. G., Thomson, J. M., Song, J. J., Hannon, G. J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, *305*, 1437-41.

Lowe, T. M. (1999). A Computational Screen for Methylation Guide snoRNAs in Yeast. *Science*, *283*, 1168-71.

Lowe, T.M., Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*, 955-64.

Ma, J. B., Yuan, Y. R., Meister, G., Pei, Y., Tuschl, T., Patel, D. J. (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, *434*, 666-70.

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

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

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Koonin, E. V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nature Reviews. Microbiology*, *9*, 467-77.

Makarova, K. S., Wolf, Y. I., Koonin, E. V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*, *41*, 4360-77.

Makarova, K. S., Wolf, Y. I., van der Oost, J., 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. *Biology Direct*, *4*, 29.

Malakhov, M. P., Mattern, M. R., Malakhova, O. A, Drinker, M., Weeks, S. D., Butt, T. R. (2004). SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *Journal of Structural and Functional Genomics*, *5*, 75-86.

Marblestone, J. G., Edavettal, S. C., Lim, Y., Lim, P., Zuo, X. U. N., Butt, T. R. (2006). Comparison of SUMO fusion technology with traditional gene fusion systems : Enhanced expression and solubility with SUMO. *Protein Science, 15*, 182-9.

Marck, C., Grosjean, H. (2003). Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea : evolutionary implications Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea : evolutionary implications. *RNA, 9*, 1516-31.

Maris, C., Dominguez, C., Allain, F. H. T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS Journal*, *272*, 2118-31.

Marraffini, L. A., Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews. Genetics*, *11*, 181-90.

Marraffini, L. A., Sontheimer, E. J. (2008). CRISPR Interference Limits Horizontal Targeting DNA, *Science, 322*, 1843-5.

Mat, W. K., Xue, H., Wong, J. T. (2008) The genomics of LUCA. *Frontiers in Bioscience, 13*, 5605-13.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography, 40*, 658 - 74.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature Reviews. Genetics*, *14*, 447-59.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, *60*, 174-82.

Mole, L. D. B., Sabatier, P. (2001). Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO Journal, 20*, 3617-22.

Mulepati, S., Bailey, S. (2011). Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *The Journal of Biological Chemistry*, *286*, 31896-903.

Nakanishi, K., Weinberg, D. E., Bartel, D. P., Patel, D. J. (2012). Structure of yeast Argonaute with guide RNA. *Nature*, *486*, 368-74.

Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M. P., Ke, A. (2012). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*, *20*, 1574-84.

Nowotny, M., Gaidamakov, S. A, Crouch, R. J., Yang, W. (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*, *121*, 1005-16.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V, Davies, C. W., Doudna, J. A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Structural & Molecular Biology*, *21*, 528-34.

Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D., Aravin, A. (2013). Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA. *Molecular Cell*, *51*, 594-605.

Omer, A. D. (2000). Homologs of Small Nucleolar RNAs in Archaea. *Science*, *288*, 517-22.

Osawa, T., Inanaga, H., Numata, T. (2013). Crystal structure of the Cmr2-Cmr3 subcomplex in the CRISPR-Cas RNA silencing effector complex. *Journal of Molecular Biology*, *425*, 3811-23.

Palmer, J. R., Baltrus, T., Reeve, J. N., Daniels, C. J. (1992). Transfer RNA genes from the hyperthermophilic Archaeon, Methanopyrus kandleri. *Biochimica et Biophysica Acta, 1132*, 315-8.

Parker, J. S., Roe, S. M., Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature Letters, 434*, 663-6.

Peters, L., Meister, G. (2007). Argonaute proteins: mediators of RNA silencing. *Molecular Cell*, *26*, 611-23.

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub,H., Randau, L. (2014). In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Research*, *42*, 5125-38.

Puttaraju, M., Been, M. D. (1995). Generation of nuclease resistant circular RNA decoys for HIV-Tat and HIV-Rev by autocatalytic splicing. *Nucleic Acids Symposium Series, 1995*, 49-51.

Raabe, C. a, Hoe, C. H., Randau, G., Brosius, J., Tang, T. H., Rozhdestvensky, T. S. (2011). The rocks and shallows of deep RNA sequencing: Examples in the Vibrio cholerae RNome. *RNA*, *17*, 1357-66.

Randau, L. (2012). RNA processing in the minimal organism Nanoarchaeum equitans. *Genome Biology*, *13*, 63.

Randau, L., Söll, D. (2008). Transfer RNA genes in pieces. *EMBO Reports*, *9*, 623-8.

Randau, L., Stanley, B. J., Kohlway, A., Mechta, S., Xiong, Y., Söll, D. (2009). A cytidine deaminase edits C to U in transfer RNAs in Archaea. *Science*, *324*, 657-9.

Rashid, U. J., Paterok, D., Koglin, A., Gohlke, H., Piehler, J., Chen, J. C. H. (2007). Structure of Aquifex aeolicus argonaute highlights conformational flexibility of the PAZ domain as a potential regulator of RNA-induced silencing complex function. *The Journal of Biological Chemistry*, *282*, 13824-32.

Reeks, J., Naismith, J. H., White, M. F. (2013). CRISPR interference: a structural perspective. *The Biochemical Journal*, *453*, 155-66.

Mojica FJ, Díez-Villaseñor C, Soria E, Juez G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology, 36*, 244-6.

Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., Backofen, R., Randau, L. (2012). Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic Acids Research*, *40*, 9887-96.

Rivas, F. V, Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., Hannon, G. J., Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nature Structural & Molecular Biology*, *12*, 340-9.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., White, M. F. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Molecular Cell*, *52*, 124-34.

Rozhdestvensky, T. S. (2003). Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Research*, *31*, 869-77.

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PloS One*, *7*, e30733.

Samson, J. E., Magadán, A. H., Sabri, M., Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews. Microbiology*, *11*, 675-87.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Research*, *39*, 9275-82.

Sauerwald, A., Sitaramaiah, D., McCloskey, J. A, Söll, D., Crain, P. F. (2005). N6-Acetyladenosine: a new modified nucleoside from Methanopyrus kandleri tRNA. *FEBS Letters*, *579*, 2807-10.

Scholz, J., Besir, H., Strasser, C., Suppmann, S. (2013). A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. *BMC Biotechnology*, *13*, 12.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Vogel, J. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, *464*, 250-5.

Shen, A., Lupardus, P. J., Morell, M., Ponder, E. L., Sadaghiani, A. M., Garcia, K. C., Bogyo, M. (2009). Simplified, enhanced protein purification using an inducible, autoprocessing enzyme tag. *PloS One*, *4*, e8119.

Sheng, G., Zhao, H., Wang, J., Rao, Y., Tian, W., Swarts, D. C., van der Oost, J. (2014). Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 652-7.

Shima, S., Hérault, D. A., Berkessel, A., Thauer, R. K. (1998). Activation and thermostabilization effects of cyclic 2,3-diphosphoglycerate on enzymes from the hyperthermophilic Methanopyrus kandleri. *Archaeal Microbiology, 170*, 469-72.

Siomi, M. C., Sato, K., Pezic, D., Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews. Molecular Cell Biology*, *12*, 246-58.

Slesarev, A. I., Mezhevaya, K. V, Makarova, K. S., Polushin, N. N., Shcherbinina, O. V, Shakhova, V. V, Kozyavkin, S. A. (2002). The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 4644-9.

Slesarev A. I., Stetter K. O., Lake J. A., Gellert M., Krah R., Kozyavkin S. A. (1993). DNA topoisomerase V is a relative of eukaryotic topoisomerase I from a hyperthermophilic prokaryote. *Nature, 364*, 735-7.

Song, J. J., Smith, S. K., Hannon, G. J., Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science*, *305*, 1434-7.

Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Stagg, S. (2013). Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Molecular Cell*, *52*, 146-52.

Staals, R. H. J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D. W., van Duijn, E., Shinkai, A. (2013). Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. *Molecular Cell*, *52*, 135-45.

Starostina, N. G., Marshburn, S., Johnson, L. S., Eddy, S. R., Terns, R. M., Terns, M. P. (2004). Circular box C/D RNAs in Pyrococcus furiosus. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 14097-101.

Su, A. A. H., Tripp, V., Randau, L. (2013). RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri. *Nucleic Acids Research*, *41*, 6250-8.

Suttle, C. A. (2007). Marine viruses, major players in the global ecosystem. *Nature Reviews. Microbiology*, *5*, 801-12.

Swarts, D. C., Jore, M. M., Westra, E. R., Zhu, Y., Janssen, J. H., Snijders, A. P., van der Oost, J. (2014). DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, *507*, 258-61.

Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., Horikoshi, K. (2008). Cell proliferation at 122°C and isotopically heavy CH_4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 10949-54.

Van der Oost, J., Westra, E. R., Jackson, R. N., Wiedenheft, B. (2014). Unraveling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews. Microbiology*, *12*, 479-92.

Van Duijn, E., Barbu, I. M., Barendregt, A., Jore, M. M., Wiedenheft, B., Lundgren, M., Heck, A. J. R. (2012). Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. *Molecular & Cellular Proteomics, 11*, 1430-41.

Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A., Boekema, E., Zhou, K. (2011). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences*, *108*, 15010.

Wagner, M., Berkner, S., Ajon, M., Driessen, A. J. M., Lipps, G., Albers, S.-V. (2009). Expanding and understanding the genetic toolbox of the hyperthermophilic genus Sulfolobus. *Biochemical Society Transactions*, *37*, 97-101.

Wang, R., Preamplume, G., Terns, M. P., Terns, R. M., Li, H. (2011). Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*, *19*, 257-64.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., Patel, D. J. (2008). Structure of the guide-strand-containing argonaute silencing complex. *Nature*, *456*, 209-13.

Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J. J., van der Oost, J., Nogales, E. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*, *477*, 486-9.

Wiedenheft, B., Sternberg, S. H., Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, *482*, 331-8.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S. M., Ma, W., Doudna, J. A. (2009). Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure*, *17*, 904-12.

Yamaguchi, Y., Park, J.-H., Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. *Annual Review of Genetics*, *45*, 61-79.

Yip, W. S. V., Vincent, N. G., Baserga, S. J. (2013). Ribonucleoproteins in archaeal pre-rRNA processing and modification. *Archaea*, *2013*, 614735.

Yoshihisa, T., Yunoki-esaki, K., Ohshima, C., Tanaka, N., Endo, T. (2003). Possibility of cytoplasmic pre-tRNA splicing : the yeast tRNA splicing endonuclease mainly localizes on the mitochondria. *Molecular Biology of the Cell*, *14*, 3266-79.

Yuan, Y.-R., Pei, Y., Ma, J.-B., Kuryavyi, V., Zhadina, M., Meister, G., Patel, D. J. (2005). Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Molecular Cell*, *19*, 405-19.

Zander, A., Holzmeister, P., Klose, D., Tinnefeld, P., Grohmann, D. (2014). Single-molecule FRET supports the two-state model of Argonaute action. *RNA Biology, 11,* 45-56.

Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., White, M. F. (2012). Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular Cell*, *45*, 303-13.

Zhu, X., Ye, K. (2012). Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. *FEBS Letters*, *586*, 939-45.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, *31*, 3406-15.

Danksagung

Ich bedanke mich sehr herzlich bei Dr. Lennart Randau für die Möglichkeit meine Doktorarbeit in seinem Labor durchzuführen und auch für eine exzellente Betreuung, Unterstützung und stete Gesprächsbereitschaft.

Herrn Prof. Dr. Rudolf Thauer, Frau Prof. Dr. Sonja Albers und Frau Prof. Dr. Eva Stuckenbrock danke ich sehr für ihre Unterstützung und Ratschläge als Mitglieder meines „IMPRS Thesis Advisory Committees“.

Ein großer Dank gilt hier nochmals Herrn Prof. Dr. Rudolf Thauer für die Übernahme der Zweitkorrektur als auch Frau Prof. Dr. Renate Renkawitz-Pohl und Herrn Prof. Dr. Torsten Waldminghaus für die spontane Zusage meiner Prüfungskommission anzugehören.

Bei Prof. Dr. Elena Conti, Dr. Christian Benda, Ajla Hrle und Judith Ebert bedanke ich mich für eine erfolgreiche Kollaboration.

Bei der gesamten AG Randau sowie der Abteilung B2 bedanke ich mich herzlich für eine angenehme und freundschaftliche Arbeitsatmosphäre sowie anregende Diskussionen als auch für die schöne Zeit neben der Laborarbeit. Besonderer Dank gilt hierbei: Andre, Basti, Hagen, Jeanine, Judith, Julia, Michi, Poppi, Sri und Vanessa.

Ganz besonders herzlich danke ich meinen Eltern, sowie Nina für die Große Unterstützung und Hilfe, ohne die all dies nicht möglich gewesen wäre.

Erklärung des Eigenanteils

Hrle, A*., **Su, A.A.H***., Ebert, J., Benda, C., Randau, L*. and Conti, E*. (* joint first/corresponding authorship) *Structure and RNA-binding properties of the Type III-A CRISPR-associated protein Csm3.* RNA Biology, 2013, 10, 1670-1678.

Das Protein Csm3 wurde von Andreas Su hergestellt und zusammen mit Akiyoshi Nakamura (Yale Universität, USA) testweise kristallisiert. Ajla Hrle, Andreas Su und Lennart Randau haben die Experimente entworfen. Die Kristallstruktur wurde von Ajla Hrle und Christian Benda gelöst. Andreas Su führte die Experimente zu Abbildung III 3A sowie RNA Bindungs- und Prozessierungsstudien durch und Ajla Hrle führte die übrigen Experimente durch. Das Manuskript wurde von Ajla Hrle, Prof. Dr. Elena Conti und Dr. Lennart Randau geschrieben sowie von allen Autoren überarbeitet.

Su, A.A.H., Tripp, V. and Randau, L. *RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri* Nucleic Acids Res, 2013, 41, 6250-6258.

Die Versuche, sowie die bioinformatischen Analysen von C/D box sRNAs und crRNAs wurden von Andreas Su durchgeführt. Die bioinformatische Analyse von möglichen C/D box sRNA Zielen wurden von Vanessa Tripp und Andreas Su durchgeführt. Bioinformatische Analysen der tRNAs wurden von Dr. Lennart Randau durchgeführt. Die Veröffentlichung wurde von Dr. Lennart Randau geschrieben und von allen Autoren überarbeitet.

Marburg, den 13.11.2014

_____

Andreas Su

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich meine Dissertation mit dem Titel: "Small RNA-guided processes in the hyperthermophilic methanogen *Methanopyrus kandleri*" selbständig und ohne unerlaubte Hilfsmittel angefertigt, sowie keine anderen als die von mir ausdrücklich angegebenen Quellen verwendet habe.

Die Dissertation wurde in der jetzigen oder ähnlichen Form bei keiner anderen Hochschule eingereicht und hat noch keinem sonstigen Prüfungszwecken gedient.

Marburg, den 13.11.2014

_____

Andreas Su