Understanding the expression stage of CRISPR-Cas defense system in *Leptospira interrogans*

A thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

AMAN PRAKASH

Under the supervision of

Prof. Manish Kumar

Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati-781039, Assam, India

November, 2022

Understanding the expression stage of CRISPR-Cas defense system in *Leptospira interrogans*

by

AMAN PRAKASH

IIT Guwahati, 2022

DOCTORAL COMMITTEE

Prof. Manish Kumar (Department of Biosciences and Bioengineering)
Supervisor

Prof. Ajaikumar B. Kunnumakkara (Department of Biosciences and Bioengineering)
Chairperson

Prof. Aiyagari Ramesh (Department of Biosciences and Bioengineering)
Member

Prof. Ashish Anand (Department of Computer Science and Engineering)
Member

DECLARATION

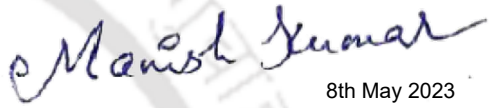
I hereby declare that the work presented in this thesis entitled "**Understanding the expression stage of CRISPR-Cas defense system in *Leptospira interrogans***" is entirely original and was carried out by me under the supervision of Prof. Manish Kumar, Indian Institute of Technology Guwahati, Assam, India.

In accordance with the standard procedure for publishing scientific observations, appropriate acknowledgments have been made wherever the research findings of other investigators have been referenced in the thesis.

Date: 1st November 2022 **Aman Prakash**

CERTIFICATE

This is to certify that the work described in this thesis entitled "**Understanding the expression stage of CRISPR-Cas defense system in *Leptospira interrogans***" is the result of investigations carried out by Aman Prakash at Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India under my supervision and this work has not been submitted elsewhere for the award of any other degree.



8th May 2023

Prof. Manish Kumar

(Thesis Supervisor)

Dedicated to my

Family

ACKNOWLEDGEMENTS

I am extremely grateful for the opportunity to pursue my doctorate degree at IIT Guwahati. I am indebted to many people who helped me during this endeavor and would like to express my sincere gratitude to them.

I am extremely grateful to my thesis supervisor Prof. Manish Kumar for the opportunity to be a part of his research group and for his unwavering faith in me. His constant support, enthusiasm, and motivation were extremely valuable during tough times. His scientific insights helped me immensely in my thesis work. His guidance in the correct direction and encouragement to pursue new research ideas motivated me in the journey of self-improvement.

I am extremely thankful to my Doctoral Committee members Prof. Ajaikumar B. Kunnumakkara, Prof. Aiyagari Ramesh, Prof. Ashish Anand for their valuable suggestions, constant encouragement, and critical assessment of my Ph.D. work, which helped me in improving my critical thinking skills.

IIT Guwahati's Department of Biosciences and Bioengineering's DCIF facilities were utilized for a number of studies. I am appreciative of the staff members for their assistance. IIT Guwahati offered the resources I needed to further my academic objectives, and the Indian Ministry of Human Resource Development (MHRD) provided the funding.

I also appreciate the administration of IIT Guwahati for quickly reopening the research laboratories following the COVID lockdowns. Even under challenging conditions, access to the research facilities was maintained, enabling students to conduct experiments.

Among the many people to whom I owe my gratitude, I would like to record my heartfelt thanks to Dr. Karukriti Ghosh and Dr. Bhuvan Dixit for their helping hand during my research work. I also record my sincere thanks to Dr. Anusua Dhara, Md. Saddam Hussain, Vineet Anand, and all the present and past lab members for their constant help and support during the entire research period.

I am appreciative of my friends Abhijeet, Anurag, Avishek, Jon, Himadree, and Monika for making my stay at IIT Guwahati a memorable one. Rajib Da's tea and stories provided the perfect short breaks from long days in the lab.

My parents have supported and loved me unconditionally. I have become a better person and professional because of their advice, guidance, and love. I am grateful to my family, especially my younger sister, who never fails to inspire and motivate me. Jimmy personified the proverb a friend in need is a friend indeed for me.

It would be hard to list everyone who has aided me in my academic endeavors. I sincerely appreciate each and every one of them for having faith in me.

Aman Prakash ***November 2022***

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

aa	amino acid
Abi	Abortive infection
ARAMP	Asgard Repeat-Associated Mysterious Proteins
ATP	Adenosine triphosphate
bp	base pair
BREX	Bacteriophage exclusion system
BSA	Bovine serum albumin
BLAST	Basic Local Alignment Search Tool
Cascade	CRISPR-associated complex for antiviral defense
CB	Cleavage buffer
CRISPR-Cas	Clustered Regularly Interspaced Short Palindromic Repeats and associated genes
CRISPRi	CRISPR interference
cDNA	complementary DNA
crRNA	CRISPR RNA
DNA	Deoxyribonuleic acid
DNase	Deoxyribonuclease
ds-DNA	double-stranded DNA
DSBs	Double-strand breaks
EL	Evidence level
ELISA	Enzyme-linked immunosorbent assay
EDTA	Ethylenediamine tetraacetic acid
EMJH	Ellinghausen-McCullough-Johnson-Harris
EMSA	Electrophoretic mobility shift assay
G-loop	Glycine rich loop
gDNA	Genomic DNA
h	Hour
HR	Homologous recombination
HRP	Horseradish peroxidase
HRAMP	Halobacterial repeat-associated mysterious proteins
IPTG	Isopropyl ß-D-1-thiogalactopyranoside
kb	Kilo base pair

kDa	Kilodalton
L	Liter
LA_Cr	*Leptospira interrogans* serovar Lai CRISPR array
LIC_Cr	*Leptospira interrogans* serovar Copenhageni CRISPR array
µg	Microgram
µM	Micromolar
mg	Milligram
MGEs	Mobile genetic elements
mM	Millimolar
MSA	Multiple sequence alignment
NAB	Nuclease activity buffer
NEC	No enzyme control
NFW	Nuclease free water
ng	Nanogram
NHEJ	Non-homologous end joining
nm	Nanometer
nM	Nanomolar
NTA	Nitrilotriacetic acid
ORF	Open reading frame
PAA	Polyacrylamide
PAM	Protospacer adjacent motif
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
pre-crRNA	Precursor CRISPR RNA
q-PCR	Quantitative real-time polymerase chain reaction
RAMP	Repeat-associated mysterious proteins
RC	Repeat consensus
RH	Random hexamers
RT-PCR	Reverse transcription-polymerase chain reaction
RPM	Rotation per minute
RNA	Ribonucleic acid
RNase	Ribonuclease

RRM	RNA Recognition Motif
RV	Repeat variant
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
ss-DNA	single-stranded DNA
SUMO	Small ubiquitin-like modifier
TALE	Transcription activator-Like effectors
TBE	Tris-borate EDTA
TMB	Trimethylbenzidine
tracrRNA	trans-activating CRISPR RNA
UV	Ultraviolet

ABSTRACT

Leptospira is a genus of spiral-shaped bacteria known as spirochetes, and its pathogenic form is notorious for causing disease in humans and animals. The function of virulent genes in *Leptospira* spp. is still confined due to the lack of efficient genetic manipulation tools. In pathogenic species of *Leptospira*, harnessing endogenous CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated proteins) system, an RNA-mediated immunity process against foreign nucleic acid, is an attractive strategy to study its pathogenesis by reverse genetics approach. However, reprogramming the endogenous CRISPR-Cas system for genome editing relies on understanding the CRISPR array processing to impart immunity against the gene of interest. Genomes of *Leptospira interrogans* harbor the CRISPR-Cas I-B locus in which the CRISPR array is flanked by two cas-operons. This study characterizes the CRISPR arrays of two widely studied pathogenic *L. interrogans* (serovars Copenhageni and Lai) whose genome sequence is available in the NCBI (National Center for Biotechnology Information). In addition, three CRISPR-associated proteins of serovar Copenhageni (LinCas6, LinCas5, and LinCas3) involved in the expression and interference of RNA-mediated immunity are characterized. Using the RT-PCR (reverse transcription-polymerase chain reaction), we account for the transcriptionally active CRISPR arrays in the direction of *cas*-operons. The recombinant LinCas6 (rLinCas6) overexpressed and purified in this study can process the precursor-CRISPR RNA (pre-crRNA) of subtype I-B to generate mature crRNA and remains bound with it. The rLinCas6 follows single turnover kinetics during the conventional cleavage of its cognate repeat RNA substrate. In rLinCas6, substitution of one of the predicted active site residues (His38) resulted in reduced cleavage activity. Biochemical analysis of the overexpressed and purified recombinant LinCas5 (rLinCas5) suggested that it is catalytically inactive on nucleic acids. However, rLinCas5 binds to the rLinCas6-crRNA complex essential for stabilizing the mature crRNA during interference. Biochemical analysis of the purified rLinCas3 demonstrated that it is a metal-dependent nuclease (DNase and RNase). The presence of nucleotide in the reaction intervenes with the rLinCas3 nuclease activity on circular single/double-stranded DNA substrate. This study features insight into CRISPR transcription, crRNA biogenesis, and the onset of the effector complex formation in *Leptospira,* which is essential for RNA-mediated interference of invading nucleic acids. In addition, this study proposes the physiological requirements of *Leptospira* CRISPR-Cas I-B during interference.

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY TABLES

CHAPTER 1

Introduction and Literature Review

1.1 Host-parasite coevolution

Bacteria and archaea are prevalent in nature and encounter various foreign mobile genetic elements (MGEs), such as plasmids, transposons, and phages (García-Aljaro, Ballesté et al. 2017). These MGEs can mediate horizontal gene transfer between prokaryotes and are responsible for the transmission of virulence genes, antibiotic resistance genes and phage resistance genes in prokaryotes (Colavecchio, Cadieux et al. 2017; Touchon, De Sousa et al. 2017). Many MGEs integrate into the host DNA and constitute up to 30% of bacterial genomes (Koonin and Krupovic 2015). Since the integration of foreign genetic material into prokaryotic genomes affects the prokaryotic life cycle, prokaryotes have evolved with several defense systems to protect themselves from invasion by MGEs (Makarova, Wolf et al. 2013). As a countermeasure, many defense systems have been identified in the host genomes, which have been classified as innate (non-specific) and adaptive (highly specific) immune systems of the prokaryotes (Koonin, Makarova et al. 2017). Some of the innate immune systems, such as restriction-modification (R-M), phage abortive infection (Abi), and bacteriophage exclusion (BREX) are being briefly described.

1.1.1 Restriction-modification (R-M) system

The R-M defense mechanism allows prokaryotes to differentiate between the host genome and invading MGEs. An R-M system employs a pair of enzymes called DNA methyltransferase (MTase) and restriction endonuclease (REase) (Blow, Clark et al. 2016). A DNA MTase methylates the host DNA at the REase site. Meanwhile, REase recognizes and cleaves the non-methylated invading DNA at the REase site. Thus, R-M systems eliminate the invading MGEs like phages and protect the host genome. However, the phages evade this sequence-specific defense through mutation of the restriction sites in invading DNA (Labrie, Samson et al. 2010).

1.1.2. Phage abortive infection (Abi)

Abi systems protect the prokaryotic populations by aborting the production of progeny phages and sacrificing the infected cells (Labrie, Samson et al. 2010). Abi systems generally target essential cellular processes like replication, transcription, and translation (Labrie, Samson et al. 2010). For instance, the Rex system in λ-lysogenic *E. coli* strains is composed of proteins RexA and RexB that function as intracellular sensor and ion channel, respectively (Labrie, Samson et al. 2010). During phage infection, RexA gets activated by recognizing a phage protein-DNA complex. This is followed by activation of the RexB that causes a drop in cellular ATP level, leading to termination of cell multiplication and abortion of phage's lytic growth (Parma, Snyder et al. 1992; Snyder 1995). The Abi system sometimes overlaps with another system called "toxin-antitoxin" (TA) which depends on the dual activity of a toxin and its antagonistic/antitoxin (Gerdes, Christensen et al. 2005). Similar to the Abi proteins, toxins of TA systems can target DNA replication and translation by inhibiting DNA gyrase and causing mRNA degradation, respectively (Gerdes, Christensen et al. 2005).

1.1.3. Bacteriophage exclusion (BREX)

Similar to the R-M systems, the BREX system's phage resistance mechanism relies on self and non-self DNA discrimination (Goldfarb, Sberro et al. 2015). Prokaryotes possessing BREX systems methylate the fifth position of a non-palindromic sequence 5′-TAGGAG-3′ in phage genomes and can prevent phage DNA replication. In contrast to the R-M systems, the BREX systems do not cleave phage DNA (Goldfarb, Sberro et al. 2015).

1.2. CRISPR-Cas adaptive immune systems

DNA-encoded CRISPR-Cas (Cluster Regularly Interspaced Short Palindromic Repeats-CRISPR-associated proteins) act as RNA-mediated adaptive immune systems and nucleic acid-targeting interference in prokaryotes (Barrangou and Horvath 2017). These systems operate alongside innate immune systems to protect the prokaryotic population against MGEs (Labrie, Samson et al. 2010). In contrast to other cellular defense mechanisms that provide generic protection against MGEs, CRISPR-Cas stores the records of earlier infections to evoke a rapid and robust response in case of reinfection (Labrie, Samson et al. 2010).

CRISPR-Cas systems have been identified in almost 90% and 50% of sequenced genomes of archaea and bacteria (Grissa, Vergnaud et al. 2007; Makarova, Wolf et al. 2015). A hallmark

of any CRISPR-Cas system is the CRISPR array which is composed of a series of direct repeat sequences (20-50 bp) interspaced with unique spacer sequences of similar lengths **(Figure 1.1)**. These spacers originate from MGEs during past infections and serve as a genetic memory (Mojica, García-Martínez et al. 2005; Pourcel, Salvignol et al. 2005). Generally, the upstream region of the CRISPR array is an AT-rich sequence called "leader" that contains the promoter for CRISPR array transcription. In close proximity to the CRISPR array, a set of *cas* genes encoding Cas proteins requisite in the functioning of the CRISPR-Cas system (Jansen, Embden et al. 2002). Generally, CRISPR immunity is driven by the Cas proteins in three distinct molecular stages: (1) adaptation or spacer acquisition, (2) Expression or CRISPR RNA (crRNA) biogenesis, and (3) interference **(Figure 1.2)**. In the first stage (adaptation), a short DNA stretch from an invading genetic element is captured and incorporated into a CRISPR array as the first spacer immediately after a leader sequence (Yosef, Goren et al. 2012). In the expression stage, the entire CRISPR array is transcribed into a precursor CRISPR RNA (pre-crRNA) driven by promoter elements of the leader sequence (Yosef, Goren et al. 2012; Charpentier, Richter et al. 2015). After transcription, the cleavage within the repeats of the pre-crRNA by ribonucleases generates mature crRNAs, each carrying a unique foreign sequence. In the final stage (interference), each crRNA forms a ribonucleoprotein complex with Cas proteins. The effector machinery guides the crRNA to the matching region of the invader nucleic acids and recruits a signature Cas protein for destruction (Barrangou, Fremaux et al. 2007; Brouns, Jore et al. 2008; Garneau, Dupuis et al. 2010; Marraffini and Sontheimer 2010; Nam, Haitjema et al. 2012). A short (2-5 bp) motif called protospacer-adjacent motif (PAM) adjacent to the target protospacer DNA directs crRNA-guided invader DNA cleavage (Deveau, Barrangou et al. 2008; Semenova, Jore et al. 2011; Sashital, Wiedenheft et al. 2012; Redding, Sternberg et al. 2015).

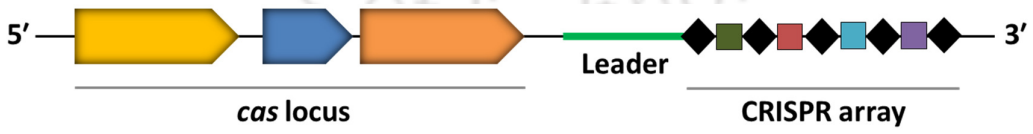


Figure 1.1. The architecture of a typical CRISPR-Cas locus in a prokaryotic genome [adapted from (Van Der Oost, Westra et al. 2014)]. CRISPR-Cas system comprises a CRISPR array containing identical direct repeats (black diamond boxes) and variable spacers (colored rectangles), an AT-rich leader region upstream to the CRISPR array (green line) and a set of genes (arrow-headed colored boxes).

To attain CRISPR immunity, many Cas proteins exhibit nuclease activity to cleave the phosphodiester bonds within nucleic acids. Nuclease can catalyze the cleavage of two types of phosphodiester bonds (5′ or 3′ of a scissile phosphate) within DNA or RNA through nucleophilic substitution. The mechanism of nuclease activity is determined by the metal-independent or metal-dependent mode of catalysis (Yang 2011). Metal-independent DNases form phosphoenzyme covalent intermediates, whereas metal-independent RNases employ 2′-OH as a nucleophile to generate 2′, 3′ cyclic phosphate intermediates (Sasnauskas, Connolly et al. 2007; Cochrane and Strobel 2008; Yang 2011). Metal-dependent nucleases (metallonucleases) catalyze the hydrolysis of the phosphodiester bonds in RNA or DNA. Metallonucleases require metal to activate the nucleophile, stabilize the transition state, and protonate the leaving group (Sasnauskas, Connolly et al. 2007; Cochrane and Strobel 2008; Dupureur 2008; Yang 2011).

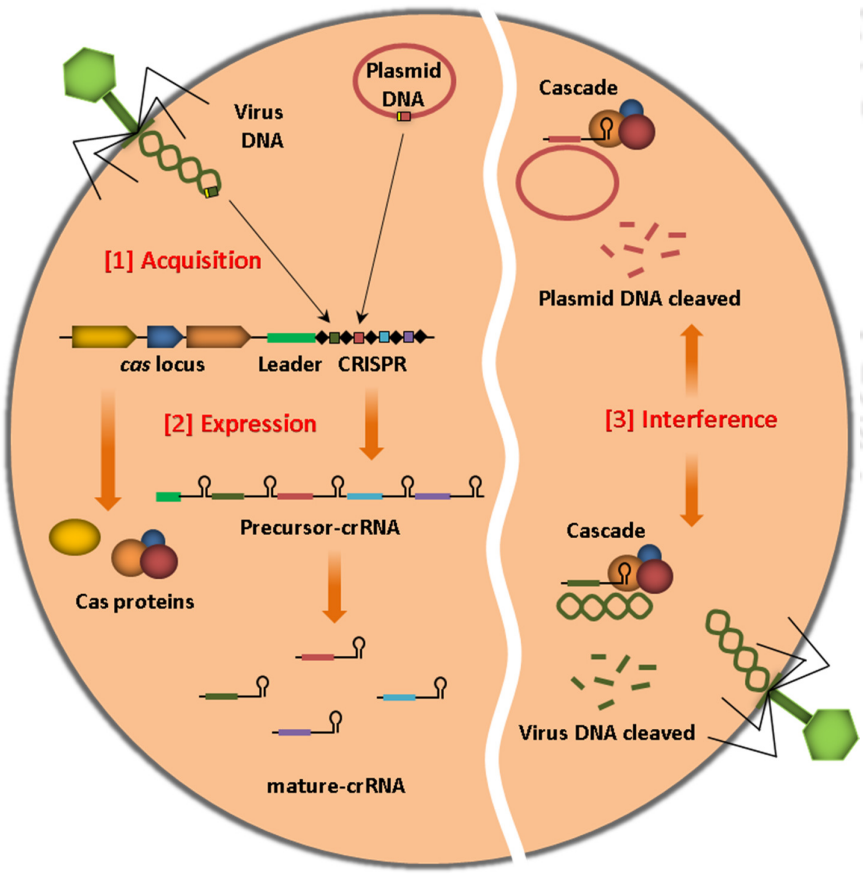


Figure 1.2. Three molecular stages of CRISPR-Cas immunity [adapted from (Sefcikova, Roth et al. 2017)]. To inactivate the invasion by MGEs, CRISPR-Cas immunity operates in three distinct molecular stages. In the first stage (adaptation or spacer acquisition), a short foreign DNA sequence (protospacer) adjacent to a PAM is selected and integrated into the CRISPR array as a spacer at the leader-repeat junction. In the second stage (expression or crRNA biogenesis), the CRISPR array is transcribed into a

long pre-crRNA, which is further processed into mature crRNAs by Cas endoribonuclease. In the final stage (interference), an effector complex formed by binding Cas proteins to the mature crRNA identifies the target sequence via complementary base-pairing between the crRNA and protospacer and PAM identification. The binding of the effector complex to the target recruits a signature Cas protein to cleave or digest the invading DNA. Black diamonds and colored rectangle boxes in the CRISPR array represent the repeats and spacers, respectively. Protospacers corresponding to the spacers is shown in phage or plasmid DNA. PAMs are demarcated by a short yellow-colored box adjacent to the protospacer region in foreign DNA. Colored circles and ovals represent Cas proteins encoded by the *cas* gene.

1.2.1 Classification of CRISPR-Cas systems

Based on the composition of effector complexes, CRISPR-Cas systems have been grouped into two classes (class 1 and 2) (Makarova, Wolf et al. 2015; Koonin, Makarova et al. 2017). Interference in class 1 systems is carried out by multi-proteins forming effector complex. In contrast, class 2 systems use a single effector protein during interference (Makarova, Wolf et al. 2015; Koonin, Makarova et al. 2017). According to the locus arrangement and the presence of signature *cas* genes that cleaves the invader DNA, these classes are further subdivided into six types (type I-VI). Each CRISPR-Cas type is further classified into multiple subtypes, often encoding subtype-specific Cas proteins (Makarova, Wolf et al. 2015; Koonin, Makarova et al. 2017). With the identification of multiple variants of Cas proteins, a total of 33 subtypes of CRISPR-Cas systems have been identified recently (Makarova, Wolf et al. 2020).

Types I, III, and IV are categorized under the class 1 CRISPR-Cas systems (**Figure 1.3),** which are most abundant in nature (Makarova, Wolf et al. 2015; Koonin, Makarova et al. 2017). The type I effector complex comprises crRNA and multiple Cas proteins called Cascade (CRISPR-associated complex for antiviral defense) (Jore, Lundgren et al. 2011). Once Cascade locates the target DNA, type I signature Cas3 protein is recruited for DNA cleavage (Sinkunas, Gasiunas et al. 2011). Cas3 exhibits nuclease and helicase activities for long-range degradation of intruder DNA (Sinkunas, Gasiunas et al. 2011). Type III system targets and cleaves crRNA complementarity RNA, followed by cleavage of ss-DNA associated with the transcription bubble. In the Type III system, Cas7 cleaves the RNA, whereas, Cas10 cleaves the ss-DNA (Taylor, Zhu et al. 2015; You, Ma et al. 2019). Type IV CRISPR-Cas system lacks an adaptation module and is mainly found in plasmids (Makarova, Wolf et al. 2020). In type IV-A, DinG helicase is essential for plasmid interference; however, the nuclease property is currently unknown (Pinilla-Redondo, Mayo-Muñoz et al. 2020).

Types II, V, and VI belong to the class 2 CRISPR-Cas systems (Makarova, Wolf et al. 2015) **(Figure 1.3)**. The type II effector protein Cas9 exhibits HNH and RuvC endonucleases that introduce ds-DNA breaks in target DNA. Along with crRNA, type II systems require an accessory non-coding RNA called trans-activating crRNA (tracrRNA) for DNA cleavage (Jinek, Chylinski et al. 2012). Type V systems have signature Cas12 as the effector protein. RuvC domain of Cas12 causes staggered and sequence-specific DNA cleavage. According to different subtypes of type V, variation of the target (DNA or RNA) and guide RNA requirements have also been observed (Yan, Hunnewell et al. 2019; Harrington, Ma et al. 2020). The type VI signature Cas13 nuclease binds crRNA to target complementary RNA. Cas13 possesses HEPN domains that degrade the located RNA target (Shmakov, Abudayyeh et al. 2015).

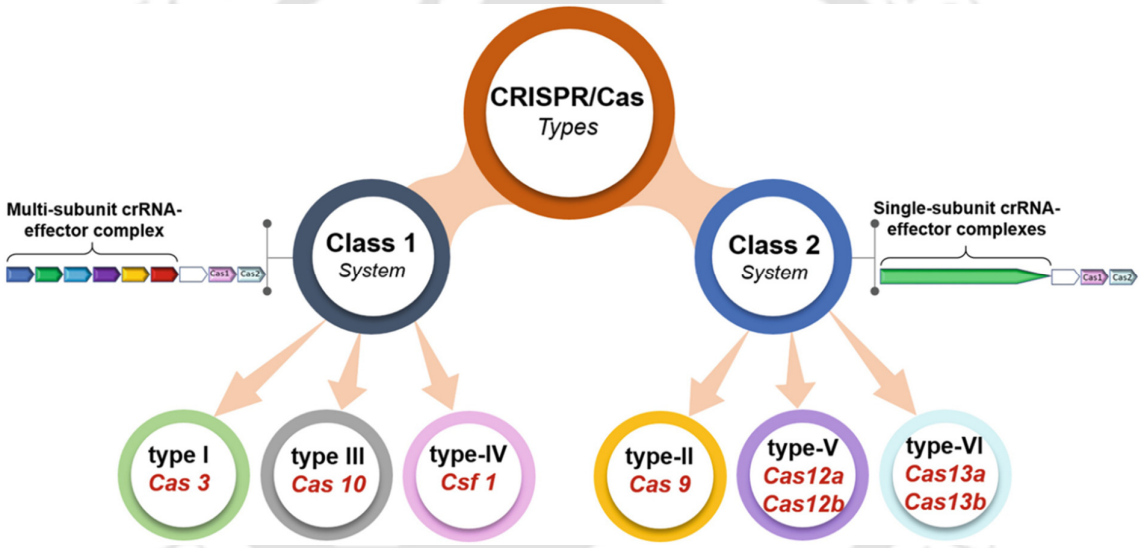


Figure 1.3. Classification of CRISPR-Cas systems [adapted from (Sahoo, Cuello et al. 2020)]. The schematic depicts two classes (class 1 and 2) and respective types (I-VI) of the CRISPR-Cas systems. The signature Cas proteins employed in the corresponding CRISPR-Cas type are narrated in the same circle.

Apart from the six main types of CRISPR-Cas system, a few CRISPR-Cas variants also exist, such as the Halobacterial and Asgard repeat-associated mysterious proteins (HRAMP and ARAMP) systems (Makarova, Karamycheva et al. 2019; Makarova, Wolf et al. 2020). The HRAMP system consists of highly diverged variants of Cas5 and Cas7 (two families of the RAMP superfamily) along with additional nucleases and uncharacterized conserved proteins. The HRAMP systems are not associated with CRISPR arrays and lack adaptation modules

(Makarova, Karamycheva et al. 2019). The ARAMP possesses diverged variants of Cas1, Cas5, Cas7 and additional nucleases (Makarova, Wolf et al. 2020).

1.3 CRISPR adaptation

During the adaptation phase, a stretch of captured DNA (pre-spacer) from MGE is directed to the CRISPR locus and integrated as a novel spacer with duplication of leader proximal repeat sequence (Wilkinson, Drabavicius et al. 2019). Essential and specific to the adaptation process, the two core *cas* genes encoding Cas1 and Cas2 proteins are relatively well conserved in most of the CRISPR-Cas systems (Brouns, Jore et al. 2008). Other common features of the adaptation process are the leader and the first repeat of the CRISPR array, implying that some functions are conserved in all CRISPR-Cas types (Makarova, Wolf et al. 2015; Amitai and Sorek 2016). According to the CRISPR-Cas type, in addition to Cas1 and Cas2, additional factors are required to generate spacers from processed pre-spacers. Cas2 is fused to a DnaQ domain in type I-E systems, which processes pre-spacers to generate suitable spacers for integration (Drabavicius, Sinkunas et al. 2018; Kim, Loeff et al. 2020). Similarly, in types I, II, and IV, Cas4 is utilized along Cas1 and Cas2 to generate appropriate spacers for integration (Lee, Zhou et al. 2018; Rollie, Graham et al. 2018). In other systems, the function of DnaQ and Cas4 can be replaced by non-CRISPR nucleases (Yoganand, Muralidharan et al. 2019).

1.4 CRISPR expression

Sequence-specific targeting of invading MGEs by crRNAs is the hallmark of the CRISPR-Cas defense system (Hille, Richter et al. 2018). Thus, the maturation of crRNAs from precursor CRISPR transcript is critical for the activity of the CRISPR-Cas system. The transcription start site of the pre-crRNAs lies within the leader sequence located upstream of the CRISPR locus. The pre-crRNA transcript is subsequently processed within the repeats to generate mature crRNAs composed of repeat segments and a spacer portion. In crRNAs, repeat segments are recognized by Cas proteins in a sequence- or structure-dependent manner and spacers are essential for target (complementary strand) binding (Hille, Richter et al. 2018). Although a common theme among the CRISPR-Cas types is the transcription of the pre-crRNA and processing, diversification of CRISPR-Cas systems in different subtypes and distinct Cas proteins led to mechanistic evolution in crRNA biogenesis (Charpentier, Richter et al. 2015).

1.4.1 Class 1 crRNA biogenesis

The family of cas6 genes encodes for RNA endonucleases responsible for processing pre-crRNAs in CRISPR-Cas types I and III systems (Carte, Wang et al. 2008). Types I and III systems exhibit remarkable similarities in the crRNA biogenesis, where Cas6 proteins process the repeat segments within the pre-crRNAs (Carte, Wang et al. 2008; Haurwitz, Jinek et al. 2010; Sashital, Jinek et al. 2011). A noteworthy exception is the type I-C system that lacks Cas6 homolog, and functionally, Cas5d (''d'' in Cas5d refers to ''Dvulg,'' the former name of Cas5 protein in type I-C) replaces Cas6 (Garside, Schellenberg et al. 2012; Nam, Haitjema et al. 2012). These nucleases cleave their cognate repeat RNA immediately downstream of the hairpin and yield mature crRNAs, each composing a full spacer unit flanked by a short repeat-derived 5′ handle (or 5′ tag) and a 3′ stem-loop of variable length (Haurwitz, Jinek et al. 2010; Gesner, Schellenberg et al. 2011; Sashital, Jinek et al. 2011; Nam, Haitjema et al. 2012) (**Figure 1.4**). The 5′ handle in crRNAs are usually 8 and 11 nts in length, depending on the system specific endoRNases (Cas6 homologs and Cas5d, respectively) utilized for crRNA biogenesis (Hochstrasser and Doudna 2015).

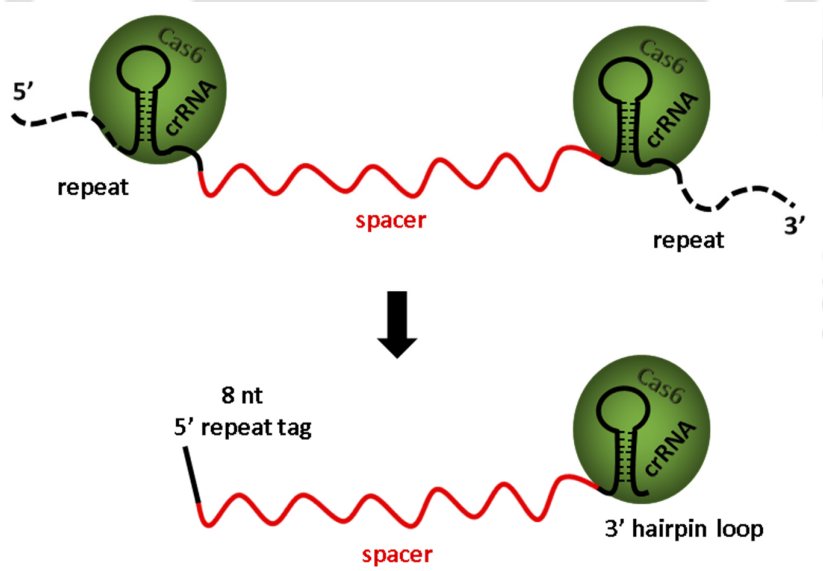


Figure 1.4. Processing of pre-crRNA by Cas6 [adapted from (Zheng, Li et al. 2020)]. The Cas6-mediated crRNA biogenesis from pre-crRNA is depicted in the schematic. Cas6 cleaves the cognate precursor within the repeat RNA segment, thus, resulting in the generation of mature crRNA that may remain bound with Cas6. Each mature crRNA contains a full spacer RNA segment flanked by repeat-derived 8 nt tag and hairpin-loop at the 5′, and 3′ ends. Repeat or repeat derived fragments, and spacer segment are demarcated by black and red colors. Green-filled circles represent Cas6 proteins.

Most of the Cas6 proteins remain associated with the crRNA after cleavage of repeat RNA segments within pre-crRNA and assemble into a ribonucleoprotein complex with other Cas proteins for CRISPR interference (Carte, Wang et al. 2008; Sashital, Jinek et al. 2011; Sternberg, Haurwitz et al. 2012; Niewoehner, Jinek et al. 2014). In contrast, all characterized homologs of Cas6a (subtype I-A) and some homologs of Cas6b (subtype I-B) dissociate from crRNA after the processing of cognate pre-crRNAs (Charpentier, Richter et al. 2015; Hochstrasser and Doudna 2015; Hille, Richter et al. 2018). In contrast to the other subtypes of type I, repeats of subtypes I-A and I-B CRISPR arrays are non-palindromic (Kunin, Sorek et al. 2007). In such cases, it was speculated that Cas6 exclusively recognizes the repeat sequence for cleavage. However, recent studies discovered that stem-loop structure in repeat RNA segments is important for undergoing cleavage and demonstrate that Cas6 remodels the repeats to form the stem-loop structure to position the requisite cleavage site (Shao and Li 2013; Shao, Richter et al. 2016; Sefcikova, Roth et al. 2017). Contrary to most monomeric Cas6 proteins, Cas6 proteins in systems with non-palindromic CRISPR repeat RNA (mainly I-A and some I-B) form dimers (Reeks, Sokolowski et al. 2013; Richter, Lange et al. 2013; Shao and Li 2013). Thus, it was suggested that the dimerization of Cas6 might be related to the remodeling function of Cas6, which is necessary for the cleavage of repeat RNA.

Repeat RNAs of type III CRISPR arrays form weakly stable stem-loop structures or are unstructured due to their non-palindromic nature (Kunin, Sorek et al. 2007). Thus, these repeats might also rely on Cas6 to remodel repeat RNA segments for canonical processing of pre-crRNAs (Hille, Richter et al. 2018). In support of this fact, Cas6 proteins from type III show high similarity to Cas6 homologs from subtypes I-A and I-B. In addition, the mechanism of crRNA biogenesis in type III is highly similar to that in subtypes I-A and I-B. In these systems, after the Cas6-mediated processing of pre-crRNAs, the processed crRNA undergoes further trimming at the 3′ end. Thus removal of the 3′ hairpin-loop from crRNAs leads to the dissociation of Cas6 from trimmed crRNAs (Carte, Wang et al. 2008; Carte, Pfister et al. 2010; Hatoum-Aslan, Maniv et al. 2011; Richter, Zoephel et al. 2012; Plagens, Tripp et al. 2014). Similar to subtype I-C, subtypes III-C and III-D lack Cas6 homologs; therefore, Cas5 proteins might be utilized for the processing of pre-crRNA. In the type IV system, the presence of *cas5* orthologs and *cas6*-like genes suggests a mechanism for crRNA biogenesis similar to other class 1 types.

Despite sharing limited sequence identity, Cas6 family members share common structural features necessary for crRNA binding. Cas6 proteins consist of N- and C-terminal repeat-associated mysterious protein (RAMP) domains separated by a cleft. These two RAMP domains form ferredoxin-like, or RNA recognition motif (RRM) folds (Haft, Selengut et al. 2005; Makarova, Haft et al. 2011; Reeks, Naismith et al. 2013). Residues from both RAMP domains can participate in catalysis. Typically, the active site is located in the cleft formed between the two RRM folds (Carte, Wang et al. 2008; Haurwitz, Jinek et al. 2010; Gesner, Schellenberg et al. 2011; Sashital, Jinek et al. 2011; Jackson, Golden et al. 2014; Mulepati, Héroux et al. 2014; Niewoehner, Jinek et al. 2014). The C-terminal RRM features a glycine-rich loop (G-loop) sequence motif that is conserved between Cas6 proteins (Makarova, Aravind et al. 2002; Makarova, Grishin et al. 2006). Generally, the G-loop follows the sequence pattern as GhGxxxxxGhG, where "h" denotes a hydrophobic residue and "xxxxx" contains at least one lysine or arginine (Haft, Selengut et al. 2005). The G-loop of Cas6 is crucial for its own folding and recognition/binding of crRNA (Wang, Preamplume et al. 2011; Wang, Zheng et al. 2012).

Cas6 nucleases cleave RNA exclusively. It is supported by an RNase A-like cleavage mechanism of Cas6 endoribonucleases where a general base extracts the proton from the 2′-hydroxyl group immediately upstream of the scissile phosphate, and a general acid donates a proton to the 5′ leaving oxygen associated with the scissile phosphate (Carte, Pfister et al. 2010; Haurwitz, Jinek et al. 2010; Haurwitz, Sternberg et al. 2012; Sternberg, Haurwitz et al. 2012). However, there is great variability in their active sites, and the catalytic residues are not always conserved, although catalytic histidines are common (Haurwitz, Jinek et al. 2010; Sternberg, Haurwitz et al. 2012; Brendel, Stoll et al. 2014; Niewoehner, Jinek et al. 2014). Active site arrangement in Cas6 enables it to cleave the cognate pre-crRNAs through a general acid-base mechanism in a metal-independent manner (Charpentier, Richter et al. 2015; Hochstrasser and Doudna 2015). Some Cas6 have high affinity towards cleaved RNA products resulting in a single-turnover cleavage reaction. However, other Cas6 may dissociate from cleaved RNA products due to a weaker binding affinity (Charpentier, Richter et al. 2015; Hochstrasser and Doudna 2015).

1.4.2 Class 2 crRNA biogenesis

Class 2 systems utilize the interference machinery or non-Cas proteins for crRNA biogenesis. Type II and subtype V-B systems require tracrRNA (trans CRISPR RNA) for CRISPR immunity (Deltcheva, Chylinski et al. 2011; Zhang, Heidrich et al. 2013; Shmakov, Abudayyeh

et al. 2015). For instance, in type II systems, binding of the effector protein Cas9 with tracrRNA-crRNA duplex further recruits the host RNase III for processing of repeat (Deltcheva, Chylinski et al. 2011). After additional processing by an unknown RNase, the effector complex comprised of tracrRNA-crRNA duplex with Cas9 participates in the interference process (Deltcheva, Chylinski et al. 2011; Jinek, Chylinski et al. 2012). In type V and VI systems, the effector proteins (Cas12 and Cas13, respectively) exhibit dual nuclease activity for crRNA biogenesis as well as target interference. Similar to type II, subtypes V-B and V-C require tracrRNA for crRNA biogenesis (Shmakov, Abudayyeh et al. 2015). Effector proteins Cas12a, Cas13a, and Cas13b (in subtypes V-A, VI-A, and VI-B, respectively) recognize and cleave the repeat RNA segment to generate crRNAs (East-Seletsky, O'Connell et al. 2016; Fonfara, Richter et al. 2016; Smargon, Cox et al. 2017).

1.5 Effector machinery assembly in type I

The type I-E Cascade (Cascade I-E) of *Escherichia coli* has been analyzed using the crystal structures of the effector complex to reveal molecular details of complex assembly (Brouns, Jore et al. 2008; Jackson, Golden et al. 2014; Zhao, Sheng et al. 2014). In *E. coli*, Cas6e endoribonuclease processes the pre-crRNA through cleavage within repeat RNA segments. After processing, Cas6e remains tightly bound to the 3′ repeat portion of crRNAs (Jore, Lundgren et al. 2011). Simultaneously, Cas7 proteins form the backbone filament by polymerizing along the spacer RNA segment of the crRNA, where the extreme Cas7 subunit at the 3′ end of the crRNA interacts with Cas6e (Jackson, Golden et al. 2014; Mulepati, Héroux et al. 2014). Concurrently, another Cascade subunit, Cas5e, caps the Cas7 filament at the 5′ end of the crRNA. In addition, Cas5e also interacts with the 5′ repeat portion of the crRNA by holding it's first 6 nt (Jackson, Golden et al. 2014). In the Cascade, Cas11 (also known as CasB or Cse2) dimer bridges the complex without interacting with crRNA. One Cas11 subunit and Cas5e contact Cas8 (also known as CasA or Cse1) at the bottom of the complex, whereas the other Cas11 subunit interacts with Cas6e (Jackson, Golden et al. 2014). In the Cascade, Cas8 and Cas11 were recognized as the largest and smallest subunits of the effector complex, respectively (Makarova, Aravind et al. 2011). This complex (approximately 405 kDa) with a stoichiometry of $Cas8_1$-$Cas11_2$-$Cas7_6$-$Cas5e_1$-$Cas6e_1$ resembles a seahorse-like architecture (Jore, Lundgren et al. 2011) **(Figure 1.5)**.

Comparison of the Cascade I-E with Cascade from other subtypes of type I revealed that they share remarkable architectural similarities. These common features include a core complex of Cas5, Cas7, and/or Cas6 with crRNA. In addition, small (Cas11) and large (Cas8) subunits are less tightly associated with the core complex, if present, depending on the subtypes of the type I system (Makarova, Aravind et al. 2011). These architectural similarities among Cascade complexes suggest a similar mechanism for complex assembly and target DNA interference (Zheng, Li et al. 2020).

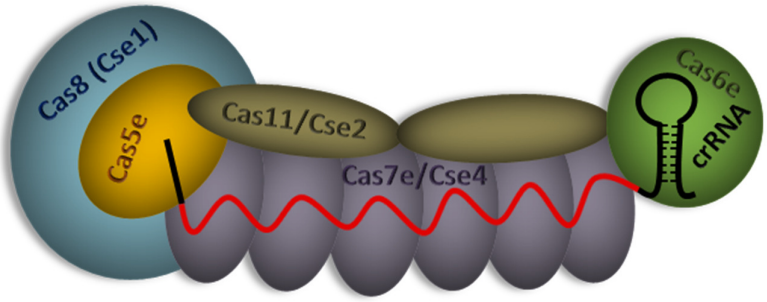


Figure 1.5. The architecture of *E.coli* Cascade I-E [adapted from (Zheng, Li et al. 2020)]. The architecture of *E. coli* Cascade I-E is depicted in the schematic. The Cas6e-bound crRNA generated in the expression stage of CRISPR-Cas acts as a scaffold for Cascade assembly. The Cascade resembles a seahorse-like architecture with a stoichiometry of Cas8/Cse1$_1$-Cas11/Cse2$_2$-Cas7/Cse4$_6$-Cas5e$_1$-Cas6e$_1$.

1.6 R-Loop formation and target DNA degradation

The Cascade recognizes the invading DNA through base pairing between the protospacer and spacer RNA segment within embedded crRNA. To avoid self-targeting at the CRISPR locus, PAM is essential for Cascade to establish a *bona fide* DNA target (Mojica, Díez-Villaseñor et al. 2009; Gudbergsdottir, Deng et al. 2011). In the *E. coli* subtype I-E system, Cas8 directly contacts with the PAM resulting in the specific binding of Cascade to the base-paired DNA segment located immediately adjacent to the PAM (Sashital, Wiedenheft et al. 2012). Type I CRISPR-Cas systems showed resilience to mismatches between the spacer of crRNAs and corresponding protospacers. Based on binding affinities between crRNA carrying mutations and DNA target, it was implied that the 1 to 8 nt sequence (also known as seed region) located immediately adjacent to the PAM within the crRNA plays an essential role in Cascade mediated recognition of DNA target (Wiedenheft, van Duijn et al. 2011). During Cascade binding, base pairing between the crRNA and target starts within the seed region and then

through the matching sequences (Hochstrasser, Taylor et al. 2014). This process results in the formation of an R-loop **(Figure 1.6)** in which the non-target (displaced) strand is bound and stabilized by the Cas11 dimer of Cascade (Hochstrasser, Taylor et al. 2014). R-loop formation causes conformational changes in the small and large subunits, leading to the recruitment of Cas3 for target degradation (Hochstrasser, Taylor et al. 2014). Once recruited by Cascade, Cas3 starts unwinding and degrading the DNA target. Generally, Cas3 comprises two fused enzymatic domains; an N-terminal HD (histidine-aspartate) nuclease domain and a C-terminal superfamily 2 helicase domain (Makarova, Grishin et al. 2006). In some type I systems, these two domains are encoded separately (domain fission) as helicase (Cas3′) and nuclease (Cas3″) (Makarova, Haft et al. 2011). The Cas3 HD domain exhibit the divalent ion-dependent nuclease activity on ss-DNA and/or RNA (Beloglazova, Petit et al. 2011; Mulepati and Bailey 2011; Sinkunas, Gasiunas et al. 2011), whereas the Cas3 helicase domain performs divalent ion- and ATP-dependent unwinding of DNA/DNA and/or DNA/RNA duplexes (Howard, Delmas et al. 2011; Sinkunas, Gasiunas et al. 2011).

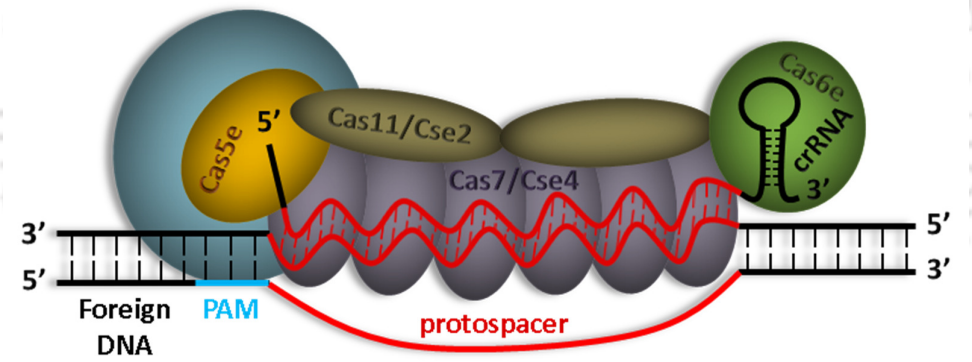


Figure 1.6. Cascade-mediated R-loop formation [adapted from (Zheng, Li et al. 2020)]. After recognizing PAM (demarcated by a blue-colored segment on DNA) on the invading DNA, crRNA in the Cascade base pairs with the target strand. Displacement of the non-target strand in the process leads to the formation of an R-loop.

The mechanism of Cas3-mediated DNA interference was uncovered by structural analyses of targeting complexes of *E. coli* (subtype I-E), where the formation of a complete R-loop structure was strictly required to recruit Cas3. Upon the formation of the complex containing Cas3 and Cascade-DNA, Cas3 nicked the non-target strand in the protospacer's 7-11 nt region. Simultaneously, the DNA target was degraded in the direction of 3′ to 5′ in the presence of ATP (Sinkunas, Gasiunas et al. 2013). It was suggested that ATP-dependent unwinding of the DNA target by the Cas3 helicase domain further provides ss-DNA region for the exonucleolytic

degradation by the Cas3 nuclease domain or other host nucleases, thus resulting in the degradation of the entire DNA target (Brouns, Jore et al. 2008; Mulepati and Bailey 2013).

1.7 Application of type I CRISPR-Cas systems

CRISPR immunity based on the type I system can be exploited for several practical applications in prokaryotic hosts. For instance, Cascade carrying a particular crRNA can be directed to a specific sequence for genome editing (using Cas3) or gene expression regulation (without Cas3). To date, genome editing remains the best-developed CRISPR application (Zheng, Li et al. 2020). In the eukaryotic cells, CRISPR-Cas mediated double-stranded breaks (DSBs) at specific locations of the genomic DNA could be repaired through either the cellular homologous recombination (HR) or the non-homologous end joining (NHEJ) repair pathways. Generally, the NEHJ repair system is error-prone, generating insertion/deletion at the target site, thus causing gene disruption or frameshift mutation. On the contrary, the HR repair system usually is error-free and precise. Most prokaryotes possess only the HR repair machinery. Nevertheless, the repair efficiency of CRISPR-caused DSBs in prokaryotes is relatively much lower, which is consistent with the CRISPR-mediated lethality observed in prokaryotes (Peng, Feng et al. 2015). However, via DNA-assisted co-expression of the HR repair system, endogenous type I mediated genome editing is now possible with high efficiencies (Li, Pan et al. 2016).

1.8 *Leptospira* and endogenous CRISPR-Cas

Leptospira is a spirochete bacteria that belongs to the family *Leptospiraceae*. *Leptospira* spp. are obligate aerobes, fastidious, motile, thin, and helically coiled with an approximate dimension of 10-20×0.15 μm (Cameron 2015). The hook-shaped ends of *Leptospira* give its distinctive question (interrogative) mark shape, as shown in a scanning electron micrograph of this bacterium (**Figure 1.7**). They can survive in organs and tissues of live or dead animals, in diluted milk, or a moist environment like swamps, soils, rivers, mud, and streams (Faine S 1999; Bharti, Nally et al. 2003; LeFebvre 2004; Adler and de la Peña Moctezuma 2010). Under laboratory conditions at 28-30°C, *Leptospira* grows slowly on solid or liquid media supplemented with vitamins B1 and B12, long-chain fatty acids, and ammonium salts (Faine S 1999; Murray, Rosenthal et al. 2009; Adler and de la Peña Moctezuma 2010). The medium Ellinghausen-McCullough/Johnson-Harris (EMJH), which contains bovine serum albumin, oleic acid, and Tween, is most commonly used for culturing leptospires. Leptospires are gram-

negative bacteria. However, their peptidoglycan chain is associated with the cytoplasmic membrane, which is exclusive to spirochetes. The *Leptospira* genome possesses two circular chromosomes with a cumulative length from 3.9 to 4.6 Mb (Picardeau, Bulach et al. 2008). Based on the structural diversity in the carbohydrate component of their lipopolysaccharide layer, leptospires were classified into 300+ serovars (Picardeau 2017). Initially, the genus *Leptospira* was divided into pathogenic and non-pathogenic species based on virulence. Phylogenetic analysis of *Leptospira* revealed its three lineages (pathogenic, intermediates, and saprophytes or non-pathogenic) that correlated with their pathogenicity level (Perolat, Chappel et al. 1998). To date, at least 35 *Leptospira* spp. have been identified, comprising 13 pathogenic, 11 intermediately pathogenic and 11 saprophytic *Leptospira* species (Vincent, Schiettekatte et al. 2019). The pathogenic species are responsible for causing disease, whereas the intermediate species can cause a very mild form of the disease in humans and animals (Bharti, Nally et al. 2003).

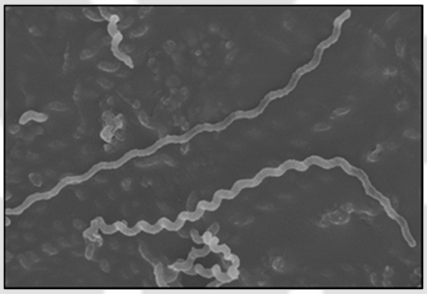


Figure 1.7. Scanning electron micrograph of *Leptospira interrogans*. The electron micrograph (20,000X magnification) shows the helically coiled shape of elongated *L. interrogans*. The typical hook-shaped ends of this bacterium are evident in the micrograph.

Infectious species of *Leptospira* causes Leptospirosis, a neglected tropical zoonotic disease that is a significant health problem worldwide (Bharti, Nally et al. 2003; Adler and de la Peña Moctezuma 2010; Karpagam and Ganesh 2020). Leptospirosis affects over 1 million people globally, with mortality of 60000 deaths yearly (Costa, Hagan et al. 2015). *Leptospira* is transmitted through direct contact with tissues and body fluids like the urine of infected animals or indirect contact with contaminated water (Bharti, Nally et al. 2003; Sharma and Yadav 2008). *Leptospira* colonize in the proximal renal tubules of carriers or infected hosts, resulting in the recurrent shedding of the bacteria through urine (Faine S 1999). Rodents like rats, mice, and moles are the primary hosts for leptospirosis. However, a broad range of other mammals, including rabbits, dogs, deer, cows, sheep, and certain marine mammals, are the secondary hosts that can carry and transmit the disease (Faine S 1999). Thus, occupations requiring

frequent contact with animals, such as veterinarians, farmers, abattoir workers, butchers, pet traders, rodent control workers and hunters, are more likely to contract Leptospirosis disease (Hartskeerl, Collares-Pereira et al. 2011; Musso and La Scola 2013). Humans are considered accidental hosts for Leptospirosis disease, which causes acute and occasionally fatal infections (Ko, Goarant et al. 2009).

Leptospira enters the body through incisions in the skin or mucous membranes of the nose, eyes, and throat. It takes 2-20 days for *Leptospira* to incubate inside the host. Depending on the species/serovar/strain of *Leptospira*, its inoculum size, and age/health of the hosts, the clinical symptoms of leptospirosis vary among individuals (Ko, Goarant et al. 2009; Adler and de la Peña Moctezuma 2010). In humans, most cases of leptospirosis are mild and self-limiting. The more severe cases of leptospirosis disease, known as Weil's syndrome, lead to multiorgan complications, including hepatic and renal dysfunction, jaundice, cardiovascular collapse, pulmonary hemorrhage and meningitis (Bharti, Nally et al. 2003). Other symptoms of leptospirosis include headache, nausea, fever, skin rashes, myalgia, and chills (Faine S 1999; Vinetz 2000). These symptoms mimic other illnesses like influenza, dengue, hepatic disease, and Hantavirus infections; therefore; Leptospirosis is often misdiagnosed in humans, leading to underestimation of the disease burden (Yang 2007; Victoriano, Smythe et al. 2009; Hartskeerl, Collares-Pereira et al. 2011).

Several studies have demonstrated that bacterial pathogenesis can be regulated by the endogenous CRISPR-cas system (Sampson, Saroj et al. 2013; Sampson and Weiss 2014). In a comparative genomics analysis of a group of *Leptospira* containing pathogenic, intermediate and non-pathogenic species, mainly pathogenic *Leptospira* spp. were identified to harbor CRISPR-Cas systems (Fouts, Matthias et al. 2016). Thus, the CRISPR-Cas systems have been hypothesized as one of the virulence factors in pathogenic *Leptospira* (Fouts, Matthias et al. 2016). Owing to the lack of efficient genetic manipulation tools, understanding *Leptospira* pathogenesis is still confined (Fernandes, Hornsby et al. 2021). So far, the homologous recombination technique is most frequently employed for targeted genetic knockout or random insertion (Croda, Figueira et al. 2008). In addition, transcription activator-like effectors (TALE) have also been used in both saprophytic and pathogenic strains for specific gene silencing. However, these techniques are costly and laborious (Pappas, Benaroudj et al. 2015). A new strategy was developed to manipulate *Leptospira* genetically, where episomal delivery of CRISPR-Cas9 (type II) using a shuttle vector (pMaOri) was achieved (Pappas, Benaroudj et

al. 2015). However, the CRISPR-Cas9 application is limited to a few bacteria because the double-strand breaks (DSBs) induced by Cas9 in the host DNA must be repaired for cell viability (Lieber 2010). In the *Leptospira* genome, owing to the absence of a NHEJ system, Cas9-induced DSBs were found to be lethal (Fernandes, Guaman et al. 2019). Thus, a CRISPR interference (CRISPRi) technique was used where an inactive (dead) variant of CRISPR-dCas9 was utilized for the genetic perturbation in *Leptospira* spp. (Shapiro, Chavez et al. 2018; Fernandes, Hornsby et al. 2021). However, CRISPRi technology is limited to gene silencing (Zhao, Shu et al. 2017). Recently, lethality due to CRISPR-Cas9-induced DSBs in *Leptospira* has been surpassed by the co-expression of the *Mycobacterium* NHEJ repair machinery (Fernandes and Nascimento 2022). However, the fastidious growth and tedious conjugation process of *Leptospira* usually lead to low efficiency of genetic manipulation (Fernandes, Hornsby et al. 2021). To overcome the obstacles of gene editing by CRISPR-Cas9 technology, harnessing endogenous CRISPR-Cas systems (type I and III) is an attractive strategy for genome editing (Li, Pan et al. 2016). The need for heterologous expression of potentially toxic proteins inside hosts can be surpassed while exploiting the endogenous CRISPR-based method (Maikova, Kreis et al. 2019). So far, endogenous CRISPR-Cas subtypes I-A, I-B, or III-B have been successfully utilized for genome editing in several archeae and *Clostridium* spp (Li, Pan et al. 2016; Pyne, Bruder et al. 2016; Cheng, Gong et al. 2017; Maikova, Kreis et al. 2019). Therefore, in the future, a comprehensive understanding of the *Leptospira* CRISPR-Cas system may enable us to produce a tool for the genetic manipulation of pathogenic *Leptospira* strains, which continues to be an arduous task.

Serovars of *L. interrogans* harbor I-B and I-C loci; however, only the I-B system was found to possess a CRISPR locus (Makarova, Haft et al. 2011; Xiao, Yi et al. 2019) **(Figure 1.8)**. Thus, it was hypothesized that the I-B system in *L. interrogans* might be adequate for CRISPR-mediated immunity. In contrast, subtype I-C probably carries out other functions in *L. interrogans* (Xiao, Yi et al. 2019). At the I-B loci of *L. interrogans*, CRISPR arrays were identified between *cas2* and *cas6* genes **(Figure 1.8)**. In other *Leptospira* species (*L. alstonii*, *L. kirschneri*, *L. weilii*, *L. santarosai*, and *L. broomii*), CRISPR arrays are positioned downstream of the entire *cas* gene cassette (Xiao, Yi et al. 2019).

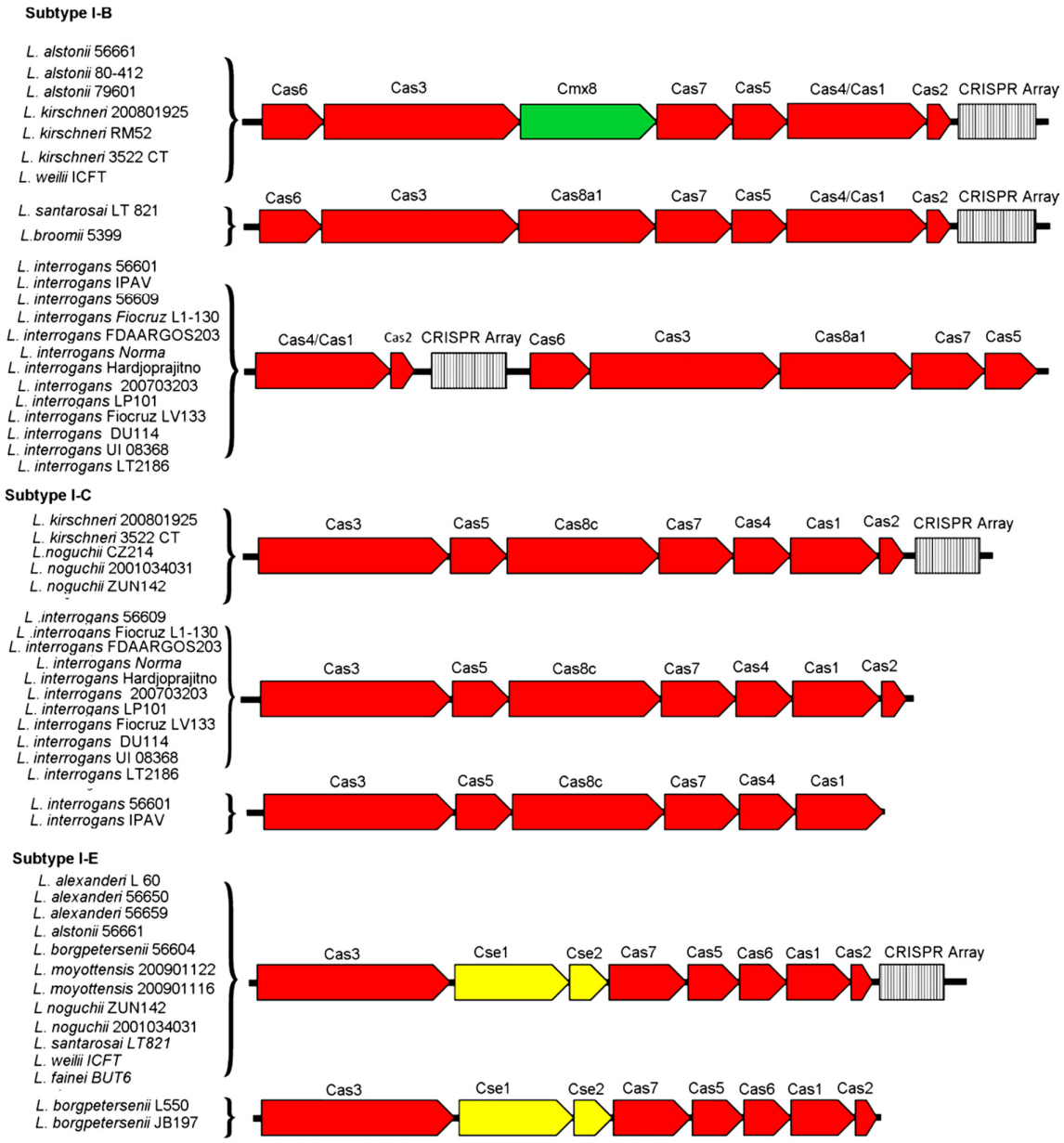


Figure 1.8. Organization of CRISPR-Cas subtypes identified in different strains of *Leptospira* [adapted from (Xiao, Yi et al. 2019)]. Cas genes are demarcated by red arrows. The CRISPR arrays are represented by black-and-white rectangular boxes. Cse and Cmx genes (Cas8 homologs) are indicated by yellow and green arrows, respectively. Strains are listed on the left side, with the corresponding CRISPR-Cas subtype on the right side.

In the genome of *L. interrogans* sv. Copenhageni strain Fiocruz L1-130, two subtypes (I-B and -C) of the type I system were predicted (Makarova, Haft et al. 2011). In subtype I-B locus of sv. Copenhageni, a CRISPR array was identified between the two independent *cas*-operons; *cas*-operon I (*cas4-cas1-cas2*) and *cas*-operon II (*cas6-cas3-cas8-cas7-cas5*) (Dixit, Ghosh et al. 2016). *In silico* analysis of CRISPR I-B loci among different serovars/strains of pathogenic

L. interrogans suggests that the two *cas* I-B operons of the subtype I-B locus span a region that contains either one or multiple CRISPR (Xiao, Yi et al. 2019) **(Figure 1.9)**. Thus, owing to the variable molecular length of the region flanked by these two *cas*-operons, CRISPR I-B loci in *L. interrogans* were termed hypervariable (Xiao, Yi et al. 2019). As highlighted above, previous studies only predicted the endogenous CRISPR arrays of *Leptospira*. However, there is a gap in understanding *Leptospira* CRISPR arrays transcription and the ability to undergo maturation or processing by the associated endoribonuclease.

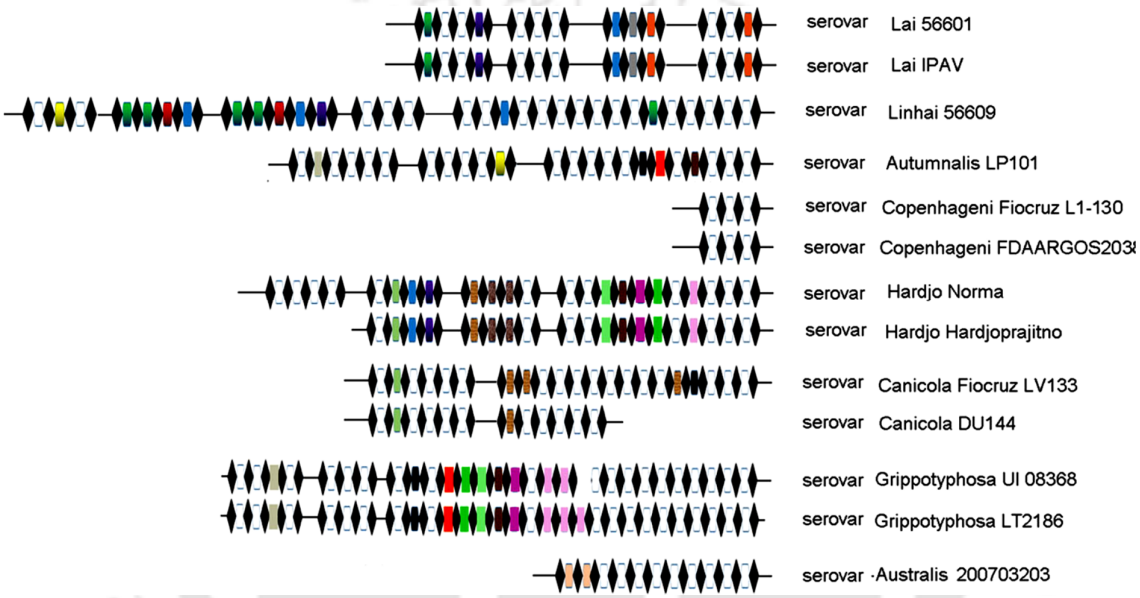


Figure 1.9. Analysis of the hypervariable region at the I-B loci in 13 strains of *L. interrogans* [adapted from (Xiao, Yi et al. 2019)]. The black-colored diamonds represent repeats, and the rectangles represent spacers. Horizontal lines represent sequences without CRISPR arrays. Colorized rectangles represent spacers that are shared within different serovars. Empty rectangles represent spacers that are unique or only shared within strains of the same serovar.

1.9 Rationale of the study

CRISPR-Cas is an RNA-mediated gene-editing technology, and due to its wide success, it is evolving rapidly. However, only a narrow range of bacterial species can overcome the DSB lethality of the nuclease utilized by expressing NHEJ machinery. Various pathogenic forms of *Leptospira* spp. that are notorious globally for causing leptospirosis disease in humans and animals lack the NHEJ machinery and thus limit the application of CRISPR-Cas. Therefore, in an alternative approach, the application of the endogenous CRISPR-Cas system is evolving. Repurposing the host's CRISPR-Cas system for genome editing depends on understanding the

RNA-mediated immunity process. Therefore, information regarding CRISPR arrays, such as CRISPR array orientation, repeat-spacer boundaries, and the PAM sequence, is required to reprogram the CRISPR-Cas systems in *Leptospira* for genetic manipulation. Understanding the CRISPR-Cas system of pathogenic *Leptospira* will help create a genetic tool to analyze gene function and pathogenesis. This thesis aimed to comprehend the knowledge of CRISPR array transcription biology in *Leptospira*. We first investigated the CRISPR array transcription in *Leptospira* and deciphered its orientation through reverse transcription-polymerase chain reaction (RT-PCR). In addition, the recombinant LinCas6 (rLinCas6) was employed to understand the processing of the precursor CRISPR I-B transcripts of *Leptospira*. To study the Cascade assembly, we analyzed the interaction of rLinCas6 and/or rLinCas5 to the crRNA under *in vitro* conditions. Furthermore, to understand the physiological requirements for the nuclease activities of the signature protein of *Leptospira* CRISPR-Cas, biochemical characterization of rLinCas3 was performed. To highlight the rationale of the proposed work, a schematic containing the key points is shown in **Figure 1.10**.

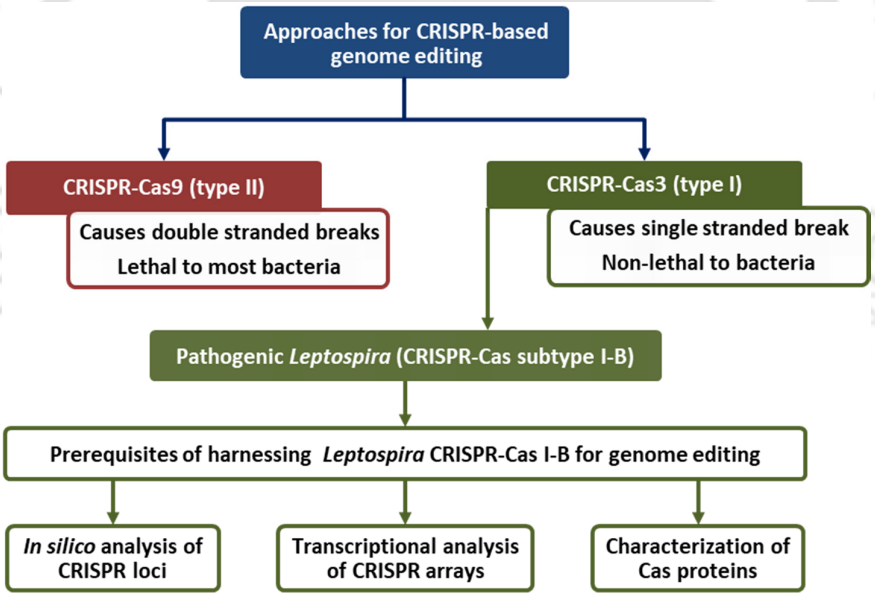


Figure 1.10. Schematic representation of the rationale of the study. A schematic diagram showing the approaches for CRISPR-based genome editing is presented to highlight the motive of the proposed work. Type I CRSIPR-Cas3 induces a non-lethal single-stranded break in bacteria rather than a lethal double-stranded break in type II CRISPR-Cas9. The prerequisites of repurposing the endogenous CRISPR-Cas I-B of *Leptospira* for genome editing are also stated.

CHAPTER 2

In silico analysis and identification of CRISPR array transcripts in *L. interrogans* serovar Copenhageni and Lai

Previous studies have identified and defined the architecture of CRISPR-Cas I-B in *L. interrogans* serovar (sv.) Copenhageni strain Fiocruz L1-130 (Makarova, Haft et al. 2011; Makarova, Wolf et al. 2015; Dixit, Ghosh et al. 2016). Moreover, using the program CRISPRFinder, CRISPR loci were predicted at the I-B of *L. interrogans* serovars (Dixit, Ghosh et al. 2016; Xiao, Yi et al. 2019). However, there was a gap in understanding *Leptospira* CRISPR array transcription and its orientation. In this chapter, CRISPR arrays in the genomes of two reference serovars of *L. interrogans* (svs. Copenhageni and Lai) were identified using the updated version of the program and compared using *in silico* approach. In addition, transcriptional analysis of CRISPR arrays positioned at the I-B locus of svs. Copenhageni and Lai were performed.

2.1 Materials and Methods

2.1.1 Bioinformatics analysis

Information of predicted CRISPR arrays in the genomes of *L. interrogans* svs. Copenhageni and Lai were retrieved via a taxonomic-based search in the CRISPRCasdb database, a repository built with CRISPRCasFinder (Pourcel, Touchon et al. 2020). Sequence alignments of multiple repeats and spacers sequences were carried out using Clustal Omega (Madeira, Park et al. 2019). Nucleotide sequences of inter-array regions at the hypervariable region of sv. Lai was extracted from the NCBI nucleotide database, and MSA was performed using the MUSCLE program (Madeira, Park et al. 2019). The ESPript program (version 3.0) was used to create visual representations of the aligned sequences (Robert and Gouet 2014). The WebLogo tool (Crooks, Hon et al. 2004) was used to construct a logo of nucleotide flanking protospacers that were retrieved via the CRISPRTarget (Biswas, Gagnon et al. 2013) tool. The OligoPerfect primer designing tool of Thermo Fisher Scientific was used to degisn the primers with or without restriction sites.

2.1.2 Bacterial strains and nucleic acid isolation

Leptospira strains (*L. interrogans* sv. Copenhageni str. Fiocruz L1-130 and *L. interrogans* sv. Lai str. 56601) were obtained from the Indian Council of Medical Research (ICMR), Regional Medical Research Centre (RMRC), Port Blair, Andaman and Nicobar Island, India. Leptospires were maintained in EMJH (Ellinghausen-McCullough-Johnson-Harris) medium (Difco) at 29°C, as described previously (Ghosh, Prakash et al. 2018; Ghosh, Prakash et al. 2018). A 7-day-old culture of *Leptospira* (500 µl) was inoculated in 9.5 ml of EMJH medium with 100 µg/ml of 5-Fluorouracil (HiMedia) and incubated for 7 days at 29°C for nucleic acids isolation.

A 7-day-old culture (10 ml) of *Leptospira* containing approximately 10^9 cells was used for genomic DNA isolation using the bacterial genomic DNA purification kit (HiMedia), as per manufacturer's instructions. Briefly, *Leptospira* culture was harvested by centrifugation, and cell lysis was performed by Proteinase K digestion and alkaline lysis. After cell lysis, the supernatant was passed through the spin column containing a fixed silica-gel membrane that binds to DNA specifically. Then the membrane was washed to remove protein contaminants and salts. Finally, the pure DNA was eluted in the elution buffer provided with the kit.

The total RNA of a 7-day-old *Leptospira* culture (10 ml, ~10^9 cells) was isolated using Trizol reagent (Invitrogen), which is an acidic solution of guanidinium thiocyanate (GITC), phenol and chloroform. According to the manufacturer's instructions, the nucleic acids and proteins were extracted with Trizol. Following centrifugation, total RNA from the upper aqueous phase was recovered after precipitation in isopropanol. Precipitated RNA was resuspended in soluble form in nuclease-free water (HiMedia). The quality and quantity of isolated total RNA were determined using agarose gel electrophoresis and spectrophotometry, respectively.

2.1.3 Reverse Transcription-Polymerase Chain Reaction (RT-PCR) and quantitative real-time PCR (q-PCR)

In the first step of the two-step RT-PCR process, the complementary first strand (cDNA) synthesis was performed with total RNA (1 µg; DNase I treated) of *Leptospira* using Verso cDNA synthesis kit (Thermo Fisher Scientific), as per manufacturer's protocol. In the first step of RT-PCR, reverse transcriptase, dNTPs, reaction buffer, and random hexamers (or specific primer) were used to generate the cDNA. In the second step of RT-PCR, this cDNA was used as a template in PCR with CRISPR spacer-specific primer pairs and PCR reagents, including Taq DNA polymerase (New England Biolabs). Additionally, a first strand synthesis reaction

was set up without reverse transcriptase that gave a "no enzyme control (NEC)" template to use in the second step of RT-PCR with the same primer pairs. The purpose of this negative control was to rule out genomic DNA contamination in the RNA that was used for cDNA synthesis. Reaction conditions in RT-PCR were according to the manufacturer's (Thermo Fisher Scientific) instructions. Each RT-PCR experiment was repeated twice to verify the reproducibility of the results.

According to established laboratory protocol (Ghosh, Prakash et al. 2018), q-PCR analysis of CRISPR arrays was performed. Briefly, a q-PCR reaction was set up that contains a 5-fold diluted cDNA template and CRISPR spacer-specific primer pairs with SYBR Green master mix (Thermo Fisher Scientific). The real-time system (BioRad) was programmed for 2 min at 50°C, 10 min at 95°C, and 40 cycles for 15 sec at 95°C and 1 min at 60°C, followed by melt curve analyses of the qPCR products. The CRISPR array transcripts were normalized with transcripts of *Leptospira 16S rRNA* (*rrs1*) using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001). The CRISPR transcripts were estimated per 10^6 copies of the *16S rRNA* of respective *Leptospira* serovars. For statistical analysis, two independent experiments, each in quadruplets, were performed.

2.2 Results and Discussion

2.2.1 Computational and transcriptional analysis of CRISPR array in serovar Copenhageni

For the improvised prediction of the CRISPR array, including their reliability and direction in *Leptospira*, the CRISPRCasdb database was analyzed. For each predicted CRISPR array, the database CRISPRCasdb project its CRISPR Id, coordinate, spacer contents, repeat consensus nucleotide sequence, transcription direction, and evidence level. A repeat consensus sequence for any predicted CRISPR array is based on the occurrence of a nucleotide at each position in corresponding repeats. Evidence level (rating 1-4) provided by the database, based on the similarity between repeats and spacer of an array, shows the reliability of that predicted CRISPR array. In sv. Copenhageni, the database CRISPRCasdb projected 11 CRISPR arrays (**Table 2.1**). The naming of these CRISPR arrays (LIC_Cr$^{1\text{-}11}$) was done based on their order (1 to 11) provided by the CRISPRCasdb.

At the I-B loci of *L. interrogans,* the CRISPR array is flanked by the two independent operons (*cas*-operon I and II) (Dixit, Ghosh et al. 2016). In this study, the inter-operonic region at the I-B locus is called the hypervariable region **(Figure 2.1)** that harbors a variable number of CRISPR arrays depending on the serovars of *L. interrogans* (Xiao, Yi et al. 2019). Out of the 11 CRISPR arrays (LIC_Cr$^{1-11}$) identified in sv. Copenhageni, a single CRISPR array (LIC_Cr^{2}), was positioned in the hypervariable region of the I-B locus **(Figure 2.1).** Other CRISPR arrays (LIC_Cr^{1} and LIC_Cr$^{3-11}$) identified in sv. Copenhageni were outside the I-B locus without *cas* genes in their vicinity. According to the database CRISPRCasdb, the array LIC_Cr^{2} contains four repeats (36 nt each) interspaced by three different spacers (34-37 nt) (**Table 2.2**), which agrees with a previous report (Dixit, Ghosh et al. 2016) using CRISPRFinder program. In this study, the naming of LIC_Cr^{2}-associated repeats (R1 to R4) and spacers (S1 to S3) were done in the direction of *cas*-operons.

Table 2.1. CRISPRCasdb analysis in the genome of sv. Copenhageni

CRISPR Id	Coordinate Start…end	Sn	Repeat consensus sequence (5′-3′)	D^n	EL
AE016823.1_1 (LIC_Cr^{1})	442,732…442,824	1	AACGCTCTTTATGAATCGCGTTG	ND	1
AE016823.1_2 (LIC_Cr^{2})	1,133,848…1,134,101	3	GTGCTCAACGCCTAACGGCATCAAAGTTATATTCAG	ND	4
AE016823.1_3 (LIC_Cr^{3})	1,451,041…1,451,345	4	TTCCTAAAGAAATAGGGAATTTAAAAAA	+	2
AE016823.1_4 (LIC_Cr^{4})	1,451,455…1,451,551	1	TTCCTAAAGAAATAGGGAATTTAAAAAA	+	1
AE016823.1_5 (LIC_Cr^{5})	1,615,099…1,615,192	1	AGGAAAGCGTTGTGTTGAGTTTT	ND	1
AE016823.1_6 (LIC_Cr^{6})	1,691,119…1,691,219	1	GAGTTCCCACAATTTACACGAGATC	ND	1
AE016823.1_7 (LIC_Cr^{7})	2,011,715…2,011,805	1	TAGGAAGTGATGCATTGAGTTCA	ND	1
AE016823.1_8 (LIC_Cr^{8})	2,797,856…2,797,947	1	ATTACGTCTCTTTGTAACTCACACTTGTT	–	1
AE016823.1_9 (LIC_Cr^{9})	3,086,747…3,086,849	1	TTTTTTAATTTCATATCAAAATCTTGATCGTTTTAAGAG	ND	1
AE016823.1_10 (LIC_Cr^{10})	3,391,097…3,391,198	1	CTGACAAATTCTAAGTTGTAAGAGTTCCCACA	+	1
AE016823.1_11 (LIC_Cr^{11})	3,991,732…3,991,827	1	AGGAAAGCGTTGTGTTGAGTTTTCC	ND	1

"+" and "−" indicates 5' to 3' direction on the sequenced and complementary strand, respectively. "ND" and "EL" stands for "not defined" and "evidence level," respectively. Sn and D^n correspond to the predicted number of spacers and direction of the respective CRISPR arrays.

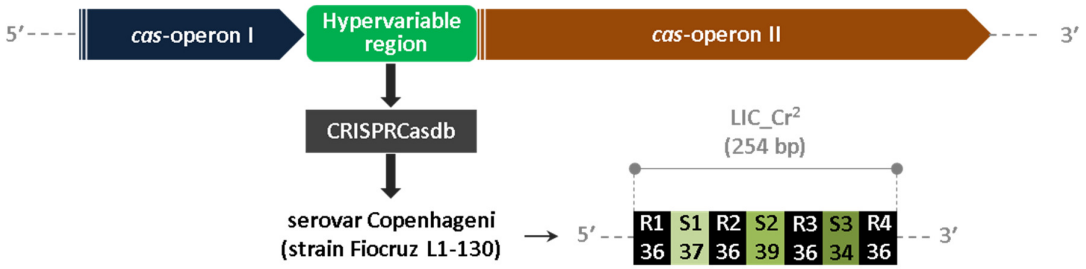


Figure 2.1. *In silico* prediction of CRISPR array at the I-B locus of serovar Copenhageni. Genomic architecture of the CRISPR-Cas I-B system of *L. interrogans*. The CRISPR-Cas I-B loci of *L. interrogans* harbor a variable region between the two *cas*-operons; operon I and II (adapted from (Dixit, Ghosh et al. 2016)). At the I-B locus of sv. Copenhageni, the database CRISPRCasdb identified a single CRISPR array (LIC_Cr^2; 254 bp) containing four repeats (R1-4) interspaced by three spacers (S1-3). The *cas* operons (blue and orange) and hypervariable region (green) are represented over a single strand in the 5' to 3' direction (left to right). The repeats (Rn; 36 bp each) and the spacers (Sn; 34-39 bp) of the CRISPR array are presented by black-filled and unique color-filled solid rectangles.

Table 2.2. Details of LIC_Cr^2-associated repeats and spacers provided by the database CRISPRCasdb

LIC_Cr^2	Coordinate start…end	Sequence (5′-3′)	Length (bp)
R1	1134101…1134066	CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC	36
S1	1134065…1134029	AAAGGATCCTTTGATCAAAAGAATTCGTCCTTGATTT	37
R2	1134028…1133993	CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC	36
S2	1133992…1133954	GGACATAGGACCAAACTTCCCATAATGTATCGTTATGGG	39
R3	1133953…1133918	CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC	36
S3	1133917…1133884	AAGGGGAAAACATTCGTCACCCCGTGAAAAACTT	34
R4	1133883…1133848	CTGAATATAACTTTGATGCCGATAGGCGGTTTAGAA	36

Since the direction of LIC_Cr^2 could not be predicted using the CRISPRCasdb (Table 2.1), its transcriptional analysis was performed using RT-PCR. For this purpose, CRISPR spacer-specific forward and reverse primers ($^{C2}S1_f$ and $^{C2}S3_r$) were designed with reference to the strand of *cas* genes **(Figure 2.2A and Table S2.1)**. This primer pair could amplify a specific 182 bp fragment (LIC_Cr^2 S1S3) in PCR using gDNA of sv. Copenhageni **(Figure 2.2B)**. To identify the transcript of LIC_Cr^2, in the first step of the RT-PCR, cDNA was synthesized from

the total RNA of sv. Copenhageni using random hexamers (RH). To rule out the gDNA contamination in isolated total RNA of sv. Copenhageni, a cDNA synthesis reaction was set without the reverse transcriptase under similar reaction conditions. In the second step of RT-PCR, the cDNA was used as a template in PCR with $^{C2}S1_f$ and $^{C2}S3_r$ primer pair. Agarose gel electrophoresis of RT-PCR product showed amplification of a 182 bp fragment using cDNA and no amplification in the negative control **(Figure 2.2A)**. This result suggested that the CRISPR array (LIC_Cr^2) is transcriptionally active in sv. Copenhageni.

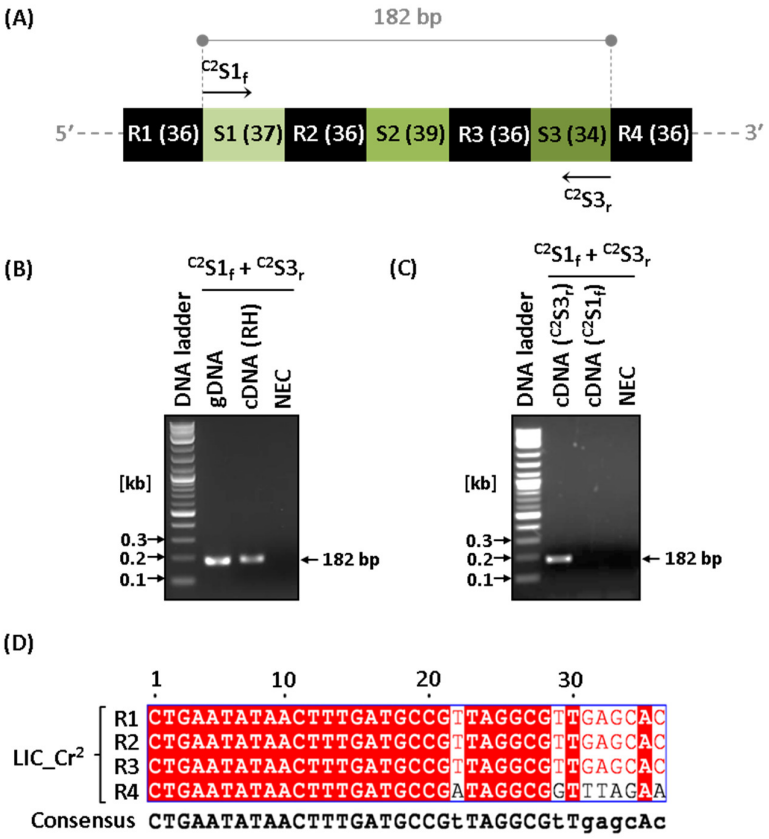


Figure 2.2. Characterization of the CRISPR I-B array of *L. interrogans* serovar Copenhageni. (A) Schematic representation of primer pairs used in the study. Spacer-specific primers used in the RT-PCR experiment are denoted by arrows on the architecture of the array LIC_Cr^2. The length of the DNA fragment expected in PCR is indicated by vertical dashed lines (grey) with the number (in bp) given over the double arrowhead. (B) Identification of CRISPR (LIC_Cr^2) array transcription. The cDNA synthesized using random hexamer (RH) served as a template in the PCR with the set of primers ($^{C2}S1_f$ and $^{C2}S3_r$) from the first and terminal spacers (arrow marked) region. An amplicon of 182 bp on 2% agarose gel was detected. The gDNA served as a positive control, whereas cDNA synthesis without reverse transcriptase served as no enzyme control (NEC) of the PCR reaction. (C) Analysis of the LIC_Cr^2 transcript orientation by RT-PCR. A two-step reverse transcriptase PCR reaction (RT-PCR) was set up using the total RNA of sv. Copenhageni. The cDNA synthesized using a single forward primer of the first spacer ($^{C2}S1_f$) or a single reverse primer of the terminal spacer ($^{C2}S3_r$) used in PCR with the primer set ($^{C2}S1_f$ and $^{C2}S3_r$). A

PCR amplicon of 182 bp was detected, with only one of two cDNAs synthesized as a template. (D) Nucleotide alignment of repeat sequences. CRISPR array (LIC_Cr^{2}) repeat DNA segments were aligned for consensus sequence in the proposed direction of transcription. Upper and lower case codes of repeat consensus denote the conserved (red-filled) and semi-conserved nucleotides in the listed repeat sequences. The sequence logo of the repeats indicates the occurrence of a particular nucleotide in CRISPR I-B repeats.

After deciphering the LIC_Cr^{2} transcription in sv. Copenhageni, the transcript orientation of LIC_Cr^{2} was elucidated using an independent RT-PCR experiment. Two separate reactions for reverse transcription were set using spacer-specific forward or reverse primer ($^{C2}S1_f$ or $^{C2}S3_r$). An amplicon of the expected size (182 bp) was evident in PCR using the $^{C2}S1_f$ and $^{C2}S3_r$ primer pair when cDNA used as a template was reverse transcribed with the reverse primer ($^{C2}S3_r$) **(Figure 2.2C)**. Moreover, the cDNA could not be synthesized using the forward primer (C2S1f), evident by no PCR amplification. This result suggested that the CRISPR array and the *cas* genes in the CRISPR-Cas I-B of *Leptospira* are transcribed co-directionally. In the direction of CRISPR-Cas I-B, sequence alignment of LIC_Cr^{2} repeats revealed variation within its terminal repeat, primarily towards the 3′ end **(Figure 2.2D)**. This type of variation in terminal repeats has also been reported previously in the CRISPR array of *Streptococcus thermophilus* (Horvath, Romero et al. 2008) and *E. coli* (Touchon and Rocha 2010).

2.2.2 *In silico* analysis of CRISPR arrays in serovar Lai

A previous *in silico* analysis in the sv. Lai genome using the program CRISPRFinder has documented 4 arrays with identical repeat consensus (28 nt) (Xiao, Yi et al. 2019). Thus, unlike a single CRISPR array at the I-B locus in sv. Copenhageni, sv. Lai contains multiple CRISPR arrays at the locus. In this study, we used a database CRISPRCasdb that employs an upgraded version of the program CRISPRFinder (CRISPRCasFinder) to predict *cas* and CRISPR loci in the prokaryotic genomes (Pourcel, Touchon et al. 2020). The CRISPRCadb projected a total of 14 CRISPR arrays (LA_Cr$^{1\text{-}14}$) in sv. Lai **(Table 2.3)**. Out of these 14 CRISPR arrays, 7 arrays (LA_Cr$^{6\text{-}12}$) were positioned in the hypervariable region (I-B locus) of the sv. Lai genome **(Figure 2.3 and Table 2.3)**. Thus, compared to the previous study (Xiao, Yi et al. 2019), three extra arrays (LA_Cr^{8}, Cr^{11,} and Cr^{12}) were identified in the hypervariable region of sv. Lai. Other 7 CRISPR arrays in sv. Lai (LA_Cr$^{1\text{-}5}$ and LA_Cr$^{13\text{-}14}$) were positioned outside the hypervariable region without *cas* genes in their vicinity.

Table 2.3. CRISPRCasdb analysis in the genome of sv. Lai

CRISPR Id	Coordinate Start…end	Sn	Repeat consensus sequence (5′-3′)	D^n	EL
AE010300.2_1 (LA_Cr[1])	1,347,143…1,347,585	6	TTCCTAAAGAAATCGGAAAACTAC	+	2
AE010300.2_2 (LA_Cr[2])	2,269,996…2,270,086	1	TGAACTCAATGCATCACTTCCTA	ND	1
AE010300.2_3 (LA_Cr[3])	2,415,309…2,415,408	1	TTAGTGGTAGTTCCTACATTTTAG	ND	1
AE010300.2_4 (LA_Cr[4])	2,677,849…2,677,942	1	AAAACTCAACACAACGCTTTCCT	ND	1
AE010300.2_5 (LA_Cr[5])	3,145,771…3,145,866	1	AGGAAAGCGTTGTGTTGAGTTTTCC	ND	1
AE010300.2_6 (LA_Cr[6])	3,163,254…3,163,495	3	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_7 (LA_Cr[7])	3,163,731…3,163,973	3	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_8 (LA_Cr[8])	3,164,138…3,164,239	1	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_9 (LA_Cr[9])	3,164,476…3,164,719	3	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_10 (LA_Cr[10])	3,164,839…3,165,154	4	CTGAATATAACTTTGATGCCGTTAGGCG	–	3
AE010300.2_11 (LA_Cr[11])	3,165,387…3,165,487	1	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_12 (LA_Cr[12])	3,165,651…3,165,821	2	CTGAATATAACTTTGATGCCGTTAGGCG	–	4
AE010300.2_13 (LA_Cr[13])	3,487,606…3,487,730	1	GATCTTGAGATAACACTTGTGGGTTGGTTATGG	ND	1
AE010300.2_14 (LA_Cr[14])	4,059,671…4,059,766	1	AGGAAAGCGTTGTGTTGAGTTTTCC	ND	1

"+" and "–" indicates 5' to 3' direction on the sequenced and complementary strand, respectively. "ND" and "EL" stands for "not defined" and "evidence level," respectively. Sn and D^n correspond to the predicted number of spacers and direction of the respective CRISPR arrays.

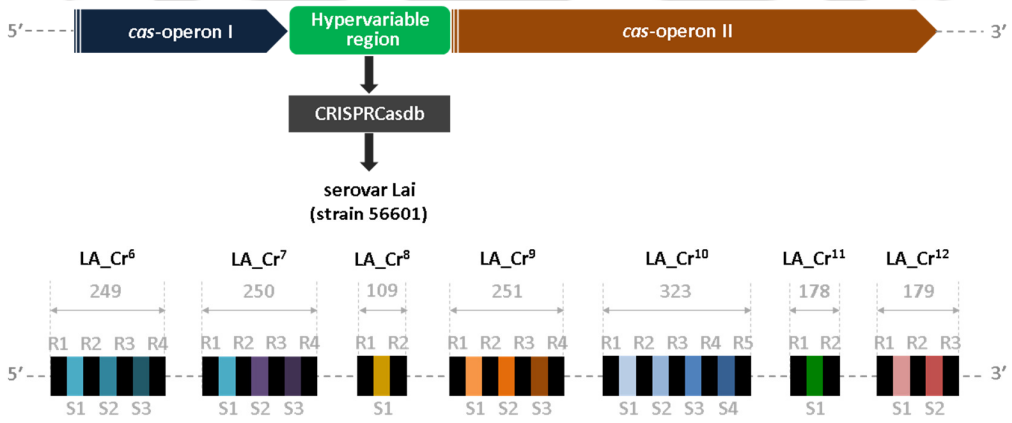


Figure 2.3. Identification of CRISPR arrays at the I-B locus of *L. interrogans* serovar Lai. CRISPRCasdb analysis in the genome of sv. Lai revealed 7 CRISPR arrays (LA_Cr[6-12]) in the hypervariable region at the I-B locus. The *cas* operons and CRISPR arrays are represented over a single strand in the 5' to 3' direction (left to right). The repeats (Rn) and the spacers (Sn) of each CRISPR array

are presented by black-filled and unique color-filled solid rectangles. The length of each CRISPR array is indicated in bp over double arrowheads.

According to the CRISPRCasdb output, each repeat of the seven CRISPR arrays (LA_Cr$^{6\text{-}12}$) was 28 nt long. In addition, the repeat consensus sequences of these arrays (LA_Cr$^{6\text{-}12}$) were 100% identical **(Table 2.3)**. Repeat consensus of CRISPR I-B arrays (LIC_Cr^{2} and LA_Cr$^{6\text{-}12}$) measured 36 nt and 28 nt in svs. Copenhageni **(Table 2.1)** and Lai **(Table 2.3)**, respectively. Interestingly, alignment of the repeat consensus of LIC_Cr^{2} (36 nt) and LA_Cr$^{6\text{-}12}$ (28 nt) revealed a deficit of 8 nt at the 3′ end in the repeat consensus of LA_Cr$^{6\text{-}12}$ **(Figure 2.4A)**. This 8 nt deficit at the 3′ end was consistent in database-defined repeats (n=28) of the seven arrays (LA_Cr$^{6\text{-}12}$) **(Table S2.2)**. Compared to the 36 nt long repeats of LIC_Cr^{2}, 28 nt long repeats in seven arrays LA_Cr$^{6\text{-}12}$ prompted us to compare the spacer sequences in LA_Cr$^{6\text{-}12}$ of sv. Lai. Multiple (n=17) spacer sequences **(Table S2.2)** of seven arrays LA_Cr$^{6\text{-}12}$ retrieved from the CRISPRCasdb were aligned. In MSA, 8 nt at the 5′ ends spacer sequences were conserved (5′-TTGAGCAC-3′) **(Figure 2.4B)**. According to the rule of thumb, similar repeats of a CRISPR array are separated by unique spacers (Mohamadi, Bostanabad et al. 2020). Hence, to follow this criterion, the conserved 8 nt sequences in database-defined spacers of seven arrays LA_Cr$^{6\text{-}12}$ have been included as a part of adjacent repeats. Such manual redefining of the repeats-spacers composition in the seven arrays (LA_Cr$^{6\text{-}12}$) resulted in the increase of repeats size (28 nt; database-defined) to 36 nt. Moreover, to maintain the repeat consistency (36 nt), the terminal repeats of each CRISPR array (LA_Cr$^{6\text{-}12}$) were extended by eight nucleotides at the 3′ end.

In this study, RT-PCR analysis of array LIC_Cr2 of sv. Copenhageni proposed a co-directional arrangement of CRISPR array and *cas* genes at the I-B locus *L. interrogans*. However, CRISPRCasdb projected the direction of each of the seven arrays LA_Cr$^{6\text{-}12}$ opposite to that of *cas* genes **(Table 2.3)**. Thus, the CRISPRCasdb erred in projecting a correct orientation CRISPR arrays at the I-B locus of sv. Lai.

To address the variation in repeat sequences of *Leptospira* CRISPR I-B arrays, multiple repeats of 8 arrays [LIC_Cr^{2} (n=4) and LA_Cr$^{6\text{-}12}$ (n=24)] were aligned. MSA of these 28 repeats generated a repeat consensus (RC) of 36 nucleotides **(Figure 2.4C).** Out of 28 repeats used in MSA, 12 repeats were identical to RC. Compared to the RC, 16 repeats possess sequence variation at one or multiple positions. These 16 repeats belonged to 8 variants of repeat [RVs (i-viii)]. Among these 8 RVs, one (i) and seven (ii-viii) RVs were identified in CRISPR I-B

arrays of svs. Copenhageni and Lai, respectively **(Figure 2.4C)**. Three (v, vi, and viii) out of eight (i-viii) RVs had a single nucleotide polymorphism **(Figure 2.4C)**. In contrast, five RVs (i-iv and vii) showed variations of 5-9 nt. This kind of sequence variation was observed either at first or terminal repeats of arrays. Thus, separate MSAs of first and terminal repeats were conducted. In MSAs, variations in nucleotide sequences were observed primarily at 5′ ends of first repeats **(Figure 2.4D)** and 3′ ends of terminal repeats **(Figure 2.4E)**. On the contrary, the MSA of repeat sequences that possessed spacers at either end showed the least sequence variation **(Figure 2.4F)**.

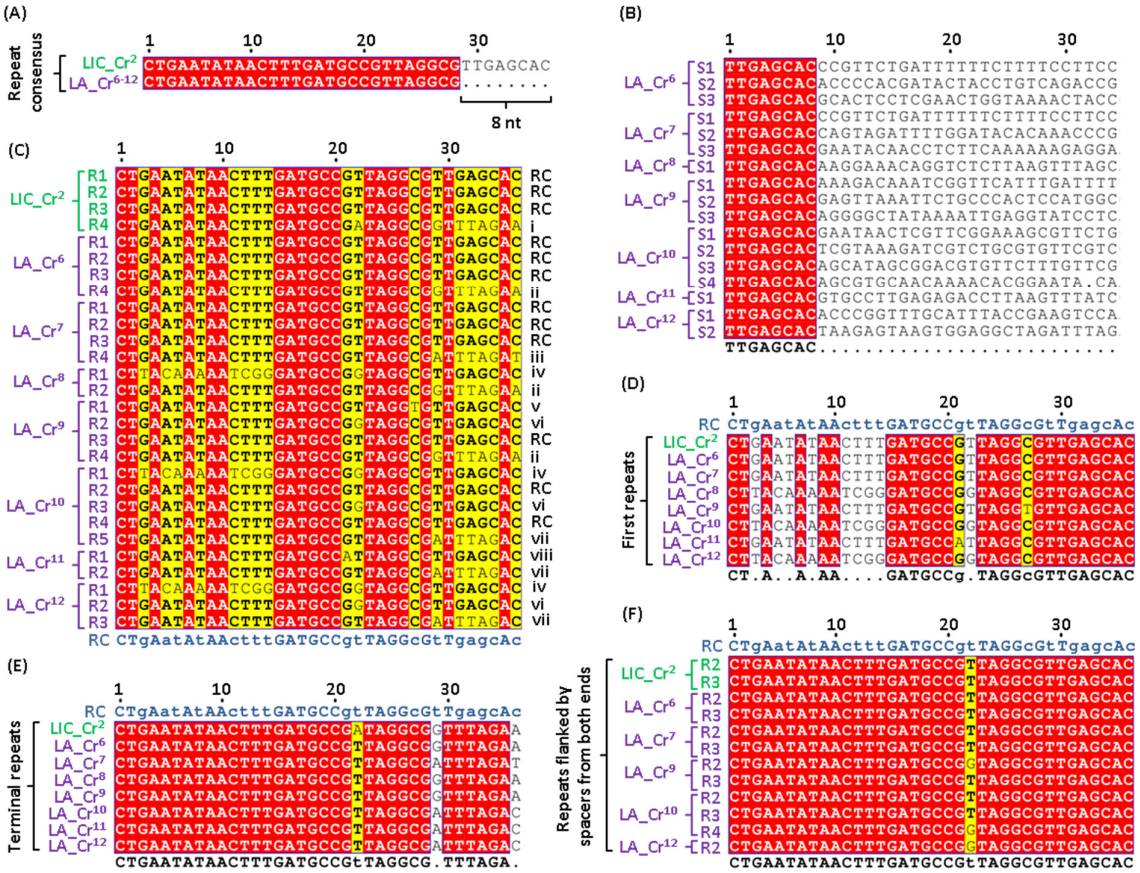


Figure 2.4. Analysis of repeat and spacer sequences of *Leptospira* CRISPR I-B arrays through multiple sequence alignments. (A) Nucleotide alignment of repeat consensus of LIC_Cr^{2} and LA_Cr$^{6-12}$. Alignment of database-defined repeat consensus of LIC_Cr^{2} and LA_Cr$^{6-12}$ revealed a deficit of 8 nt at 3' end of LA_Cr$^{6-12}$ repeats (B) MSA of database-defined spacer sequences of sv. Lai. Alignment of spacer sequences (n=17) of CRISPRs arrays LA_Cr$^{6-12}$) shows identical 8 nt sequences at the 5' end of spacers (red filled box). (C) MSA of repeat sequences. Database-defined repeats of LIC_Cr^{2} and curated repeats of LA_Cr$^{6-12}$) were aligned. The alignment showed conservation and polymorphism in the 1st to 36th nucleotide position of repeat sequences. Repeat consensus (RC) and repeat variants (RVs; i-viii) are indicated at the bottom and right to the aligned repeat sequences. MSA of first repeats (D), terminal

repeats (E), and repeats flanked by spacers at both ends (F). Red and yellow colors in the alignment-graphics represent 100% and more than 70% identity, respectively. Consensus sequences of MSA performed in D-F were shown below the alignment, where upper/lower case denotes conserved/semi-conserved nucleotides. Consensus nucleotides in the alignments are represented by bold letter codes. Dots (.) in consensus sequences shows no conservation at that particular position.

2.2.3 Analysis of the hypervariable region at the I-B locus of *Leptospira*

Discontinuous arrangement of multiple CRISPR arrays at the I-B locus of sv. Lai prompted us to explore the inter-array sequences in the hypervariable region. Subsequent arrays in the series LA_Cr$^{6\text{-}12}$ (108-323 bp) are separated by inter-array regions that range from 112 to 229 bp (**Figure 2.5A**).

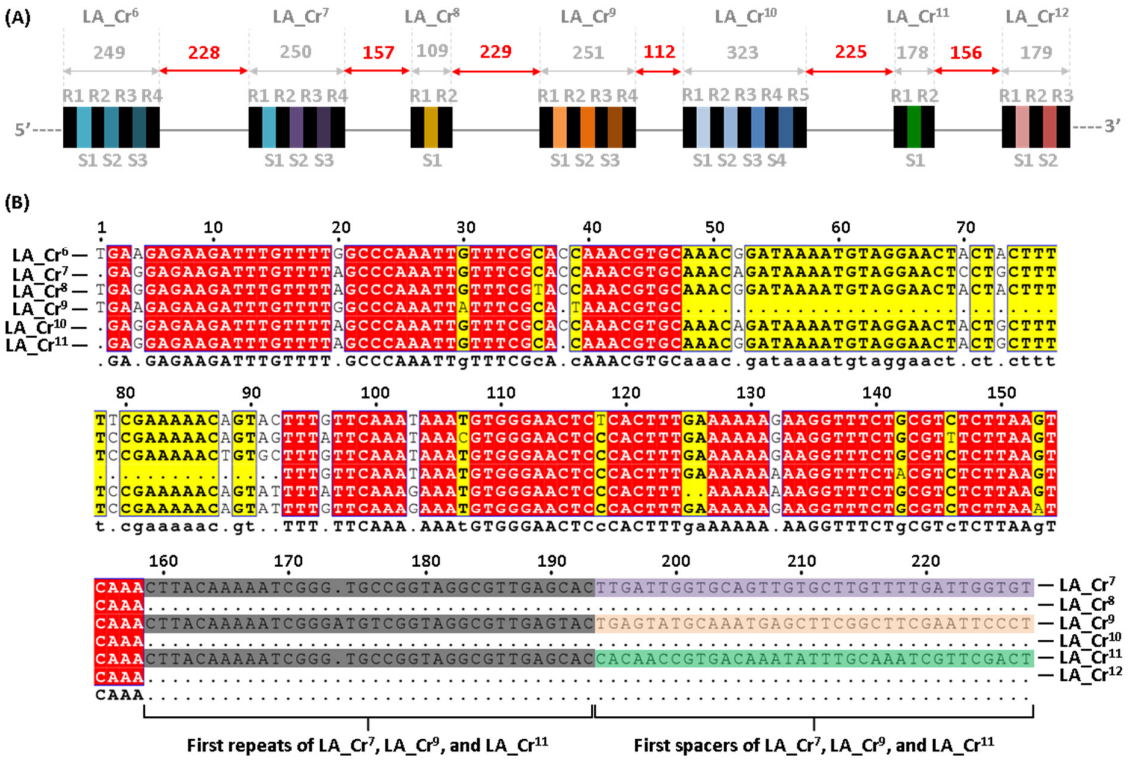


Figure 2.5. Analysis of inter-array regions at the hypervariable region of serovar Lai. (A) The architecture of subtype I-B array locus in serovar Lai. CRISPR arrays and inter-array regions are drawn to the scale in the architecture. The length of CRISPR (grey) and inter-array regions (red) are indicated by numbers (in bp) given over the double arrowhead on top of the architecture. (B) Multiple sequence alignment (MSA) of regions flanked by two adjacent CRISPR arrays. Grey and unique colored regions at the 3' end of alignment represent the first repeat-spacer units (redefined now) of LA_Cr^{7}, Cr^{9}, and Cr^{11}. Red and yellow highlighted nucleotides in the alignment-graphic represent 100% and more than 70% identity, respectively. Consensus sequences of MSAs are shown below the alignment where upper and lower case denotes conserved and semi-conserved nucleotides, respectively.

The alignment of the five inter-array sequences (LA_Cr^{6}-Cr^{7}, LA_Cr^{7}-Cr^{8}, LA_Cr^{8}-Cr^{9}, LA_Cr^{9}-Cr^{10}, and LA_Cr^{10}-Cr^{11}) showed high similarity in the 155-158 nt from 5′ end **(Figure 2.5B)**. However, beyond 155-158 nt, three inter-arrays sequences (LA_Cr^{6}-Cr^{7}, LA_Cr^{8}-Cr^{9}, and LA_Cr^{10}-Cr^{11}) possessed additional repeat (35-36 bp), and the spacer-like sequences (36 bp). The repeat-like sequences were similar, with more than 94% identity to one of the RVs (iv). Moreover, these newly identified repeat-spacer-like sequences were positioned immediately upstream of arrays LA_Cr^{7}, Cr^{9}, and Cr^{11}. Therefore, we included these newly identified repeat-spacer units as a part of their adjacent CRISPR arrays and thus, the boundaries of arrays LA_Cr^{7}, Cr^{9}, and Cr^{11} have been redefined **(Figure 2.5B, 2.6 and Table S2.3)**. The inclusion of three new repeat-spacer units in the arrays resulted in two additional RVs (ix and x) in the list of previously observed RVs (i-viii) **(Table 2.4)**.

Table 2.4. CRISPR I-B repeats variants identified in svs. Copenhageni and Lai

Repeat types	CRISPR repeats	Sequences (sense, 5′-3′)
RC (repeat consensus)	LIC_Cr^{2} R1, 2, and 3 LA_Cr^{6} R1, 2, and 3 LA_Cr^{7} R2, 3, and 4 LA_Cr^{9} R4 LA_Cr^{10} R2, and 4	CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC
i	LIC_Cr^{2} R4	CTGAATATAACTTTGATGCCGATAGGCGGTTTAGAA
ii	LA_Cr^{6} R4 LA_Cr^{8} R2 LA_Cr^{9} R5	CTGAATATAACTTTGATGCCGTTAGGCGGTTTAGAA
iii	LA_Cr^{7} R5	CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAT
iv	LA_Cr^{8} R1 LA_Cr^{10} R1 LA_Cr^{12} R1	CTTACAAAAATCGGGATGCCGGTAGGCGTTGAGCAC
v	LA_Cr^{9} R2	CTGAATATAACTTTGATGCCGTTAGGTGTTGAGCAC
vi	LA_Cr^{9} R3 LA_Cr^{10} R3 LA_Cr^{12} R2	CTGAATATAACTTTGATGCCGGTAGGCGTTGAGCAC
vii	LA_Cr^{10} R5 LA_Cr^{11} R3 LA_Cr^{12} R3	CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAC
viii	LA_Cr^{11} R2	CTGAATATAACTTTGATGCCATTAGGCGTTGAGCAC
ix	LA_Cr^{7} R1 LA_Cr^{11} R1	CTTACAAAAATCGGG-TGCCGGTAGGCGTTGAGCAC
x	LA_Cr^{9} R1	CTTACAAAAATCGGGATGTCGGTAGGCGTTGAGTAC

Underlines and hyphens show polymorphism and deletion of nucleotide, respectively, in repeat variants compared to the repeat consensus.

For the development of memory in prokaryotes against MGEs, integration of protospacer DNA into the CRISPR array is the primary stage of the CRISPR-Cas immunity, which depends on the leader and consensus nucleotides within repeats (Mosterd, Rousseau et al. 2021). The length of the leader sequences has been reported to range from 100 to 500 nt where the first 10-43 nt at the leader-repeat junction is critical for adaptation (Yosef, Goren et al. 2012; Carte, Christopher et al. 2014). Furthermore, mutation of the repeat's 8 nt at the leader junction hampers the adaptation process (Grainy, Garrett et al. 2019). In this study, the identification of transcripts from 7 CRISPR arrays (LA_Cr^{6} to Cr^{12}) suggested that these arrays in sv. Lai is controlled by a single leader. In the hypervariable region of sv. Lai, the nucleotide sequences flanking the 5′ end of the first array's repeat (LA Cr^{6}) differ from those of the remaining six CRISPR arrays (LA Cr$^{7-12}$). Therefore, such variations at the first repeats may disrupt the adaptation process in six CRISPR arrays (LA_Cr$^{7-12}$) of sv. Lai. Thus, we hypothesize that integration of a new spacer exclusively occurs at the first CRISPR array (LA_Cr^{6}) during adaptation in CRISPR-Cas subtype I-B of sv. Lai. However, further research is needed to confirm this notion.

After analyzing the variation in CRISPR I-B arrays in sv. Copenhageni and Lai, the hypervariable region between both serovars, were compared. The region between *cas*-operon I and proximal repeat (R1 of LA_Cr^{6} and LIC_Cr^{2}) in both serovars is 365 bp long. This region is highly conserved; with 99.7% sequence identity and 100% query coverage **(Figure 2.6)**. The region between operon II and proximal repeat (LA_Cr^{12} R3 and LIC_Cr^{2} R4) is 851 and 460 bp long in sv. Lai and Copenhageni, respectively. They share 98.7% sequence identity and 81% query coverage. Within these regions, around 244 bp upstream to *cas*-operon II and 77 bp downstream to the proximal repeat of *cas*-operon II are highly similar between svs. Lai and Copenhageni **(Figure 2.6)**. Moreover, the DNA segment of 77 bp (denoted by the red line) downstream to LIC_Cr^{2} showed high similarity with the DNA segment of 77 bp downstream of each seven arrays LA_Cr$^{6-12}$ **(Figure 2.6)**. This identical feature in the hypervariable region of sv. Copenhageni and Lai suggests that the downstream DNA segment (~77 bp) of each CRISPR I-B array in *Leptospira* is conserved. Apart from CRISPR I-B arrays, identical genes *LA3183* and *LIC10940* encoding hypothetical proteins were also identified in the hypervariable region of svs. Copenhageni and Lai.

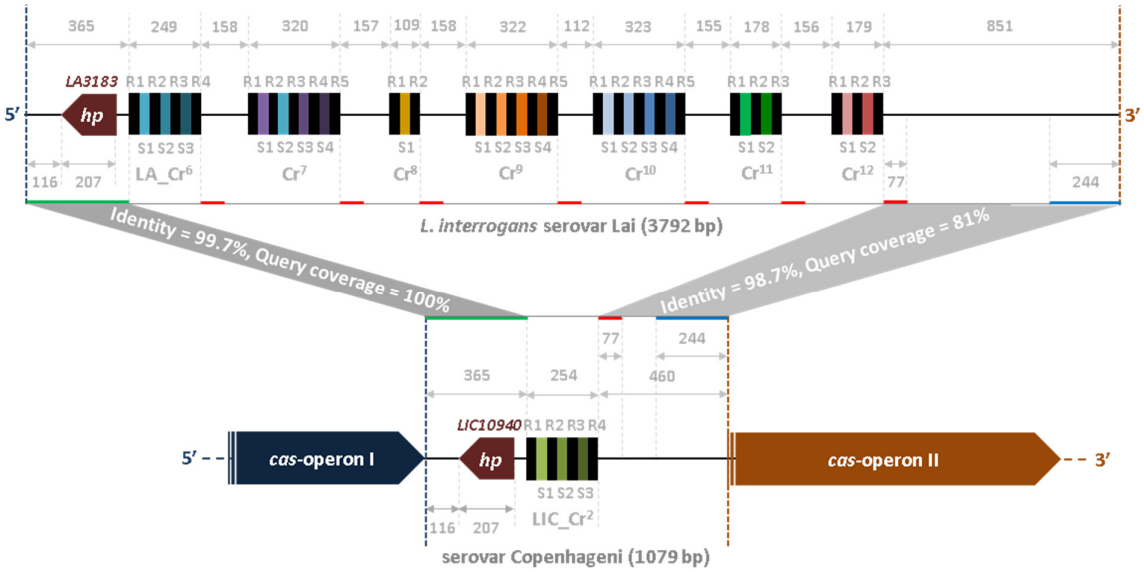


Figure 2.6. Comparative analysis of hypervariable region at subtype I-B loci between serovars Lai and Copenhageni. The region between *cas*-operon I and *cas*-operon II in serovar Lai (3792 bp; top panel) and Copenhageni (1079 bp; bottom panel) are drawn to scale in the direction of the CRISPR-Cas I-B (5'-3'). Highly similar regions shared between svs. Lai and Copenhageni are demarcated by the same color-coded (green, red, and blue) thick horizontal lines along thin lines (solid grey) that correspond to the hypervariable region in *Leptospira* serovars. Identical genes *LIC10940* and *LA3183* encoding hypothetical proteins (*hp*) are demarcated by pentagons (dark red). In the CRISPR arrays, black and unique color-filled rectangles represent repeat and spacer regions, respectively.

Since the CRISPR boundaries of sv. Lai genome have been redefined, the rectified spacer sequences **(Table S2.3)** were utilized to identify PAMs for the *Leptospira* I-B system by *in silico* approach. Feeding of 23 spacers DNA of sv. Lai into the CRISPRTarget program with a cut-off score of 25 resulted in 53 hits as possible protospacers. In the CRISPRTarget output, spacers RNA were found to align with *Leptospira* phages (LnoZ_CZ214, LinZ_10, Lin_34, and LbrZ_5399) and viral genome fragments derived from metagenomic samples. Immediately upstream to the 37 out of 53 (~70%) predicted protospacers, a trinucleotide sequence (5′-ATG-3′) was conserved, as apparent in the sequence logo of nucleotides flanking protospacer sequences **(Figure 2.7)**. Therefore, we speculate that the PAM sequence 5′-ATG-3′ may be employed for the interference against MGEs in *Leptospira*. In a study performed elsewhere (Xiao, Yi et al. 2019), a weakly conserved PAM sequence (5′-TAC-3′) for the *L. interrogans* subtype I-B was identified through the same approach. Such variability in the predicted PAM could be due to feeding database-defined length or orientation of spacers during *in silico* analysis. Recently, a computational pipeline (Vink, Baijens et al. 2021) has predicted the PAM

sequence 5′-ATG-3′ for the *Leptospira* subtype I-B system, which agreed well with our result presented in this study.

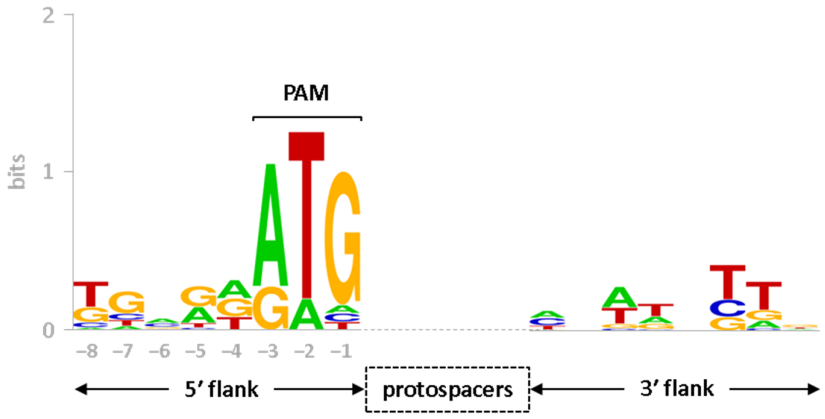


Figure 2.7. Sequence logo of protospacer-flanks. Conservation of nucleotides flanking protospacers at the 5' and 3' (8 nt each) is depicted in the form of a sequence logo. Conserved nucleotides at −3 to −1 position (5'-ATG-3') in the upstream of protospacers represent the consensus PAM sequence. The Y axis indicates the information content of the sequence in bits. The height of the nucleotide code in the logo represents the nucleotide conservation at each position in the protospacer flanks.

2.2.4 Transcriptional analysis of CRISPR I-B arrays in serovar Lai

The transcriptional analysis of CRISPR I-B arrays (LA_Cr$^{6\text{-}12}$) of sv. Lai was performed using RT-PCR. The primer pairs employed in RT-PCR were synthesized so that they annealed to the spacers of each array **(Table S2.1 and Figure 2.8A)**. The specificity of these primer pairs was first checked by PCR using the genomic DNA of sv. Lai. The expected size of DNA amplicons [177 bp (LA_Cr^{6} S1S3), 249 bp (LA_Cr^{7} S1S4), 250 bp (LA_Cr^{9} S1S4), 251 bp (LA_Cr^{10} S1S4), and 107 bp (LA_Cr^{11} S1S2 and LA_Cr^{12} S1S2)] were obtained after agarose gel electrophoresis of the PCR products (**Figure 2.8A and Figure 2.8B, left panel**). With the same sets of primers, similar sizes of amplicons were also evident in RT-PCR, where cDNA made using random hexamers was a template (**Figure 2.8B, middle panel**). No amplification in control RT-PCR reactions suggested that RNA was free of DNA contamination (**Figure 2.8B, right panel**). Thus, the active transcription of six out of seven CRISPR arrays (LA_Cr$^{6\text{-}7}$ and LA_Cr$^{9\text{-}12}$) at the hypervariable region of sv. Lai was confirmed.

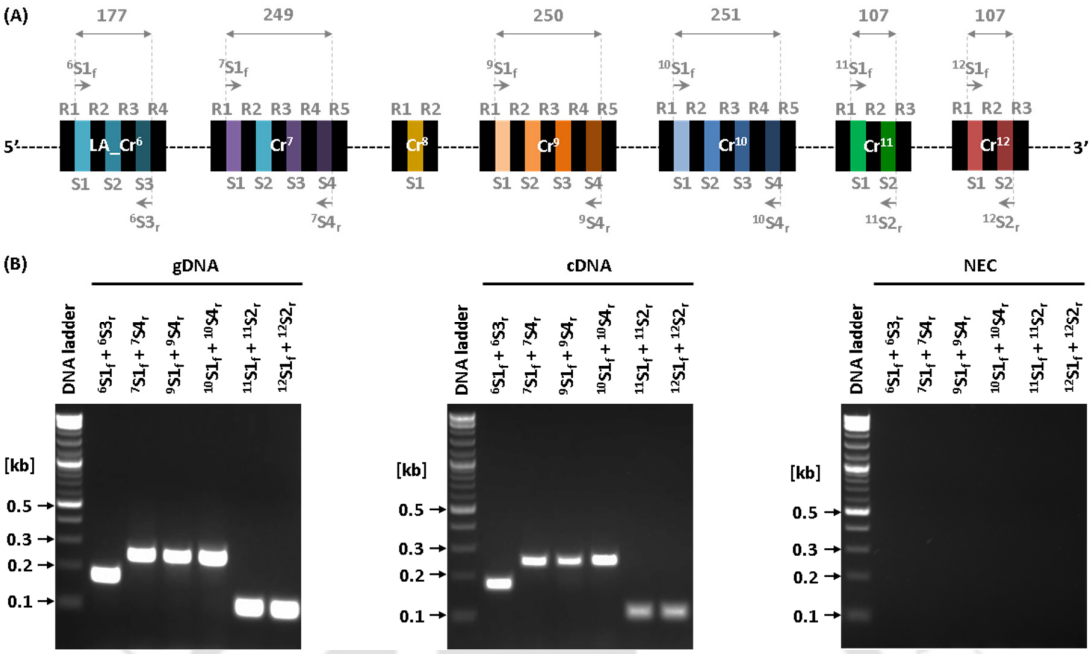


Figure 2.8. RT-PCR of CRISPR I-B arrays of sv. Lai. (A) Schematic representation of primer pairs used in the RT-PCR analysis. CRISPR spacer-specific primer pairs used in the RT-PCR experiment are depicted along the architecture of the hypervariable region (I-B) of sv. Lai. The length of DNA fragments within CRISPR arrays that were expected in PCR with the corresponding primer pair are indicated by vertical dashed lines (grey). Numbers given over double arrowheads on top of the architecture denote the size of DNA fragments (in bp) to be amplified in PCR. (B) Identification of CRISPR array transcription in sv. Lai. PCR using genomic DNA (gDNA; positive control) of sv. Lai with the first and terminal spacer-specific primer pair of each CRISPR (except LA_Cr^{8}) (left panel). PCR using cDNA template synthesized from total RNA of serovar Lai with random hexamers (middle panel). PCR with RNA (cDNA synthesis reaction without reverse transcriptase) is a no enzyme control (NEC) of the RT-PCR experiment (bottom panel). PCR products were resolved on 2% agarose gel.

Due to the discontinuous arrangement of CRISPR I-B arrays in sv. Lai, it was unclear whether the seven CRISPR arrays (LA_Cr$^{6\text{-}12}$) are transcribed as multiple independent pre-crRNA or as a single long pre-crRNA. Therefore, an additional RTPCR experiment was performed to amplify consecutive arrays with partial overlapping regions, as presented graphically in **Figure 2.9A**. The partial overlapping DNA fragments were amplified using the primer pairs enlisted in **Table S2.1** and the cDNA (prepared using random hexamer) as a template**.** In RT-PCR products, the amplicons of three overlapping CRISPR regions [LA_Cr^{6} S1-Cr^{7} S4 (655 bp), LA_Cr^{7} S2-Cr^{9} S4 (955 bp), and LA_Cr^{9} S2-Cr^{10} S4 (685 bp)] were detected on the agarose gel **(Figure 2.9B)**. This result indicated a continuous transcription from LA_Cr^{6} to LA_Cr^{12} in sv. Lai. To confirm this hypothesis, an additional RT-PCR experiment was performed using cDNA generated with a single primer (^{12}S2$_r$) specific to the spacer of the terminal array

(LA_Cr^{12}). PCR with this cDNA and primer pair ($^{6}S1_f$ and $^{6}S3_r$) specific to the first array (LA_Cr^{6}) could amplify a DNA amplicon of 177 bp **(Figure 2.9C)**. Thus, seven CRISPR I-B arrays (LA_Cr$^{6\text{-}12}$) were transcribed as a single long pre-crRNA in sv. Lai.

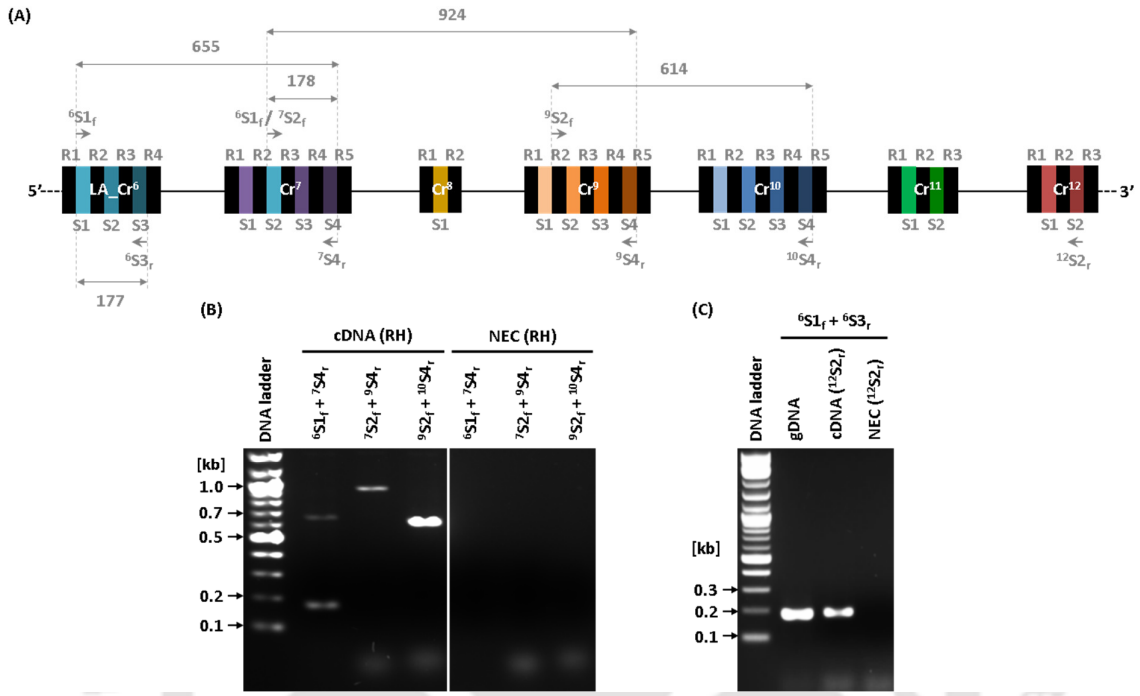


Figure 2.9. RT-PCR of CRISPR I-B arrays to decipher pre-crRNA in sv. Lai. (A) Schematic representation of primer pairs and binding positions used to amplify cDNA of serovar Lai. The primers used in RT-PCR analysis and corresponding annealing regions are indicated over the graphic for clarity. LA_Cr^{6} to LA_Cr^{12}, including CRISPR flanking regions, are drawn to scale in the direction of CRISPR-Cas I-B (5'-3'). Similar repeats and distinct spacers in CRISPRs are represented by black and unique color-filled rectangles, respectively. Vertical dashed lines (grey) and numbers (in bp) given over the double arrowhead indicate DNA regions and corresponding lengths that could be amplified in RT-PCR. (B) Identification of transcription from one CRISPR to another in sv. Lai. The aforementioned primer pairs were used in RT-PCR with cDNA (right panel) or a control template (right panel) prepared using random hexamers. (C) RT-PCR of first array LA_Cr^{6} using cDNA prepared with spacer specific primer of terminal CRISPR of LA_Cr$^{6\text{-}12}$ series. A reverse transcription reaction was set up using a spacer-specific reverse primer of terminal array LA_Cr^{12}. This cDNA template was used in RT-PCR for the amplification of LA_Cr^{6}. PCR products were resolved on 2% agarose gel.

After deciphering the active transcription of CRISPR I-B arrays in svs. Copenhageni and Lai, the abundance of pre-crRNAs was assessed using quantitative real-time PCR (q-PCR). Primer pairs chosen for this qPCR experiment were based on the amplicon size of ~180-250 bp **(Figure 2.10A)**. With these primer pairs and cDNA synthesized using random hexamers, a q-

PCR analysis of CRISPR I-B arrays (LIC_Cr^{2} of sv. Copenhageni, LA_Cr^{6}, Cr^{7}, Cr^{9}, and Cr^{10} of sv. Lai) was performed. The number of pre-crRNAs in both serovars was nearly 8 copies per 10^{6} copies of *16S rRNA* transcripts of respective serovars **(Figure 2.10B)**. In addition, the relative number of pre-crRNAs in serovar Lai substantiates the transcription of the seven arrays (LA_Cr$^{6\text{-}12}$) jointly as a single precursor unit.

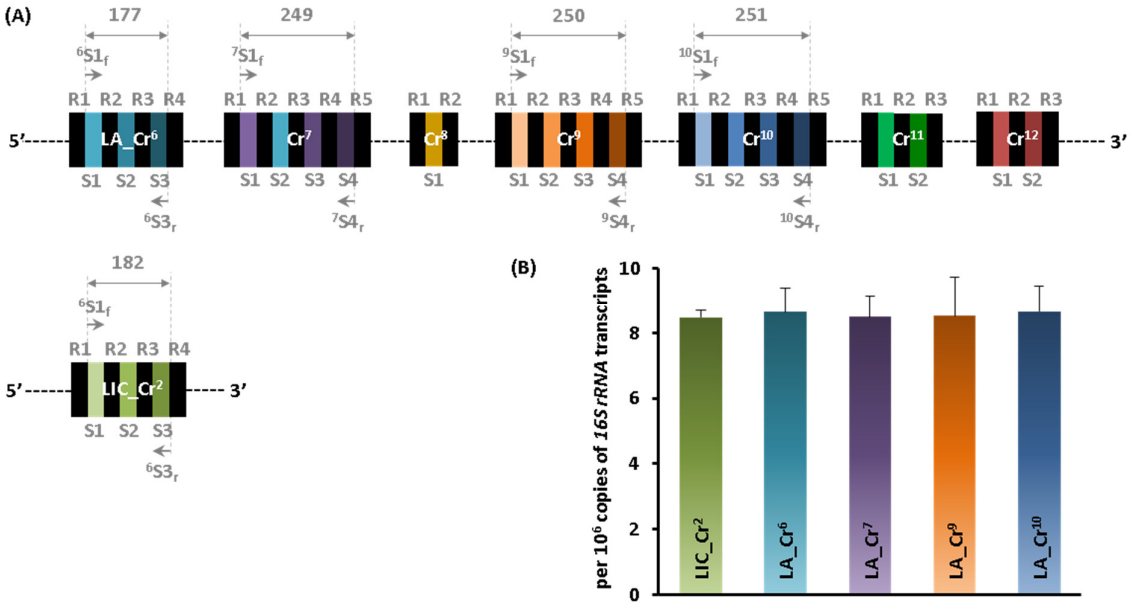


Figure 2.10. Quantitative real-time PCR of CRISPR I-B arrays in serovars Copenhageni and Lai. (A) Schematic representation of primer pairs and binding positions used to amplify cDNA of serovar Lai. For clarity, the primers used in q-PCR analysis and corresponding annealing regions are indicated over the graphic. CRISPR arrays of svs. Copenhageni and Lai are drawn to scale in the direction of CRISPR-Cas I-B (5'-3'). Similar repeats and distinct spacers in CRISPRs are represented by black and unique color-filled rectangles, respectively. Vertical dashed lines (grey) and numbers (in bp) given at the apex of the double arrowhead over the architecture indicate DNA regions and corresponding lengths that could be amplified in q-PCR. (B) Quantification of CRISPR I-B array transcripts in sv. Copenhageni and Lai. Transcripts of CRISPR arrays (LIC_Cr^{2}, LA_Cr^{6}, Cr^{7}, Cr^{9,} and Cr^{10}) are quantified using q-PCR per 10^{6} copies of *16 S rRNA* transcripts of respective *Leptospira* serovars. Results are indicative of two independent experiments, each performed in quadruplets.

An outer membrane lipoprotein LipL32, with a copy number of approximately 38000 per cell, is the most abundant protein of *Leptospira* (Murray 2013). Compared to LipL32, other known outer membrane proteins such as Loa22 (Ristow, Bourhy et al. 2007), FlaB (Lin, Surujballi et al. 1997), LipL41 (Shang, Summers et al. 1996), and LipL21 (Cullen, Haake et al. 2003) (~30000, 20000, 10500, 8800 copies per cell, respectively) present at a considerably lower

level in *Leptospira* (Murray 2013). Moreover, Copies of leptospiral *16S rRNA* transcripts were reported to be nearly 1000 folds more than the LipL32 transcripts (Backstedt, Buyuktanir et al. 2015). Thus, we can correlate that about 8 copies of pre-crRNA were detected per 10^3 copies of LipL32 (10^6 copies of *16S rRNA*) transcripts in *Leptospira*. These comparisons indicate the presence of pre-crRNAs at the basal level in *Leptospira*. Immediate processing of a larger fraction of expressed pre-crRNA by native LinCas6 endoribonuclease is one of the plausible factors which resulted in extremely low copies of full-length CRISPR transcripts. In *Streptococcus thermophiles* (Young, Dill et al. 2012) and *Thermus thermophiles* (Agari, Sakamoto et al. 2010), phage/viral infection has been shown to upregulate the expression of CRISPR-Cas. Under similar circumstances, it is thus possible that the CRISPR transcription in *Leptospira* may also upregulate upon encountering invading DNA elements. However, further research is needed to pinpoint the factors affecting CRISPR transcription in *Leptospira*.

2.3 Conclusion

In this chapter, we identified the CRISPR locus in two reference genomes of *L. interrogans* (svs. Copenhageni and Lai) using a CRISPR database. To understand the variability in CRISPR I-B locus among *Leptospira*, we compared the corresponding genomic segments between the two serovars. Based on our comparative analysis of CRISPR arrays between the two genomes, we redefined the CRISPR I-B arrays in sv. Lai. We also performed RT-PCR and q-PCR to decipher the transcriptional status of CRISPR I-B arrays in *Leptospira*.

In the serovars of *L. interrogans*, two independent *cas* I-B operons span hypervariable regions that accommodate the CRISPR locus containing a single array or multiple arrays. To identify the CRISPR arrays at the I-B locus of two reference serovars of *L. interrogans* (sv. Copenhageni str. Fiocruz L1-130 and sv. Lai str. 56601), we utilized the CRISPRCasdb database that is a repository of computationally predicted CRISPR-Cas loci through the upgraded version of CRISPRFinder program (CRISPRCasFinder). At the I-B locus of *L. interrogans* sv. Copenhageni, the CRISPRCasdb database revealed a single CRISPR array with unknown orientation. Using RT-PCR, the CRISPR I-B array was found transcriptionally active in sv. Copenhageni. In addition, transcription of this array was identified in the direction of *cas*-operons. Thus, a co-directional arrangement of *cas* genes and CRISPR array was ascertained at the I-B locus in *Leptospira*. It also indicated the location of the leader [towards

cas2 of the *cas*-operon I (*cas4-cas1-cas2*)] that drives the transcription of the CRISPR I-B array in *Leptospira*.

At the I-B locus of sv. Lai, the CRISPRCasdb database revealed seven CRISPR arrays between the two *cas*-operons. The orientation of each of these seven arrays was projected in the opposite direction of *cas* I-B operons, which disagreed with the proposed co-directional arrangement of CRISPR-Cas I-B in *Leptospira*. Analysis of spacer sequences associated with CRISPR I-B arrays of sv. Lai suggested a mishap in the database-annotated repeat-spacer boundaries. Moreover, analysis of inter-array sequences revealed three repeat-spacer units that were not apparent in the database. Thus, we failed to apply the CRISPR direction and boundaries projected by the CRISPRCasdb in the genome of *L. interrogans* sv. Lai. Therefore, based on sequence analyses, we redefined the CRISPR arrays at the I-B locus of sv. Lai.

Alignment of repeats composing CRISPR I-B arrays in svs. Copenhageni and Lai revealed sequence variations, mainly in first and terminal repeats. Apart from repeats identical to the repeat consensus, 10 repeat variants were found to compose the CRISPR arrays (I-B) of svs. Copenhageni and Lai. Comparative analysis of hypervariable regions between svs. Copenhageni and Lai indicated that the sequence between cas-operon I (3′ end of *cas2*) to its proximal repeat is more conserved than that of cas-operon II (5′ end of *cas6*) to its proximal repeat. In addition, in both serovars, around 77 bp regions downstream of every CRISPR I-B array are highly conserved. Sequence variation in the hypervariable region among the serovars of *L. interrogans* appears mainly due to CRISPR-associated spacers. Using the spacer sequences of CRISPR I-B arrays in sv. Lai, a conserved trinucleotide 5′-ATG-3′ immediately adjacent to the 5′ ends of targets (protospacers), was predicted. It suggests that sequence 5′-ATG-3′ as a PAM could be employed during the interference process in *Leptospira*.

Like the active transcription of CRISPR I-B array in sv. Copenhageni, transcript expression from the CRISPR I-B arrays of sv. Lai was also evident in RT-PCR. Moreover, in sv. Lai, a continuous transcription of all seven CRISPR I-B arrays, was demonstrated through RT-PCR. Thus, we concluded that a single leader drives the transcription of multiple CRISPR arrays identified in the hypervariable region of sv. Lai. Such continuous transcription of CRISPR arrays in sv. Lai would generate a long precursor transcript corresponding to multiple CRISPR arrays and inter-array sequences. Through q-PCR analysis, around 8 copies of these precursor transcripts (per 10^6 copies of *16S rRNA* transcripts) were recorded in sv. Lai. Moreover, a similar number of precursor CRISPR I-B array transcripts were also estimated in sv.

Copenhageni. These q-PCR data suggested an active but basal-level transcription of CRISPR I-B arrays in *Leptospira*.

CHAPTER 3

Cloning, expression, purification of recombinant proteins (rLinCas6, rLinCas5, and rLinCas3), and *in vitro* synthesis of pre-crRNAs

In the previous chapter, using the RT-PCR experiments, we have identified that the direction of CRISPR I-B arrays transcription in *L. interrogans* is the same as that of *cas* (I-B)-operons.

However, CRISPRCasdb database analysis in the genome of a reference strain (*L. interrogans* serovar Lai) projected the transcription of CRISPR I-B arrays in the direction opposite to that of *cas*-operons. Thus, to functionally validate the orientation of CRISPR arrays, we intend to utilize the system-specific CRISPR endoribonuclease (LinCas6). We also sought to demonstrate the formation of ribonucleoprotein complex involving crRNA, LinCas6, and LinCas5 that signifies the onset of Cascade formation in *Leptospira*. In addition, we aspire to characterize the LinCas3 nuclease activities to investigate the physiological requirements during target degradation. For the purposes mentioned above, in this chapter, we have cloned the open reading frames of genes encoding LinCas6, LinCas5, and LinCas3 and purified the overexpressed recombinant Cas proteins (rLinCas6, rLinCas5, and rLinCas3). Moreover, polyclonal antisera against rLinCas6, rLinCas5, and rLinCas3 were generated in mice to utilize in the characterization of respective Cas proteins. In addition, to generate the system-specific substrates, pre-crRNAs were synthesized *in vitro* after cloning the CRISPR arrays directionally under the control of the vector's T7 promoter.

3.1 Materials and Methods

3.1.1 Bacterial strains, culturing media, and growth condition

Procured pathogenic *Leptospira* strains were grown in an EMJH medium for genomic DNA isolation, as described in section 2.1.2. Strains of *E. coli* [DH5α and BL21 (DE3)] were cultured at 37ºC in Luria-Bertani (LB) broth medium or LB agar with or without kanamycin or ampicillin [100 µg/ml (HiMedia)] for transformation, and expression studies.

3.1.2 Cloning of *cas* ORFs and CRISPR arrays

The coding DNA sequences (CDS) of *LIC10939* (*cas6*; 633 bp), *LIC10935* (*cas5*; 624 bp), and *LIC10938* (*cas3*; 2253 bp) were amplified in PCR using genomic DNA of sv. Copenhageni as template. After that, genes *LIC10939*, *LIC10935*, and *LIC10938* were cloned individually in the pET28a vector (Novagen). According to the manufacturers' instructions, genes *LIC10935* and *LIC10938* were also cloned individually in the pET-SUMO vector (Invitrogen). A substitution mutation variant of *LIC10939,* where histidine coding 38th codon was mutated to code for alanine, was generated using a site-directed mutagenesis kit (NEB). Full-length [LIC_Cr^{2} R1R4 (254 bp)] and miniature CRISPR arrays [LA_Cr^{6} R2R4 (178 bp) and LA_Cr^{12}

R3R4 (107 bp)] were amplified in PCR with genomic DNA of svs. Copenhageni and Lai, respectively. These CRISPR arrays were cloned separately in the pTZ57RT vector (Thermo Scientific). For the cloning purpose, restriction enzymes (*Nhe*I, *Xho*I, *Kpn*I, and *Hin*dIII) used for the double digestion of inserts (*LIC10939*, *LIC10935*, *LIC10938*, and CRISPR arrays) and vectors (pET28a and pTZ57RT) were purchased from NEB or Thermo Scientific. The set of primer pairs (forward and reverse) used for cloning and mutagenesis purposes are mentioned in **Table 3.1**.

Table 3.1. Primer pairs used in this chapter

Name	Sequence (5′-3′)	Purpose
*LIC10939*_F | *Nhe*I *LIC10939*_R | *Xho*I	CTAGCTAGCATGTCCATTCCGAACGTCA CCGCTCGAGTCACGAAGCTTTACTCTCTACTTTATT	Cloning of *LIC10939* (*cas6*), *LIC10935* (*cas5*), and *LIC10938* (*cas3*) in pET28a
*LIC10935*_F | *Nhe*I *LIC10935*_R | *Xho*I	CTAGCTAGCATGGATCCATTGATTCTTTACTACGA CCGCTCGAGTTAAGAGGGTCGGAGTTTAAACC	
*LIC10938*_F | *Nhe*I *LIC10938*_R | *Bam*HI	CTAGCTAGCGTGATCCTTCTCGCAAAATCAT CCGGGATCCTCAATTCCATTCTCCGCCTCG	
*LIC10935*_F *LIC10935*_R	ATGGATCCATTGATTCTTTACTACGA TTAAGAGGGTCGGAGTTTAAACC	Cloning of *LIC10935* (*cas5*) and *LIC10938* (*cas3*) in pET-SUMO
*LIC10938*_F *LIC10938*_R	GTGATCCTTCTCGCAAAATCAT TCAATTCCATTCTCCGCCTCG	
LIC10939^{H38A}_F *LIC10939*^{H38A}_R	CCCGAATTAGCCGAACACGAT ACAAAGATGGCAAATGGC	Mutagenesis of *LIC10939* (H38A)
*LA3183*_R (P1) LA_Cr7 S4 reverse (P2)	CTATCTAATGATTGGCCAACGC TACGCCGGTTCCTCTTTTTTG	Template (P1P2 and P3P4) generation and cloning of CRISPR arrays [LIC_Cr^{2} R1R4, LA_Cr^{6} R2R4, and LA_Cr^{12} R2R4] in pTZ57R/T
LA_ Cr12 S1 forward (P3) LA3189 upstream reverse (P4)	ACCCGGTTTGCATTTACCGAAG TCATTTTTCGGATTCCATTTTATT	
LIC_Cr^{2}_F | *Hin*dIII LIC_Cr^{2}_R | *Kpn*I	CCCAAGCTTCTGAATATAACTTTGATGCCGTTAGG CGGGGTACCTTCTAAACCGCCTATCGGC	
LA_Cr$^{6/12}$_F | *Hin*dIII LA_Cr$^{6/12}$_R | *Kpn*I	CCCAAGCTTCTGAATATAACTTTGATGCCGTTAGG CGGGGTACCTTCTAAACCGCCTATCGGC	

In the primer name, F and R (after the underscore sign) stand for forward and reverse.

3.1.3 Protein overexpression and purification

Bacterial strain *E. coli* BL21 (DE3) was used to express the rLinCas6 (~25.5 kDa) with N-terminal vector (pET28a) derived 23 residues (~2.5 kDa) containing 6×his tag. The rLinCas6 expression was induced in 1 L culture of *E. coli* BL21 in LB broth medium using 0.5 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) for 4 h at 37 °C and 200 rpm (rotation per minute). The rLinCas6 was purified according to the recommendation for purification of polyhistidine-containing recombinant protein with the Ni-NTA (Nickel-nitrilotriacetic acid) purification system (Thermo Fisher Scientific) with few modifications. Briefly, the induced cells expressing rLinCas6 were harvested through centrifugation (10 min at 5000 rpm and 4 °C), and pellets were washed in 40 ml of PBS [phosphate buffered saline (137 mM NaCl, 2.7 mM

KCl, 8 mM Na_2HPO_4, and 2 mM KH_2PO_4 at pH 7.4)]. The washed cell pellet was resuspended in 10 ml of cold native lysis buffer (50 mM Tris-HCl pH 8.0, 10% glycerol, 300 mM NaCl, and 1% Triton-X-100) and sonicated for 10 min with 5-sec on-off cycles. To remove cellular debris and insoluble proteins from the lysate, the resulting homogenate was centrifuged at 12,000×*g* for 20 min, and then the collected supernatant was filtered (0.2 microns, Axiva). Soluble rLinCas6 from the filtered supernatant was purified by Nickel (Ni)-affinity column (gravity) chromatography using Ni-NTA resins (Invitrogen). For this purpose, around 10 ml of soluble lysate was allowed to pass (three times) through equilibrated (in lysis buffer) Ni-NTA resins in the purification column. After that, resins were washed with two column volumes of native wash buffer (300 mM NaCl and 50 mM Tris-HCl pH 8.0) containing 20 mM imidazole. Then Ni-NTA resin was rewashed with two column volumes of native wash buffer containing 50 mM imidazole. The rLinCas6 was then eluted out using 2 ml of native elution buffer [50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10% glycerol, and 250 mM imidazole]. The collected rLinCas6 elute fractions were analyzed on SDS-PAGE to confirm the presence and purity of rLinCas6. After this, the purified rLinCas6 was dialyzed in a D-tube dialyzer (10 kDa cut-off; Sigma) against a dialysis buffer [50 mM Sodium phosphate (pH 7.4), 150 mM NaCl] and concentrated up to 1 mg/ml in a centricon (10 kDa cut-off; Corning Life Sciences). Aliquots of rLinCas6 (50 µl per tube) were stored as protein stocks at −80°C until further use. The yield of purified rLinCas6 (2 mg/L) was estimated using the Bradford reagent (Thermo Scientific).

The rLinCas5 (~37 kDa) with N-terminal vector (pET-SUMO) derived 6×his-SUMO tag (~13 kDa) was over-expressed in *E. coli* BL21 cells in 1 L LB medium using 0.5 mM IPTG for 4 h at 37 °C and 200 rpm. The rLinCas5 was purified using the denaturing method due to the expression of rLinCas5 as insoluble inclusion bodies in *E. coli* BL21. The induced cells were harvested, washed in PBS, and resuspended in 10 ml denaturing lysis buffer (50 mM phosphate buffer pH 7.8, 300 mM NaCl, and 8 M urea). After sonication, insoluble proteins and cellular debris were separated from the soluble fraction through centrifugation and filtration, as described in the rLinCas6 purification procedure. The collected supernatant was then passed (three times) through the purification column containing equilibrated (in denaturing lysis buffer) Ni-NTA resins. Subsequently, the resins were washed with two column volumes of denaturing wash buffer [50 mM phosphate buffer (pH 6.0), 500 mM NaCl, 8 M urea]. Two more washing (one column volume each) of Ni-NTA resins were performed with denaturing wash buffer containing 50 and 70 mM imidazole. The rLinCas5 bound with resin was then eluted out with 5 ml of denaturing elution buffer [50 mM phosphate buffer (pH 4.0), 500 mM

NaCl, 8 M urea, and 300 mM imidazole]. After confirming the presence and purity of eluted rLinCas5 on SDS-PAGE, the purified rLinCas5 was dialyzed in a D-tube dialyzer against a dialysis buffer [50 mM Sodium phosphate (pH 7.4), 300 mM NaCl, and 10% glycerol]. The dialyzed rLinCas5 was then stored in aliquots at −20°C until further use. The yield of purified rLinCas5 was around 1 mg/L.

The expression of rLinCas3 [6×His-SUMO-LinCas3 (~98 kDa)] was induced in *E. coli* BL21 strain in 1 L of LB broth medium using 0.5 mM IPTG for 20 h at 18°C and 200 rpm. The induced cells were harvested, washed in PBS, resuspended in 10 ml of native lysis buffer, and sonicated. Then insoluble proteins and cellular debris were separated from the soluble fraction through centrifugation and filtration, as described in the rLinCas6 purification procedure. The rLinCas3 from soluble fraction was then purified by the native method using an automated liquid chromatography system [NGC system (BioRad)] according to the operating manual. Elutes of rLinCas3 were collected in 10 ml native elution buffer containing 50-100 mM imidazole. The NGC-purified rLinCas3 was incubated with 10 mM EDTA for 1 h on ice and then subjected to size exclusion chromatography to attain a homogeneous and apo form of rLinCas3. For this purpose, the chromatography column [Superdex Increase 200 column (GE Healthcare)] was first equilibrated with two column volume of equilibration buffer [50 mM sodium phosphate (pH 7.4), 300 mM NaCl, and 2% glycerol] and rLinCas3 was then resolved in SEC. Collected fractions during SEC were examined on SDS-PAGE for the presence of pure rLinCas3. For size estimation of protein collected in SEC elute, the column was calibrated with gel filtration molecular weight markers (Sigma), according to the manufacturer's protocol. Finally, the SEC-purified rLinCas3 was concentrated using centricon in a storage buffer [50 mM sodium phosphate (pH 7.4), 300 mM NaCl, and 10% glycerol] before storing at −20°C until further use. The yield of purified rLinCas3 was around 1.75 mg/L.

3.1.4 Generation of polyclonal antibodies against purified recombinant proteins

Around 15-20 µg (per mouse) of purified recombinant proteins/antigens (rLinCas6, rLinCas5, and rLinCas3) were used to immunize 4-6 weeks old female BALB/c mice. Emulsion of antigens (50-200 µl per mouse) in Freund's complete adjuvant (Santa Cruz Biotechnology) was used for primary immunization of a group of 4-6 mice (one group for each antigen) through subcutaneous injection. A control mouse was injected with an equal volume of PBS emulsified in the same adjuvant. On the 14th and 24th day of primary immunization, two booster injections

of antigens emulsified in Freund's incomplete adjuvant (Santa Cruz Biotechnology) were given to the respective group of mice. On the 10th day of the second booster, blood was collected from each mouse through retro-orbital bleeding. Mice were then sacrificed using atlanto-occipital dislocation method as described before (Kumar, Yang et al. 2010). Sera obtained from collected blood were used for determining antibody-titer by immunoassays before experimental use. Antibodies generation in mice was performed in the Department of Veterinary Microbiology, College of Veterinary Science, Assam Agriculture University Guwahati, India, after approval from the Institutional Animal Ethics Committee.

3.1.5 Enzyme-linked immunosorbent assay (ELISA)

A 96-well microtiter plate (Tarsons) was coated (50 µl per well) with rLinCas6, rLinCas5, or rLinCas3 (400 ng/well) for 2 h at 37°C. After coating, the well's surface was blocked with 3% bovine serum albumin (BSA; 100 µl per well) for 2 h at 37ºC. After three washes with 200 µl per well of PBS containing 0.05% Tween-20 (PBS-T), 50 µl per well of anti-LinCas6/ LinCas5/LinCas3 antibodies at 1:100-10000 dilutions were added in respective wells of the microtiter plate and incubated for 2 h at 37ºC. As a negative control of this ELISA experiment, pre-immune serum collected from mice before the primary immunization was used at similar dilutions. After three washes, 50 µl per well of goat anti-mouse antibody (HRP-conjugated; 1:5000) was probed and incubated for 1 h at 37 ºC. After that, wells were rewashed before adding TMB (Tetramethyl Benzidine) peroxidase substrate (Thermo Scientific; 50 µl per well). After incubation for 10 min at 37ºC, the reaction was terminated with 1 M H_2SO_4 (50 µl per well). Finally, the absorbance was measured at 450 nm wavelength using an ELISA plate reader (Tecan). Data are represented as mean ± standard error mean (SEM) of two independent experiments, each performed in triplicate.

3.1.6 Generation of RNA substrates

According to the manufacturer's recommendation, an RNA synthesis kit (NEB) was used for *in vitro* synthesis of the pre-crRNAs. Briefly, the recombinant plasmids [pTZ57R/T-(LIC_Cr^2 R1R4)/(LA_Cr^6 R2R4)/(LA_Cr^{12} R2R3) from overnight grown culture (20 ml) of *E. coli* DH5α were isolated. The isolated plasmids were linearized using the *Kpn*I restriction enzyme and gel purified. After that, 1 µg of each linearized plasmid was used as a template in RNA synthesis reaction (20 µl) containing reagents provided with the kit [reaction buffer, NTPs (ATP, GTP, UTP, and CTP; 10 mM each), and T7 polymerase mix]. After incubation for 4 h

at 37°C, the reaction was diluted to 100 µl with nuclease-free water (NFW), DNase I buffer, and DNase I enzyme (NEB) and incubated for 15 min at 37°C. Then the reaction volume was adjusted to 200 µl by adding 80 µl NFW and 20 µl of 3 M sodium acetate (pH 5.2; Himedia). After this, the aqueous phase was extracted via two times extractions with an equal volume of 1:1 phenol/chloroform mixture (Himedia). The collected aqueous phase was mixed with one volume of 100% isopropanol and incubated overnight at −20°C. Precipitated RNA was pelleted down by centrifugation and washed with 1 ml of cold 70% ethanol. After air drying of ethanol, the washed RNA precipitate was resuspended in 200 µl of NFW. Each of these pre-crRNAs contains a vector (pTZ57R/T) derived 10 nt sequences (5′GGGAAAGCUU3′) at its 5′ end. The quality of synthesized RNA was examined by urea-polyacrylamide gel electrophoresis before using it as a substrate in assays with recombinant proteins. A similar procedure was followed for the synthesis of *luciferase* mRNA. For this, a FLuc control template provided with the kit was used as a template in the RNA synthesis reaction.

3.2 Results and discussion

3.2.1 Cloning of *LIC10939/cas6*, overexpression, and purification of rLinCas6

To understand the expression stage of CRISPR-Cas defense in *Leptospira*, the gene *LIC10939/cas6* involved in the crRNA biogenesis in *L. interrogans* sv. Copenhageni, was purified after overexpression in a heterologous system (*E. coli*). For this purpose, the open reading frame of *LIC10939* (633 bp, including the stop codon) was PCR-amplified from the genomic DNA of sv. Copenhageni using the gene-specific primer pair **(Table 3.1)** having restriction sites for *Nhe*I and *Xho*I **(Figure 3.1A)**. The *LIC10939* was cloned in the pET28a vector, and the clone was confirmed by double digestion of the recombinant pET28a-*cas6* plasmid using *Nhe*I and *Xho*I restriction enzyme, as evident from an insert fall out of 633 bp on the agarose gel **(Figure 3.1B)**. Using the construct pET28a-*cas6*, the rLinCas6 (~25.5 kDa, including N-terminal 6×his tag) was overexpressed in *E. coli* BL21 cells and purified using Ni-affinity column chromatography **(Figure 3.1C)**. In addition to the wild-type version of rLinCas6, a mutated version of rLinCas6 (rLinCas6^{H38A}) was generated. The clone was confirmed by culture PCR of transformant colonies **(Figure 3.1D)** and sequencing of plasmid pET28a-*cas6*^{H38A}. After that, the rLinCas6^{H38A} was expressed and purified **(Figure 3.1E),** similar to the procedure followed for rLinCas6.

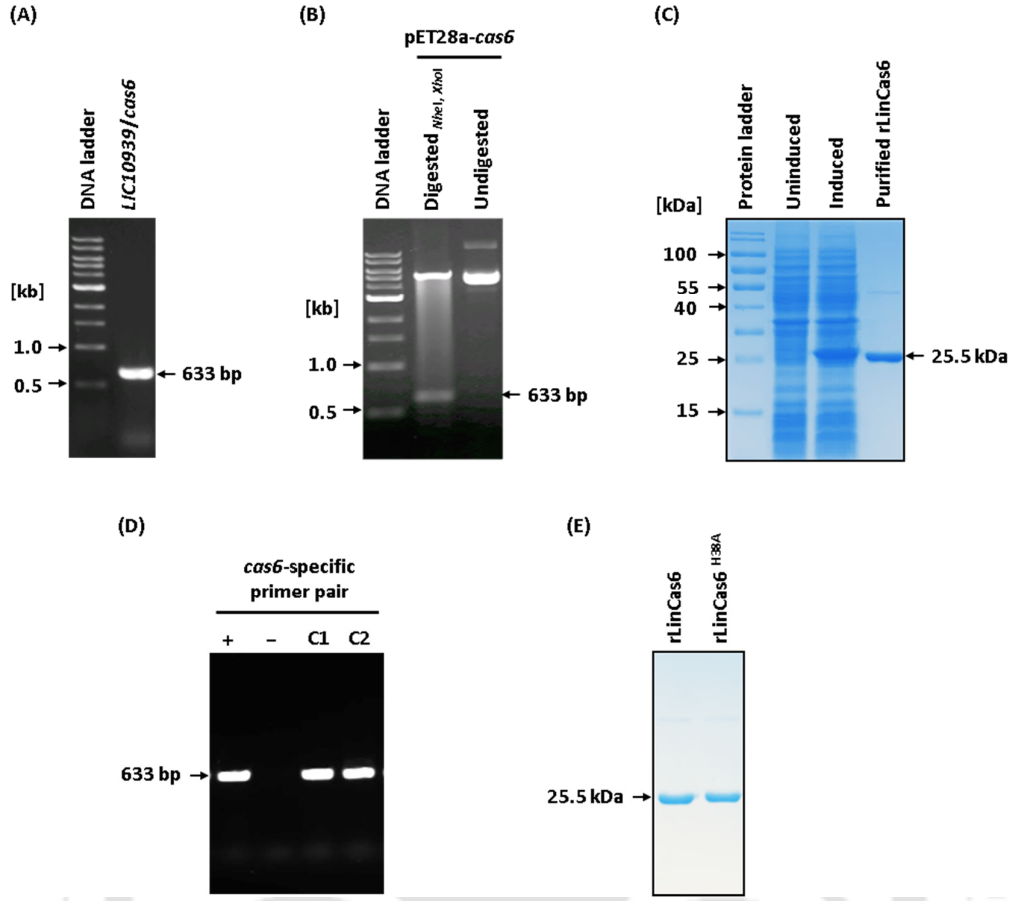


Figure 3.1. Overexpression and purification of rLinCas6 and rLinCas6^{H38A}. (A) PCR amplification of *LIC10939*. The PCR-amplified *LIC10939* gene (633 bp) was resolved on 1% agarose gel. (B) Confirmation of *LIC10939* clone in pET28a vector. The plasmid pET28a-*cas6* was double digested using *Nhe*I and *Xho*I and resolved on the agarose gel. The undigested plasmid pET28a-*cas6* was also resolved on the gel as a control. (C) SDS-PAGE analysis of uninduced and induced cell lysate along with purified rLinCas6. Whole cell lysate of uninduced and induced *E. coli* BL21 cells harboring pET28a-*cas6* construct were resolved on 12% SDS-polyacrylamide gel. Affinity-purified rLinCas6 (2 µg) was also resolved on the gel to examine its purity. (D) Substitution mutation of histidine coding 38th codon of *LIC10939* to alanine. Two of the transformants colonies (C1 and C2) were screened in culture PCR using cas6-specific primer pair. "+" and "−" indicate positive (with genomic DNA template) and negative (no template) control of PCR. (E) SDS-PAGE analysis of affinity-purified rLinCas6^{H38A} (2 µg). An equivalent amount of rLinCas6 was also resolved on the gel as a protein marker.

3.2.2 Cloning of *LIC10935/cas5*, overexpression, and purification of rLinCas5

To characterize and study the interaction of LinCas5 to the crRNA or LinCas6 bound crRNA, the gene *LIC10935/cas6* encoding LinCas5 in *L. interrogans* sv. Copenhageni was purified after overexpression in a heterologous system (*E. coli*). Initially, the gene *LIC10935* (624 bp,

including the stop codon), after PCR amplification **(Figure 3.2A)**, was cloned in the pET28a vector **(Figure 3.2B)**. Expression of rLinCas5 (~26 kDa, including N-terminal 6×his tag) was evident in the whole cell lysate of induced *E. coli* BL21 harboring the construct pET28a-*cas5* **(Figure 3.2C)**. Almost all overexpressed rLinCas5 was observed in the insoluble fraction when induced *E. coli* BL21 cells were lysed in the native lysis buffer **(Figure 3.2D)**. Hence, the rLinCas5 was purified by the denaturing method using Ni-affinity chromatography **(Figure 3.2E, left panel)**. To renature the purified rLinCas5 in its native conformation, dialysis of elutes against the native buffer was performed. Within 2 h of dialysis at 4 °C, elutes of rLinCas5 turned cloudy, suggesting the precipitation of rLinCas5 during dialysis. After dialysis, analysis of the soluble portion and insoluble pellet on SDS-PAGE indicated precipitation of the entire rLinCas5 that was present in elutes **(Figure 3.2E, right panel)**. Moreover, varying the dialysis buffer's pH (6.0-10.0) could not help avoid the precipitation of rLinCas5.

Due to the persistent insolubility problem of rLinCas5, it was decided to overexpress the LinCas5 with a fusion tag (SUMO; Small Ubiquitin-like Modifier). The SUMO fusion tag is known to enhance the solubility of recombinant protein (Costa, Almeida et al. 2014). The PCR-amplified *LIC10935*/*cas5* ORF **(Figure 3.3A)** using primer pair without restriction site overhangs was cloned in the pET-SUMO vector. The positive pET-SUMO-*cas5* clone in the correct orientation to the T7 promoter of the vector was confirmed through culture PCR of *E. coli* DH5α transformant colonies **(Figure 3.3B)**.

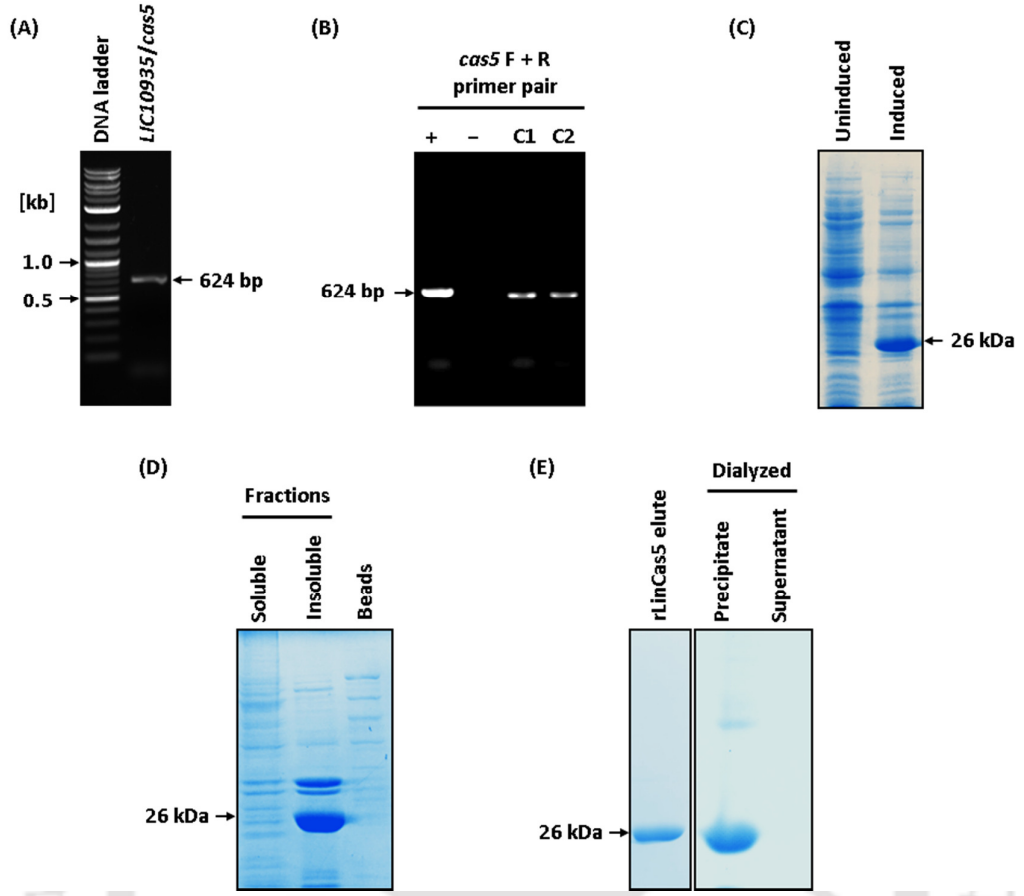


Figure 3.2. Expression and purification of rLinCas5 (N-terminal 6×his-tagged). (A) PCR amplification of *LIC10935*. The PCR-amplified *LIC10935* gene (624 bp) was resolved on 1% agarose gel. (B) Confirmation of *LIC10939* clone in pET28a vector. The transformants *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with gene-specific primer pair. "+" and "−" indicate positive (with genomic DNA template) and negative (no template) control of PCR. (C) SDS-PAGE analysis of uninduced and induced cell lysate. Whole cell lysate of uninduced and induced *E. coli* BL21 cells harboring pET28a-*cas5* construct were resolved on 12% SDS-polyacrylamide gel. (D) SDS-PAGE analysis of fractions after lysis of induced *E. coli* BL21 cells. Soluble (supernatant) and insoluble (pellet) fractions obtained after lysis of Induced cell pellet in the native lysis buffer were resolved on 12% SDS-polyacrylamide gel (middle panel). Ni-NTA resin (beads) incubated with the soluble fraction was also resolved on the gel. (E) Purification and dialysis of rLinCas5. The purified rLinCas5 (~2 µg, 26 kDa) by the denaturing method was resolved on 12% SDS-polyacrylamide gel (left panel). Supernatant and insoluble pellet obtained after dialysis of rLinCas5 were resolved on SDS-polyacrylamide gel (right panel).

Using the construct pET-SUMO-*cas5*, overexpression of the rLinCas5 (~37 kDa) was evident in *E. coli* BL21 cells when induced at both 18 and 37°C **(Figure 3.3C, left panel)**. However, even with the SUMO tag, the overexpressed rLinCas5 was still observed in inclusion bodies of *E. coli* BL21 cells **(Figure 3.3C, middle panel)**. Hence, it was purified by the denaturing method using Ni-affinity chromatography **(Figure 3.3C, right panel)**. The rLinCas5 (6×his-

SUMO-tagged LinCas5), purified by denaturing method, was renatured through dialysis in a native buffer containing 300 mM NaCl, and no precipitation was observed **(Figure 3.3D, left panel)**. However, at a lower concentration of NaCl (150 mM) in the dialysis buffer, precipitation of rLinCas5 was observed. To generate the untagged version of rLinCas5, SUMO-protease treatment was given to the dialyzed lot of rLinCas5. Although the 6×his-SUMO tag could be removed from rLinCas5, precipitation of untagged rLinCas5 was observed in the process **(Figure 3.3D, right panel)**. Owing to this problem, the tagged version of rLinCas5 (6×his-SUMO-tagged LinCas5) was used to characterize LinCas5 in various downstream experiments.

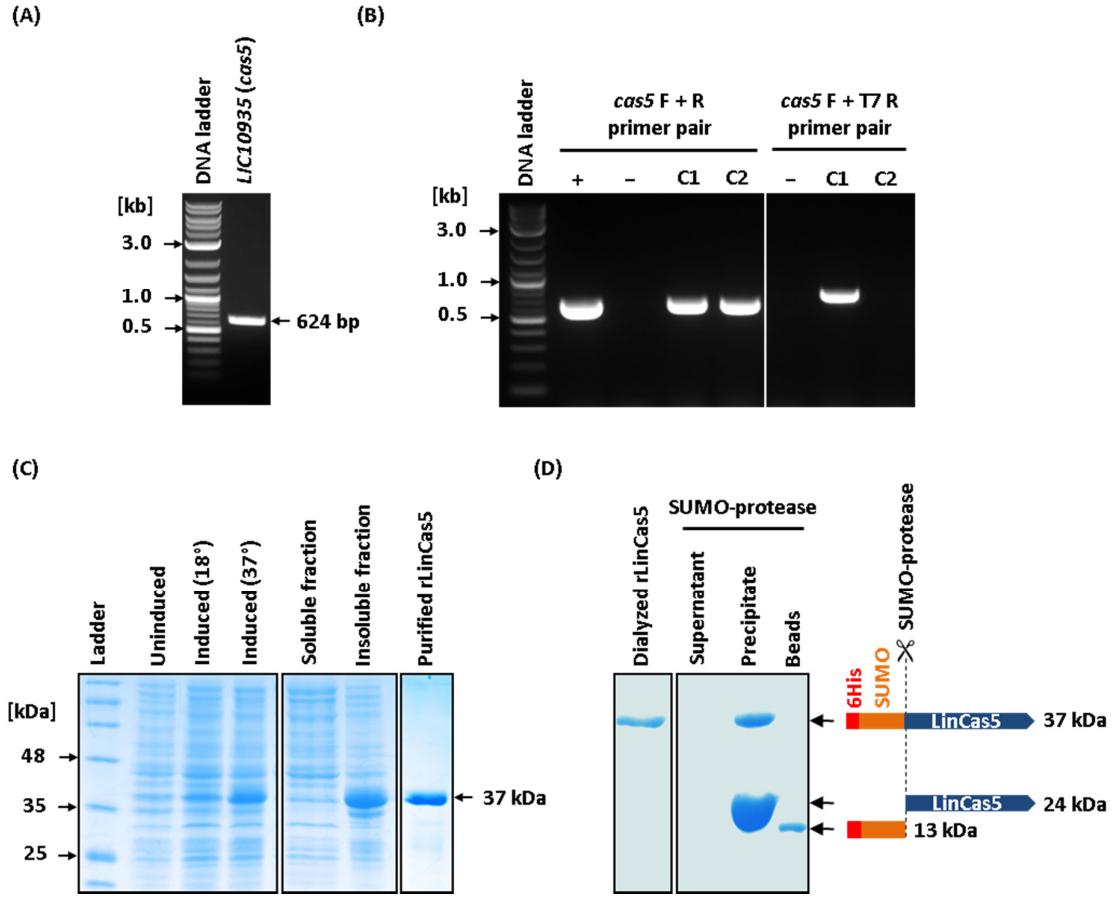


Figure 3.3. Overexpression and purification of rLinCas5 (N-terminal 6×his-SUMO-tagged). (A) PCR amplification of *LIC10935*. The PCR-amplified *LIC10935* gene (624 bp) was resolved on 1% agarose gel. (B) Confirmation of *LIC10935* clone in the correct orientation in pET-SUMO vector. The transformants *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with gene-specific primer pair (left panel) or gene-specific forward and vector-specific T7 reverse primer pair (right panel). "+" and "–" indicate positive (with genomic DNA template) and negative (no template) control of PCR. (C) Purification of rLinCas5. The whole cell lysate of uninduced and induced E. coli BL21 cells harboring the pET-SUMO-*cas5* construct was resolved on 12% SDS-polyacrylamide gel (left panel). Soluble (supernatant) and

insoluble (pellet) fractions obtained after lysis of Induced cell pellet in the native lysis buffer were resolved on 12% SDS-polyacrylamide gel (middle panel). The purified rLinCas5 (~2 µg, 37 kDa) by the denaturing method was resolved on 12% SDS-polyacrylamide gel (right panel). (D) SUMO-protease treatment of dialyzed rLinCas5. SDS-PAGE analysis of dialyzed rLinCas5 (left panel). The reaction containing SUMO-protease and rLinCas5 was incubated with Ni-NTA resin (beads), and elute was collected. The insoluble pellet obtained in elute were separated, and the supernatant was collected. After that, supernatant, precipitate, and beads were resolved on 12% SDS-polyacrylamide gel (right panel). Schematic representation of SUMO-protease mediated cleavage of 6×his-SUMO-tag (red and orange filled rectangles) from the rLinCas5 (~37 kDa) was shown right to the gel image. The blue-filled rectangle represents the untagged rLinCas5 (~24 kDa).

To evaluate the secondary structures and folding of rLinCas5 (dialyzed), circular dichroism (CD) spectroscopy was attempted. However, the high salt concentration (300 mM NaCl) in the buffer led to saturation of the detector signal, and the spectrum in the far-UV (ultraviolet) region was not observed.

3.2.3 Cloning of *LIC10938/cas3*, overexpression, and purification of rLinCas3

To examine the physiological requirements of LinCas3 for its nuclease activity, the gene *LIC10938/cas3* encoding LinCas3 in *L. interrogans* sv. Copenhageni was purified after overexpression in a heterologous system (*E. coli*). Initially, the gene *LIC10938* (2253 bp, including the stop codon), after PCR-amplification **(Figure 3.4A)**, was cloned in the pET28a vector **(Figure 3.4B)**. However, the expression of rLinCas3 (~87 kDa, including N-terminal 6×his tag) was not observed in the whole cell lysate of induced *E. coli* BL21 harboring the construct pET28a-*cas3* **(Figure 3.4C)**. Such difficulty in the overexpression of recombinant Cas3 protein has been reported previously, which was resolved later by expressing the Cas3 with an N-terminal fusion tag (Sinkunas, Gasiunas et al. 2011; Westra, van Erp et al. 2012; Mulepati and Bailey 2013). Thus, we changed the expression vector to pET-SUMO for the heterologous overexpression of rLinCas3. The PCR-amplified *LIC10938/cas3* ORF **(Figure 3.5A)** using primer without restriction site overhangs was cloned in the pET-SUMO vector through TA-cloning. The positive pET-SUMO-*cas3* clone in the correct orientation to the T7 promoter of the vector was confirmed through culture PCR of *E. coli* DH5α transformant colonies **(Figure 3.5B)**. Using the construct pET-SUMO-*cas3*, overexpression of the rLinCas3 (~98 kDa) was evident in *E. coli* BL21 cells **(Figure 3.5C, left panel)**. After affinity purification from the lysate of induced *E. coli* BL21cells, SDS-PAGE analysis suggested the

presence of additional proteins (around 17 kDa) along with rLinCas3 (~98 kDa) in elute **(Figure 3.5C, right panel)**. This is probably due to the cleavage of a fraction of expressed rLinCas3 from its N-terminal end. To purify the rLinCas3 (98 kDa) from the affinity purified elute, size exclusion chromatography was performed. With reference to the standard curve **(Figure 3.5D, left panel)** obtained after SEC of protein markers, the pure rLinCas3 was eluted (96 kDa) at around its monomeric size (98 kDa), while the co-purified protein was eluted at around 23 kDa **(Figure 3.5D, right panel)**. A deviation between the molecular weight of peaks observed in SEC (~96 and 23 kDa) and protein bands in SDS-polyacrylamide gel (~96 and 17 kDa) **(Figure 3.5C, right panel)** could be due to chromatographic error. Fractions containing pure rLinCas3 were pooled together **(Figure 3.5E)** for storage and to use in the biochemical assays. In addition, an untagged version of rLinCas3 (~85 kDa) was generated using SUMO-protease **(Figure 3.5F)**.

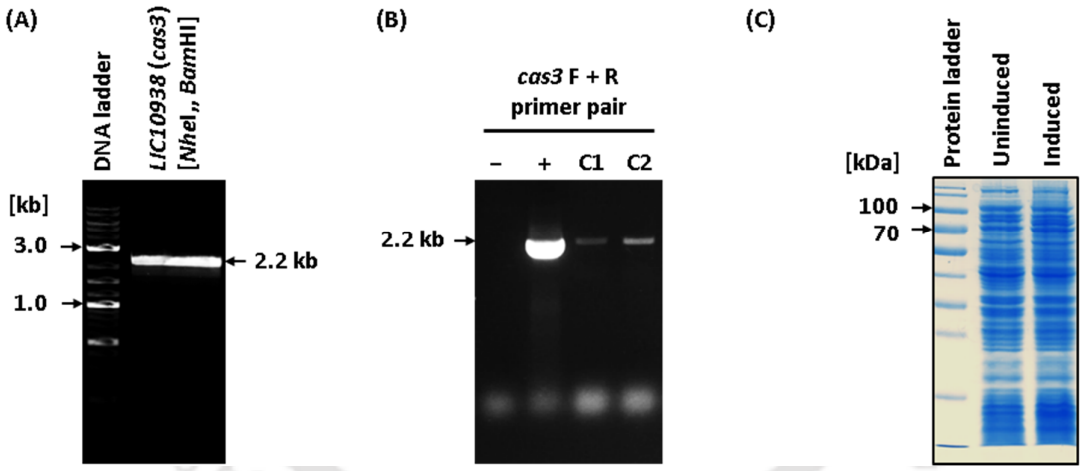


Figure 3.4. Cloning of ORF *LIC10938*/*cas3* encoding LinCas3. (A) Double digestion PCR-amplified *LIC10938*. The PCR-amplified *LIC10938* gene (2253 bp) was double digested using *Nhe*I and *Xho*I restriction enzymes and resolved on 1% agarose gel for gel purification. (B) Confirmation of *LIC10938* clone in pET28a vector. The transformants *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with gene-specific primer pair. "+" and "−" indicate positive (with genomic DNA template) and negative (no template) control of PCR. (C) SDS-PAGE analysis of uninduced and induced cell lysate. Whole cell lysate of uninduced and induced *E. coli* BL21 cells harboring pET28a-*cas3* construct were resolved on 12% SDS-polyacrylamide gel.

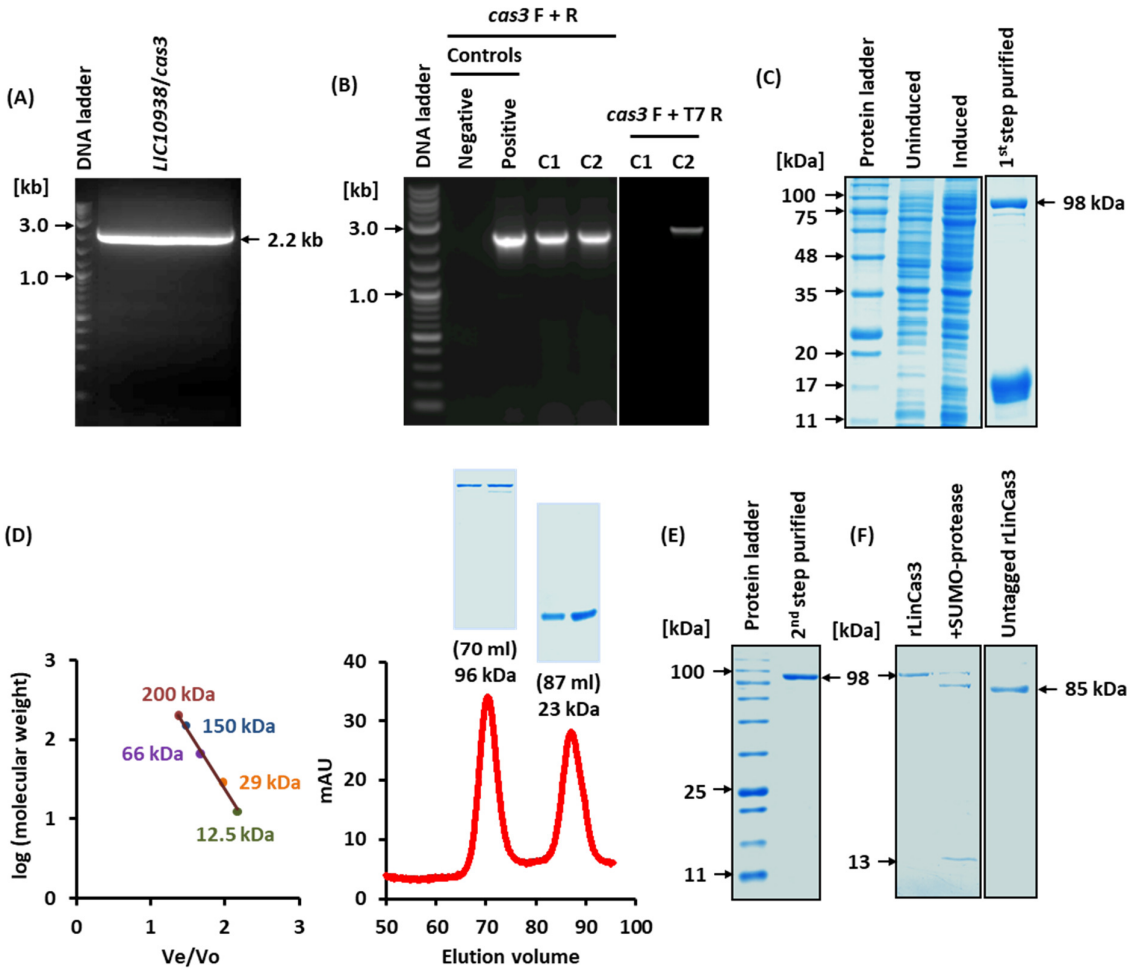


Figure 3.5. Overexpression and purification of rLinCas3 (N-terminal 6×his-SUMO-tagged). (A) PCR amplification of *LIC10938*. The PCR-amplified *LIC10938* gene (2.2 kb) was resolved on 1% agarose gel. (B) Confirmation of *LIC10938* clone in the correct orientation in pET-SUMO vector. The transformant *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with a gene-specific primer pair (left panel) or gene-specific forward and vector-specific T7 reverse primer pair (right panel). Positive and negative control of PCR was performed with genomic DNA and without a template, respectively. (C) Affinity purification of rLinCas3. Whole cell lysate of uninduced and induced *E. coli* BL21 cells harboring pET-SUMO-*cas3* construct (left panel) and affinity-purified rLinCas3 (right panel) were resolved on 12% SDS-polyacrylamide gel. (D) Size exclusion chromatography of affinity purified rLinCas3. A standard curve (logarithm of molecular weight vs. ratio of elution volume to the void volume) was plotted after the SEC run of protein markers (12.5-200 kDa) (left panel). Ve and Vo correspond to elution and void volumes, respectively. SEC profile of affinity-purified rLinCas3 was presented as mAU (280 nm) vs. elution volume (ml). The SDS-polyacrylamide gel shows the proteins in SEC elutes collected at ± 1 ml from the peak points. (E) SDS-PAGE of a pure lot of rLinCas3 purified through SEC. (F) Purification of untagged rLinCas3. After treatment of tagged rLinCas3 (98 kDa) with SUMO-protease, the reaction was incubated with Ni-NTA resin and eluted out (untagged rLinCas3; 85 kDa). The dialyzed untagged rLinCas3 was resolved on 12% SDS-polyacrylamide gel.

3.2.4 Detection of rLinCas6, rLinCas5, and rLinCas3 in ELISA

To aid in the characterization of recombinant Cas proteins, polyclonal antibodies against each purified protein (rLinCas6, rLinCas5, and rLinCas3) were raised in mice. To detect recombinant proteins by corresponding antibodies, indirect ELISA was performed. Microtiter plate coated with rLinCas6, rLinCas5 or rLinCas3 were probed with anti-LinCas6, anti-LinCas5 or anti-LinCas3 sera, respectively, in the dilution range of 1:100-1:10000. Serum-dilution dependent detection of rLinCas6 **(Figure 3.6A)**, rLinCas5 **(Figure 3.6B)** and rLinCas3 **(Figure 3.6C)** suggested that antigen-specific antibodies were generated in mice during immunization. However, pre-immune serum showed cross-reactivity with the rLinCas6 antigen **(Figure 3.6A)**. Such cross-reactivity of pre-immune serum was not observed with rLinCas5 **(Figure 3.6B)** and rLinCas3 antigens **(Figure 3.6C)**. Nevertheless, ELISA data analysis indicated that at least 1:1000 dilution of generated immune sera would be sufficient in the immunoblotting experiment to detect respective antigens.

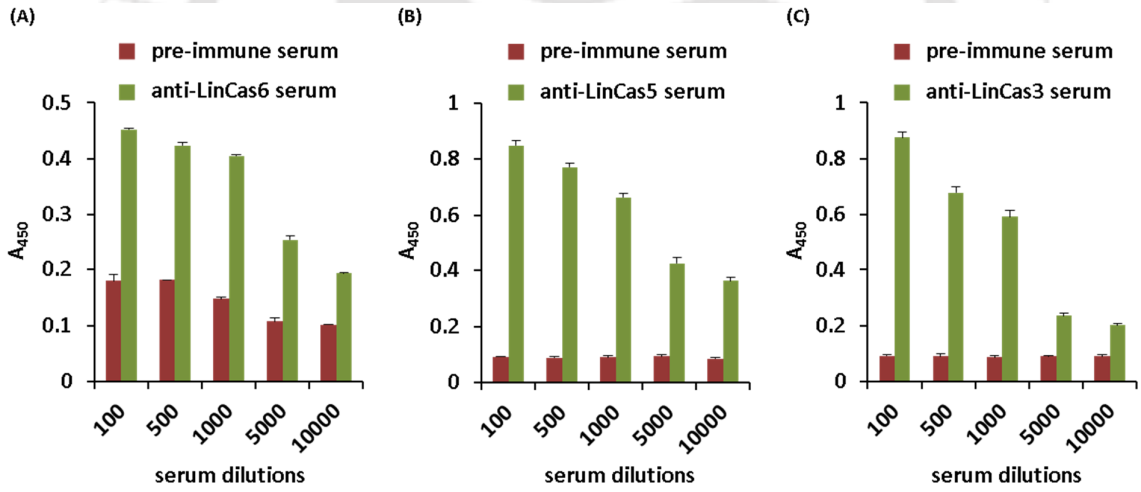


Figure 3.6. Detection of recombinant proteins through ELISA. The immune sera obtained from mice were used in ELISA at 1:100-1:10000 dilutions to detect (A) rLinCas6, (B) rLinCas5, and (C) rLinCas3 (green bars). Sera obtained from mice before the immunization were used as controls (red bar) in these ELISA experiments. Data are represented as mean ± standard error mean (SEM) of two independent experiments performed in triplicate.

3.2.5 *In vitro* synthesis of pre-crRNAs

To understand the processing event during crRNA biogenesis in *Leptospira*, system-specific pre-crRNA substrates were synthesized *in vitro*. At the I-B locus of sv. Copenhageni, a single CRISPR array (LIC_Cr^2) transcribing in the direction of *cas*-operons, was deciphered in

chapter 2. To synthesize the pre-crRNA of LIC_Cr^2, the full-length array (R1R4; 254 bp) was cloned in the appropriate orientation. For this purpose, the DNA fragment of the first to terminal repeat was PCR-amplified using the genomic DNA template of sv. Copenhageni with primer pair carrying restriction sites of *Hin*dIII and *Kpn*I **(Table 3.1)**. This PCR product (insert) and plasmid pTZ57R/T were double digested **(Figure 3.7A)**, ligated, and transformed in competent *E. coli* DH5α. The positive clone of LIC_Cr^2 R1R4 in pTZ57R/T was confirmed by double digestion of plasmid **(Figure 3.7B)** isolated from a transformant *E. coli* DH5α culture. After verifying the clone, the plasmid construct pTZ57R/T-LIC_Cr^2 was linearized using the *Kpn*I restriction enzyme **(Figure 3.7C)**, gel purified, and used as a template in the RNA synthesis reaction. After DNase I treatment, the *in vitro* synthesized LIC_Cr^2 RNA (264 nt, including vector-derived 10 nt sequence at the 5′ end) was further purified from the RNA synthesis reaction by phenol-chloroform extraction **(Figure 3.7D)**.

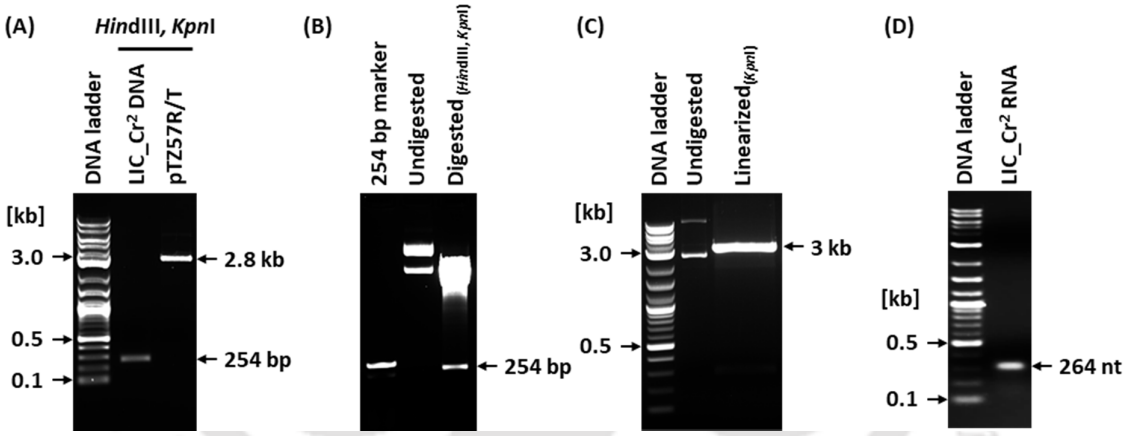


Figure 3.7. *In vitro* synthesis of the full-length LIC_Cr^2 RNA. (A) Double digestion of insert LIC_Cr^2 and vector pTZ57R/T. The full-length array LIC_Cr^2 (254 bp) amplified in PCR and vector pTZ57R/T (2.8 kb) isolated from an overnight grown culture of *E. coli* DH5α were double digested using *Hin*dIII and *Kpn*I enzymes and resolved on 1% agarose gel for extraction. (B) Confirmation of LIC_Cr^2 clone in pTZ57R/T vector. The plasmid isolated from transformant *E. coli* DH5α culture was double digested using *Hin*dIII and *Kpn*I enzymes and resolved on 1% agarose. The undigested plasmid pTZ57R/T-LIC_Cr^2 was also resolved on the agarose gel as a control. (C) Linearization of plasmid pTZ57R/T-LIC_Cr^2 using *Kpn*I. (D) *In vitro* synthesized full-length LIC_Cr^2 RNA (264 nt) was resolved on 1% agarose gel.

Unlike a single CRISPR I-B array (LIC_Cr^2) in sv. Copenhageni, sv. Lai possesses 7 CRISPR I-B arrays (LA_Cr$^{6\text{-}12}$), as described in chapter 2. To synthesize the pre-crRNAs of sv. Lai, a miniature version of the first [LA_Cr^6 (R2R4); 178 bp] and terminal [LA_Cr^{12} (R2R3); 107 bp] CRISPR arrays, were cloned separately in the appropriate orientation. To clone LA_Cr^6 (R2R4), a DNA template was first amplified in PCR with the P1 and P2 primer pair using the

genomic DNA of sv. Lai **(Figure 3.8A and B)**. This was done to avoid the amplification of other CRISPR arrays (LA_Cr$^{7-11}$) in PCR using repeat-specific primer pairs. Using the P1P2 DNA template, another set of PCR was performed with repeat-specific primer pair containing restriction site (*Hin*dIII and *Kpn*I) overhangs **(Figure 3.8C)**. DNA amplicon of LA_Cr^{6} (R2R4; 178 bp) was gel purified, double digested, and cloned in the pTZ57R/T vector. Then the transformant *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with vector T7-specific forward and terminal repeat-specific reverse primer pair **(Figure 3.8D)**. Plasmid pTZ57R/T-LA_Cr^{6} isolated from the positive clone was linearized, gel purified, and pre-crRNA (188 nt) was synthesized **(Figure 3.8E)**, as aforementioned for LIC_Cr^{2}.

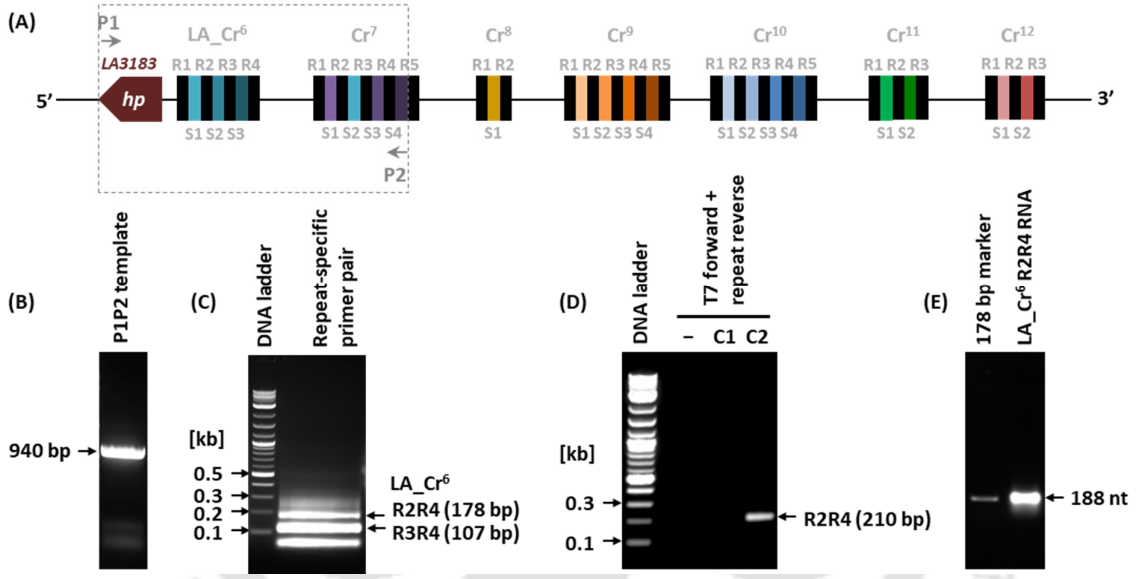


Figure 3.8. *In vitro* synthesis of the miniature LA_Cr^{6} (R2R4) RNA. (A) Schematic of CRISPR I-B locus in sv. Lai. The architecture of CRISPR I-B locus of sv. Lai, including the gene *LA3183* encoding a hypothetical protein (hp), is drawn to scale in the direction of the CRISPR array (5′-3′). Arrowheads and a dashed grey box indicate the binding regions of forward (P1), reverse primer (P2), and amplified fragment in PCR. (B) Generation of P1P2 DNA template. The PCR-amplified P1P2 DNA template (940 bp), with P1 and P2 primer pair using genomic DNA template of sv. Lai was resolved on 1% agarose gel for gel purification. (C) PCR amplification of array LA_Cr^{6}. Using the P1P2 DNA template, the PCR-amplified product with repeat-specific primer pair was resolved on 1% agarose gel. Regions of LA_Cr^{6} amplified in PCR are demarcated right to the gel image. (D) Confirmation of LA_Cr^{6} R2R4 clone in pTZ57R/T vector. The transformants *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with vector-specific T7 forward and repeat-specific reverse primer pair. "–" indicates the negative control (without DNA template) of PCR. (E) *In vitro* synthesized miniature LA_Cr^{6} RNA (188 nt) was resolved on 1% agarose gel.

Similarly, to clone the miniature LA_Cr^{12} (R2R3), a DNA template was first generated from the genomic DNA of sv. Lai in PCR with P3 and P4 primer pair **(Figure 3.9A and B)**. Using

the P3P4 DNA template, the DNA fragment of LA_Cr^{6} (R2R4; 178 bp) was PCR-amplified with repeat-specific primer pair containing restriction site (*Hin*dIII and *Kpn*I) overhangs (**Figure 3.9C**). The double-digested miniature LA_Cr^{6} array was then cloned in the pTZ57R/T vector, and transformant *E. coli* DH5α colonies positive for pTZ57R/T-LA_Cr6 construct were screened through culture PCR (**Figure 3.9D**). To synthesize the pre-crRNA of LA_Cr^{12} R2R3 (117 nt) (**Figure 3.9D**), plasmid pTZ57R/T-LA_Cr^{6} was linearized and used as a template in the RNA synthesis reaction, as described previously for LIC_Cr^{2}.

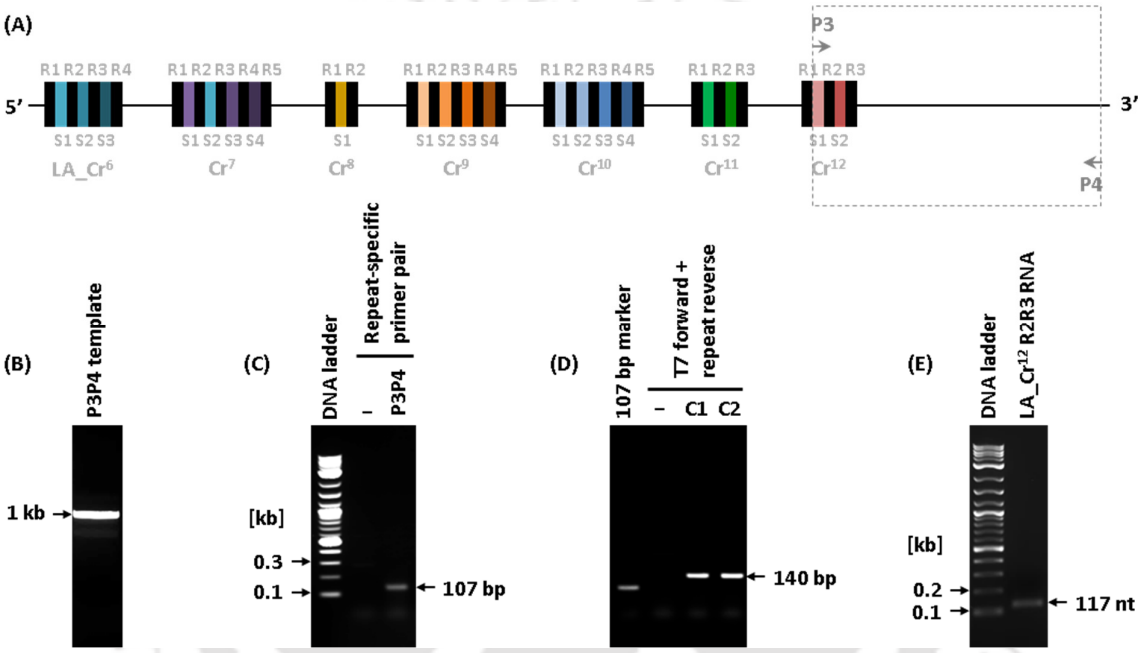


Figure 3.9. *In vitro* synthesis of the miniature LA_Cr^{12} (R2R3) RNA. (A) Schematic of CRISPR I-B locus in sv. Lai. The architecture of CRISPR I-B locus of sv. Lai is drawn to scale in the direction of the CRISPR array (5′-3′). Arrowheads and dashed grey box indicate the binding regions of forward (P3) and reverse (P4) primer pair and amplified fragment in PCR. (B) Generation of P3P4 DNA template. The PCR-amplified P3P4 DNA template (958 bp), with P3 and P4 primer pair using genomic DNA template of sv. Lai was resolved on 1% agarose gel for gel purification. (C) PCR amplification of array LA_Cr^{12}. Using the P3P4 DNA template and repeat-specific primer pair, the PCR-amplified product was resolved on 1% agarose gel. (D) Confirmation of LA_Cr^{12} R2R3 clone in pTZ57R/T vector. The transformants *E. coli* DH5α colonies (C1 and C2) were screened in culture PCR with vector-specific T7 forward and repeat-specific reverse primer pair. "–" indicates the negative control (without DNA template) of PCR. (E) *In vitro* synthesized miniature LA_Cr^{12} RNA (117 nt) was resolved on 1% agarose gel.

In *L. interrogans* sv. Lai, we have observed a continuous transcription of CRISPR I-B arrays (from LA_Cr^{6} to Cr^{12}), including the inter-array regions (Chapter 2, section 2.2.4). Therefore, an ideal substrate to understand rLinCas6-mediated crRNA biogenesis in sv. Lai would be a

pre-crRNA containing sequences of subsequent CRISPR arrays separated by corresponding inter-array region. However, for cloning to synthesize the respective transcript, PCR amplification of such DNA fragments using repeat-specific primers would require optimization in primer sequences and annealing temperatures. This is due to the presence of many similar repeat segments within the desired DNA fragment. In this study, thus, *in vitro* transcribed pre-crRNA substrates are limited to the transcripts derived from a single CRISPR array (LA_Cr^{6} R2R4 and LA_Cr^{12} R2R3). Further work is required in future to synthesize precursor transcript containing CRISPR arrays and inter-array regions for a comprehensive understanding of LinCas6-mediated crRNA biogenesis in *L. interrogans* sv. Lai.

3.3 Conclusion

In this chapter, we cloned the *cas* gene *LIC10939*, *LIC0935* and *LIC10938* encoding for LinCas6, LinCas5, and LinCas3, respectively, in *L. interrogans* sv. Copenhageni. To characterize the LinCas6, LinCas5, and LinCas3, a recombinant form of these proteins was overexpressed heterologously and purified using affinity chromatography. Against these recombinant Cas proteins, polyclonal antisera were raised in mice to utilize them as a tool in the characterization of corresponding Cas proteins. To employ specific RNA substrates in RNase assays, CRISPR I-B arrays of sv. Copenhageni and Lai were cloned. Using the vector-array plasmid constructs, pre-crRNAs were *in vitro* synthesized.

The gene *LIC10939*/*cas6* was cloned in the pET28a vector between *Nhe*I and *Xho*I restriction sites. Using the pET28a-*cas6* plasmid construct, the rLinCas6 (N-terminal 6×his-tagged) was overexpressed in *E. coli* BL21. After overexpression, a soluble form of rLinCas6 was purified by the native method using affinity chromatography with a yield of ~2 mg/L.

The gene *LIC10935*/*cas5* was cloned in the pET28a vector between *Nhe*I and *Xho*I restriction sites. Using the pET28a-*cas5* plasmid construct, the rLinCas5 (N-terminal 6×his-tagged) was overexpressed in *E. coli* BL21. Almost the entire expressed rLinCas5 was insoluble; thus, it was purified by the denaturing method using affinity chromatography. However, precipitation of the purified rLinCas5 was observed during dialysis. Moreover, changing of dialysis condition (pH 6.0-10.0) could not help avoid the precipitation of rLinCas5. Due to the persistent solubility problem of 6×his-tagged rLinCas5, the gene *LIC10935*/*cas5* was cloned in the pET-SUMO vector to fuse a solubility-enhancing tag (SUMO) in rLinCas5. Overexpressed rLinCas5 (N-terminal 6×his-SUMO-tagged) was still observed in the insoluble

fraction; however, it could be dialyzed in soluble form after purification by denaturing method. The yield of purified rLinCas5 was estimated to be around ~1 mg/L. The 6×his-SUMO-tag from rLinCas5 was cleaved using SUMO-protease. However, a soluble form of untagged rLinCas5 could not be purified because of its precipitation after removing the 6×his-SUMO tag.

The gene *LIC10938*/*cas3* was cloned in the pET28a vector between *Nhe*I and *Bam*HI restriction sites. However, overexpression of the rLinCas3 (N-terminal 6×his-tagged) in *E. coli* BL21was not observed. Thus, the gene *LIC10938*/*cas3* was cloned in the pET-SUMO vector. After overexpression, a soluble form of rLinCas3 was purified by the native method using affinity chromatography. Co-purification of an additional protein (~17-23 kDa) was observed in the rLinCas3 elute. It was probably due to the cleavage of a fraction of overexpressed rLinCas3 from its N-terminal end. To eliminate this co-purified protein, affinity-purified elute was resolved in SEC, and pure rLinCas3 (yield ~1.75 mg/L) was collected at a peak corresponding to its monomeric size. After the SUMO-protease treatment of pure rLinCas3, an untagged version of rLinCas3 was also purified using affinity chromatography.

Purified recombinant Cas proteins; rLinCas6 (N-terminal 6×his-tagged), rLinCas5 (N-terminal 6×his-SUMO-tagged), and rLinCas3 (N-terminal 6×his-SUMO-tagged) were used as antigens to generate polyclonal antisera in mice. In ELISA, the detection of each antigen using generated corresponding antibodies was observed.

CHAPTER 4

Characterization of rLinCas6, rLinCas5, and rLinCas3

In the *Leptospira* CRISPR-Cas subtype I-B locus, *cas*-operon II contains genes encoding LinCas6, LinCas3, LinCas8, LinCas7, and LinCas5 proteins (Dixit, Ghosh et al. 2016). In the I-B system, Cas6 is responsible for crRNA biogenesis, whereas Cas5, Cas7, and Cas8 interact with crRNA to form a Cascade (Maier, Stachler et al. 2019). Once Cascade binds to the crRNA complementary sequence on MGEs, the signature Cas3 protein destroys the target through its combined activities of helicase and nuclease (Mulepati and Bailey 2013). Thus, in the I-B system of *Leptospira*, Cas proteins encoded by cas-operon II are presumed accountable for crRNA biogenesis and interference processes. In Chapter 2, the transcription of CRISPR I-B arrays into pre-crRNAs was deciphered in svs. Copenhageni and Lai. To understand the crRNA biogenesis and initiation of Cascade formation in *Leptospira*, rLinCas6 and rLinCas5 were characterized in this chapter. Before utilizing endogenous Cascade-Cas3 for CRISPR-based applications like genome editing in *Leptospira*, the functionality of LinCas3 needs to be elucidated. Hence, biochemical characterization of nuclease activity in rLinCas3 was performed to understand the physiological requirements of LinCas3 in the CRISPR immunity of *Leptospira*.

4.1 Materials and Methods

4.1.1 Oligonucleotide substrates used in activity assays in this study

Oligonucleotide substrates used in this study **(Table 4.1)** were outsourced from IDT (Integrated DNA Technologies). Sequences of these oligonucleotides were based on CRISPRCasdb-defined repeat/spacer sequences of the array LA_Cr^{2}. In addition, *in vitro* synthesized unlabeled pre-crRNAs of full-length LIC_Cr^{2} and miniature LA_Cr^{6} (R2R4) and LA_Cr^{12} (R2R3) were used as substrates in this study. To generate the unlabeled mature crRNA, miniature pre-crRNA of LA_Cr^{12} (0.1 µM) was incubated with rLinCas6 (1 µM) for 1 h at 37°C. After this, the processed RNA fragments, including mature crRNA, were purified via phenol-chloroform extraction and precipitation followed by re-suspension in NFW, as described previously in section 3.1.6.

Table 4.1. Custom synthesized 5′ fluorescent-labeled RNA oligos used in the study

RNA oligos	**Sequence (5′-3′)**
Consensus repeat RNA of LIC_Cr^{2} (repeat RNA; sense)	CUGAAUAUAACUUUGAUGCCGUUAGGCGUUGAGCAC
Consensus repeat RNA of LIC_Cr^{2} (repeat RNA; antisense)	GUGCUCAACGCCUAACGGCAUCAAAGUUAUAUUCAG

LIC_Cr[3] repeat RNA consensus (sense as per CRISPRCasdb)	UUCCUAAAGAAAUAGGGAAUUUAAAAAA
LA_Cr[1] repeat RNA consensus (sense as per CRISPRCasdb)	UUCCUAAAGAAAUCGGAAAACUAC
mature LIC_Cr[2] RNA (crRNA or mature crRNA)	UUGAGCACAAGGGGAAAACAUUCGUCACCC CGUGAAAAACUUCUGAAUAUAACUUUGAU GCCGAUAGGCG

4.1.2 EMSA and SEC of DNA bound with rLinCas6

Agarose gel-based EMSA of consensus repeat DNA [unlabeled sense LIC_Cr[2] (1-5 µM)] or plasmid DNA [circular and linear pTZ57R/T vector (100 ng each)] was performed after incubation with rLinCas6 (5 µM) in the cleavage buffer [CB; 20 mM HEPES-KOH (pH 8.0), 250 mM KCl, 2 mM $MgCl_2$, and 1 mM DTT] for 1 h at 37°C. An equivalent amount of DNA was incubated in the CB under similar reaction conditions to serve as a negative control of EMSA. Reactions, including controls, were resolved on 1% agarose gel and visualized after staining with EtBr [Sisco Research Laboratories (SRL)] or SYBR Gold (Invitrogen). Size exclusion chromatography (SEC) was performed using a Superdex Increase 200 column (GE Healthcare #28-9909-44) on AKTAprime plus (GE Healthcare). The column was first equilibrated with CB and then calibrated with gel filtration molecular weight markers (Sigma-Aldrich #MWGF200), as per the manufacturer's instructions. Affinity-purified rLinCas6 (4 µM) incubated (37°C, 1 h) with or without repeat DNA (4 µM) in 1 ml of CB was resolved in the chromatography column. The EMSA experiment was repeated to verify the reproducibility of the results.

4.1.3 RNase cleavage assays with rLinCas6

Cleavage assays using ribonuclease rLinCas6 (50-4000 nM) were performed on various RNA substrates [5′ fluorescent-labeled consensus repeat RNA of LIC_Cr[2], LIC_Cr[3], and LA_Cr[1] (250 nM each), and pre-crRNA of LIC_Cr[2] R1R4, LA_Cr[6] R2R4, or LA_Cr[12] R2R3 (100 ng each)] in CB for an hour at 37°C unless stated otherwise. As a control of the assay, under a similar condition, an equal amount of RNA substrate in the absence of recombinant protein was incubated in CB. After incubation, each reaction, including control, was stopped by mixing an equal volume of 2×formamide-based loading dye (1×; 47.5% formamide, 0.01% SDS, and 0.5 mM EDTA with or without 0.01% bromophenol blue and 0.005% Xylene Cyanol) and heat denaturation for 5 min at 95°C. After that, denatured samples were resolved onto denaturing urea (8 M) 10-20% polyacrylamide gel [PAA; acrylamide:bisacrylamide-19:1) in 0.5×TBE

buffer [tris base (50 mM), boric acid (50 mM), and EDTA (1 mM) with pH 7.6], as described previously (Sokolowski, Graham et al. 2014). A low-range ss-RNA, microRNA ladder (NEB), or a mixture of three 5′ fluorescent-labeled RNA oligos (36, 28, and 24 bases) was resolved on the gel to estimate the size of the RNA fragments. After electrophoresis, the gel was visualized directly or after staining in SYBR Gold. Each RNase assay of rLinCas6 was repeated at least once, and the results were only given if repeatability was obtained.

4.1.4 Single turnover assay with rLinCas6

The purified rLinCas6 (500 nM) was incubated with 5′ fluorescent-labeled consensus repeat RNA of LIC_Cr2 (250 nM) for the various duration (1-30 min) at 37°C, as described previously (Sokolowski, Graham et al. 2014). An equal amount of substrate served as a control of the assay under a similar condition with the maximum incubation time (30 min). The reaction was stopped, resolved onto denaturing 8 M urea 20% PAA, and imaged without staining, as mentioned in section 4.1.2. Intensity of the bands observed on gel was quantified using Image Lab software (Bio-Rad). The fraction of RNA cleaved (for time point 1-10 min) vs. time plot was fitted in the exponential curve using OriginLab software. The natural logarithm of the fraction of uncleaved RNA was plotted against time (1-5 min) until saturation was attained, and the rate constant (K_{cat}) of the cleavage assay was estimated, as described previously (Jesser, Behler et al. 2019). The slope of the fitted line resulting from the plot corresponds to $-K_{cat}$. Data obtained from two independent assays were used for plotting the exponential curve and determination of K_{cat} value.

4.1.5 EMSA of repeat RNA or mature crRNA bound with rLinCas6

Increasing concentrations of the rLinCas6 (50-1000 nM) were mixed with a fixed amount of 5′ fluorescent-labeled repeat RNA of LIC_Cr2 (250 nM each) and incubated in CB for 1 h at 37°C. Similarly, 5′ fluorescent-labeled mature crRNA of LIC_Cr2 (250 nM each) was incubated with increasing concentrations of rLinCas6 (500-4000 nM). Post-incubation, native PAGE loading dye (1×; 0.5×TBE and 5% glycerol, and 10 mM EDTA with or without bromophenol blue) was mixed with each reaction and electrophoresed onto native PAA gel (10-20%) for two hours at 200 V. This step was performed in a cold room (4°C) to rule out heat generation during native-PAGE. The gels were visualized as described in section 4.1.2. This EMSA experiment was repeated to verify the reproducibility of the results.

4.1.6 SEC of rLinCas6-crRNA complex and detection of macromolecules in elutes

Around 150 µg of nickel-affinity purified rLinCas6 that was incubated (37°C, 1 h) with or without pre-crRNA (300 µg) in 1 ml of cleavage buffer was run in the equilibrated and calibrated chromatography column, as mentioned in section 4.1.1. Collected elutes (20 µl) at peak points were immunoblotted using the polyclonal antibodies against rLinCas6, as described elsewhere (Ghosh, Prakash et al. 2018). Briefly, elutes fractions were resolved on 12% SDS-PAGE, transferred on the nitrocellulose membrane, and probed with mouse anti-LinCas6 immune serum (1:1000). The immunoblot was developed by adding secondary antibodies [HRPO-conjugated anti-mice antibodies (1:5000)] and enhanced chemiluminescence. For detecting crRNAs in the SEC elute, the collected fraction was concentrated (10×) through Corning Spin-X UF Concentrator (MERCK), and 20 µl of concentrated elute was resolved onto denaturing 8 M urea 20% PAA, as described in the section 4.1.2.

4.1.7 RNase assay of wild-type (WT) rLinCas6 and rLinCas6^{H38A}

Increasing concentrations (0.05-5 µM) of rLinCas6 or rLinCas6^{H38A} were incubated separately with 5′ fluorescent-labeled consensus (sense and antisense) repeat RNA (250 nM) or unlabeled pre-crRNA (100 ng) of LIC_Cr^{2} for 1 h at 37°C. Reactions after incubation were terminated, resolved on 10-20% denaturing PAA gel, and imaged directly or after staining in SYBR Gold, as described in section 4.1.2. Each RNase assay of rLinCas6^{H38A} was repeated to verify the reproducibility of the results.

4.1.8 RNase assay of rLinCas5

The rLinCas5 (2 µM) was incubated with 100 ng of circular and linear plasmid DNA (pTZ57R/T), *luciferase* RNA, or pre-crRNA (LIC_Cr^{2} R1R4) in CB buffer for 1 h at 37°C. As a control of the assay, under a similar condition, an equal amount of DNA/RNA substrate in the absence of recombinant protein was incubated in CB. Post-incubation, reactions, and controls were resolved on agarose or denaturing PAA gels.

4.1.9 EMSA of pre-crRNA and mature crRNA incubated with rLinCas5

EMSA of pre-crRNA [LA_Cr^{12} R2R3 (117 nt, 100 ng)], unlabeled mature crRNA (100 ng), or 5′ fluorescent-labeled mature crRNA (250 nM) incubated (1 h, 37°C) with rLinCas5 (1-4 µM)

was performed on native 8-15% PAA gel. Electrophoresis was performed in a cold room (4°C) to prevent heat generation during native PAGE. The gels were visualized as described in section 4.1.2. These EMSA experiments were repeated to verify the reproducibility of the results.

4.1.10 EMSA of crRNA bound with rLinCas6 and rLinCas5, and immunoblotting

First, the reaction (5 µl) containing rLinCas6 (0.5 and 1 µM) and 5′ fluorescent-labeled crRNA (500 nM) was incubated in CB for 30 min at 37°C. Then increasing concentration (0.5-2 µM) of rLinCas5 was added to the reaction containing the rLinCas6-crRNA complex. After an additional 30 min incubation at 37°C, 1 µl of 10× native PAGE loading dye was added to the reaction and electrophoresed immediately on native 10% PAA gel. For immunoblot analysis, a reaction of rLinCas6 (0.5 µM) and crRNA (500 nM), to which 0.5 µM of rLinCas5 was added, was scaled up (from 10 to 30 µl) and native PAGE was performed. Following this, the native gel was incubated (at room temperature) with SDS-PAGE running buffer and transferred to a nitrocellulose membrane in transfer buffer containing tris base (5 mM), glycine (190 mM), and methanol (20%), as described previously (Roelofs, Suppahia et al. 2018). The membrane was probed with anti-LinCas6/LinCas5 (1:1000), and a blot was developed, as described in section 4.1.6. This overall experiment was repeated to verify the reproducibility of the results.

4.1.11 Nuclease activity assay of rLinCas3

The rLinCas3 (250 nM) was incubated with 100 ng of circular ss-DNA [ΦX174 virion DNA (NEB)], circular ds-DNA (isolated pTZ57R/T plasmid from *E. coli* DH5α), linear ds-DNA [linearized pTZ57R/T plasmid using *Kpn*I (NEB) and purified using gel extraction kit (Thermo Scientific)] or *luciferase* mRNA [*in vitro* synthesized using RNA synthesis kit (NEB)]. These reactions were performed in nuclease activity buffer (NAB; 10 mM Tris-HCl pH 7.5 and 60 mM KCl) in the absence or presence of metal ions [Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} (10 mM each)]. To elucidate the metal ion concentration for optimal rLinCas3 nuclease activity, 1 µM of rLinCas3 was incubated with 100 ng of circular ss-DNA in NAB containing 1 µM-100 mM of Ni^{2+}/Mg^{2+} ion. The concentration-dependent nuclease assays of rLinCas3 (50 or 100-500 or 1000 µM) were performed on nucleic acid substrates (circular ss-DNA, circular/linear ds-DNA, *luciferase* mRNA, or DNA/RNA oligonucleotides) in NAB in the presence of Ni^{2+}/Mg^{2+} ion (2 mM). After incubation for 1 h at 37°C, DNA loading dye was

added in each reaction, including control, and samples were resolved on 1% EtBr mixed agarose gel. Each assay for the biochemical characterization of rLinCas3 was repeated to verify the reproducibility of the results.

4.1.12 Bioinformatics analysis

The secondary structure of repeat RNA consensus with minimum free energy (MFE; ΔG°) was evaluated in the RNAalifold web server (Bernhart, Hofacker et al. 2008). The Clustal Omega program (Madeira, Park et al. 2019) was used for generating identity matrices and multiple sequence alignments of LinCas6/LinCas3 proteins with their respective orthologs. MSAs were visualized using Jalview (Waterhouse, Procter et al. 2009).

4.2 Results and Discussion

4.2.1 Characterization of rLinCas6

4.2.1.1 LinCas6 is a DNA-binding protein

The activity of Cas6 on DNA is not explored in the existing literature due to the specialized role of Cas6 in the expression stage of CRISPR immunity. Thus, the activity of rLinCas6 was first investigated on repeat DNA (single-stranded and unlabelled). For this purpose, the rLinCas6 (4 µM) was incubated (1 h, 37°C) with repeat DNA (consensus repeat of LIC_Cr[2]; 1 µM) in nuclease activity buffer (NAB). In the presence of rLinCas6, electrophoretic migration of repeat DNA was restricted on agarose gel **(Figure 4.1A, left panel)**. The addition of SDS (denaturing agent) in the DNA-protein reaction mixture resulted in retrieval of the repeat DNA mobility, similar to that of the control. Incubation of the equivalent amount of rLinCas6 with an increasing concentration of repeat DNA (2-5 µM) resulted in a migration shift of repeat DNA on agarose gel **(Figure 4.1A, right panel)**. Such restriction and shift in migration of DNA indicated that rLinCas6 has a binding affinity towards repeat DNA. To understand the rLinCas6 binding specificity to the DNA substrate, the rLinCas6 (4 µM) was incubated with the non-specific DNA (circular or linear plasmid DNA; 100 ng). On gel electrophoresis of the DNA-protein mixture, a similar arrest in the mobility of the circular **(Figure 4.1B, left panel)** or linear ds-DNA **(Figure 4.1B, right panel)** was detected. These results suggested that rLinCas6 possesses a non-specific binding affinity to DNA.

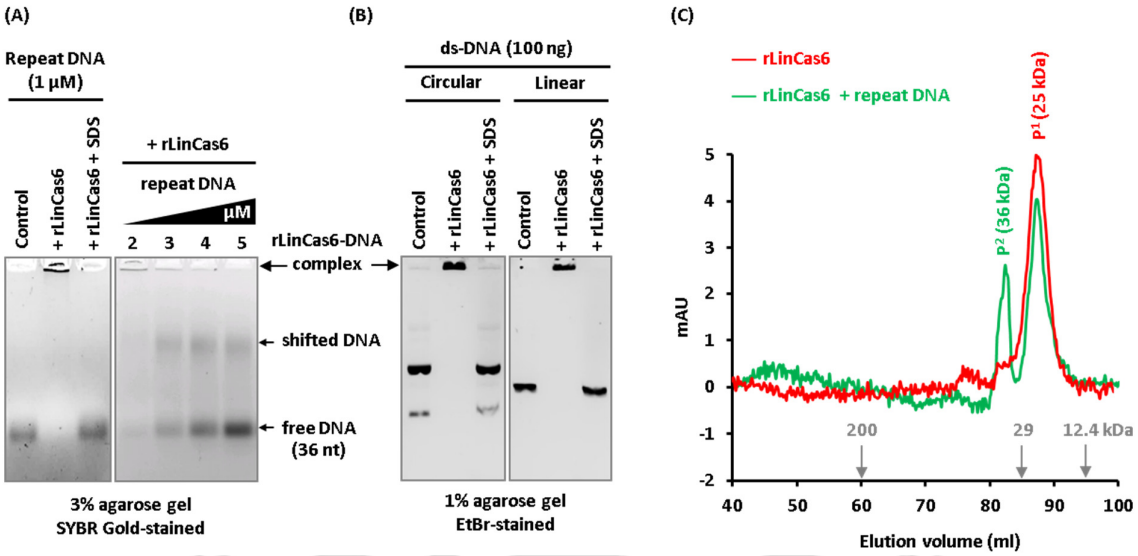


Figure 4.1. Binding of rLinCas6 to DNA. (A) The rLinCas6 binds to cognate repeat DNA. One micromolar of repeat DNA incubated with rLinCas6 (4 µM) for an hour was resolved onto an agarose gel (left panel). Retention of nucleoprotein complex was observed in gel pocket after agarose gel electrophoresis. Migration of repeat DNA got restored after the addition of SDS, similar to that in control (no protein). When higher concentrations of repeat DNA (2-5 µM) were included with the rLinCas6, a proportion of repeat DNA demonstrated a shift in migration on agarose gel (3%) electrophoresis (right panel). (B) The rLinCas6 binds to non-specific plasmid DNA. On 1% agarose gel, migration of circular (left panel) and linear (right panel) plasmid DNA got restricted when incubated with rLinCas6 (4 µM) for 1 h at 37°C. The addition of SDS in reaction after incubation restored the migration of plasmid DNA. (C) Size exclusion chromatography of rLinCas6 and repeat DNA complex. The pure rLinCas6 (300 µg) or rLinCas6 (500 µg) incubated with repeat DNA (20 µl of 100 µM) was resolved in SEC. Running profiles (mAU at 280 nm vs. elution volume) of pure rLinCas6 (red) and rLinCas6-repeat DNA (green) indicate their elution at P1 (~25 kDa) and P2 (~36 kDa) peaks, respectively.

To validate the formation of the rLinCas6-DNA complex, size exclusion chromatography of rLinCas6 incubated with or without repeat DNA was performed. Upon SEC, the resolved product of rLinCas6 and repeat DNA reaction displayed two peaks at ~82 and 87 ml elution volume (**Figure 4.1C**). Based on the standard molecular weight marker used in SEC, the peak obtained at ~82 ml (~36 kDa) indicated the elution of rLinCas6-DNA complex (~25+11 kDa). Elution of the pure and free rLinCas6 (~25 kDa) at the peak of ~87 ml was observed. This SEC analysis validated that rLinCas6 binds to DNA. In the type I-C system, Cas5d replaces the role of Cas6 in crRNA biogenesis. Cas5d from *Streptococcus pyogenes* (SpCas5d) and *Xanthomonas oryzae* (XoCas5d) have been demonstrated to interact with non-specific DNA through agarose-based electrophoretic mobility shift assay. Based on this, it was suggested that in addition to the expression stage of CRISPR immunity, Cas5d may participate in the

adaptation and/or interference stages (Koo, Ka et al. 2013). Similarly, we speculate that LinCas6 may have a role other than crRNA biogenesis in *Leptospira*.

4.2.1.2 LinCas6 cleaves the cognate repeat RNA (sense) canonically

The Cas6 protein of a CRISPR-Cas system (mostly in type-I and -III) canonically cleaves the cognate repeat sequences, independent of metal-ion and ATP (adenosine triphosphate), to generate mature crRNAs from precursor CRISPR transcript (Brouns, Jore et al. 2008; Carte, Wang et al. 2008; Carte, Pfister et al. 2010; Gesner, Schellenberg et al. 2011; Jore, Brouns et al. 2012; Reimann, Alkhnbashi et al. 2017; Jesser, Behler et al. 2019). This orthodox property of Cas6 was exploited to validate the functional orientation of the CRISPR I-B array in *Leptospira*. Thus, cleavage or nuclease activity of the rLinCas6 was investigated on the commercially synthesized CRISPR repeat RNA **(Table 4.1)** substrate of 36 nt (5′ fluorescent-labeled). The sequence of this repeat RNA was a consensus RNA sequence of LIC_Cr[2] repeats in the sense direction, as previously proposed through RT-PCR. The rLinCas6 (4 µM) was incubated (1h, 37°C) with 5′ fluorescent-labeled repeat RNA (250 nM), and the reaction was resolved on denaturing urea polyacrylamide gel. Analysis of gel without staining showed an excised 5′ fluorescent-labeled fragment of 28 nt, suggesting that the CRISPR repeat substrate was cleaved upstream to 8 nt from its 3′ end **(Figure 4.2A, left panel)**. In addition, heat-denatured (Δ; 95°C for 5 min) rLinCas6 was inactive on repeat RNA, suggesting that the active rLinCas6 only is responsible for the observed cleavage of repeat RNA. Such cleavage pattern of CRISPR repeat was also recorded through *in vitro* assay using other cognate Cas6 orthologs (Charpentier, Richter et al. 2015; Behler and Hess 2020). The predicted structure of repeat RNA ($\Delta G° = -9.05$ kcal/mol) exhibited a stable hairpin-stem loop due to the presence of the palindromic sequence **(Figure 4.2A, bottom panel)**. It hinted at the cleavage of repeat RNA towards the 3′ end (downstream) of the stem (after G28) by rLinCas6.

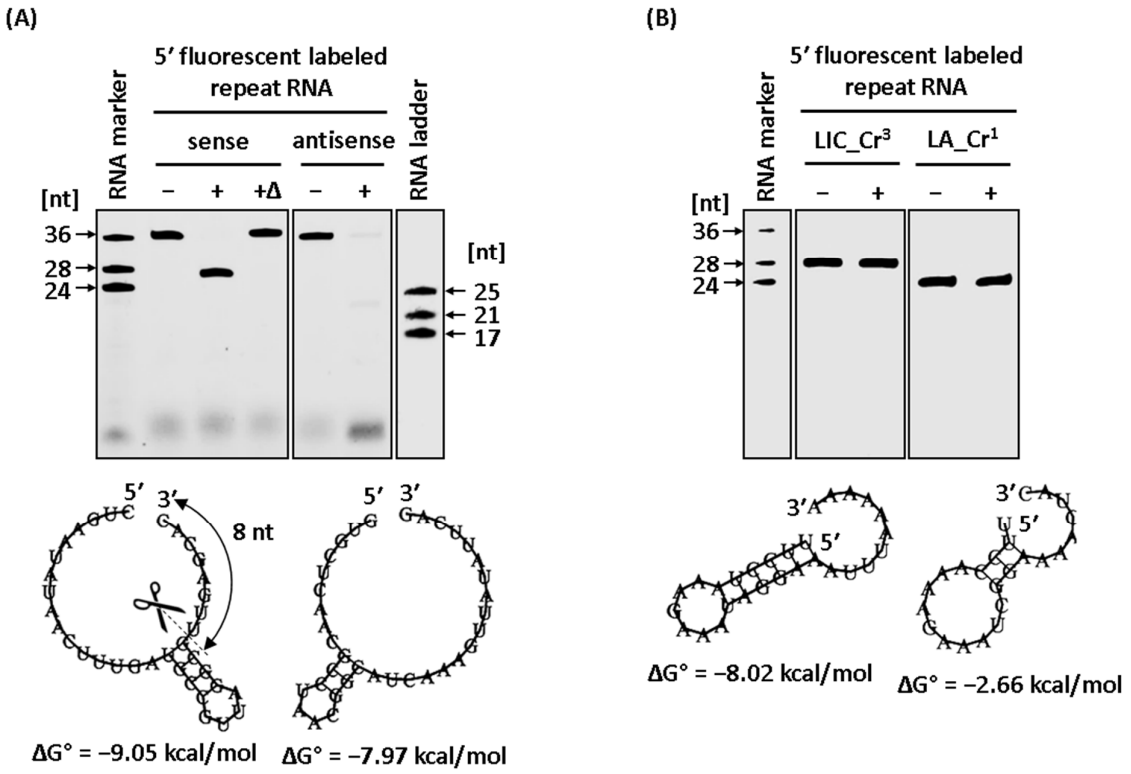


Figure 4.2. The activity of rLinCas6 on 5' fluorescent-labeled synthetic repeat RNAs. (A) *In vitro* cleavage of sense and antisense LIC_Cr[2] repeat RNA by rLinCas6. Synthetic 5' fluorescent-labeled sense repeat RNA (250 nM) was incubated with (+) or without (−) rLinCas6 (4 μM) for 1 h, resolved onto denaturing 8 M urea 20% PAA gel, and visualized under UV light. The rLinCas6 cleaves the repeat RNA upstream of 8 nt from its 3' end, laying at the stem-loop end (after G28). The heat-denatured (Δ) rLinCas6 did not show cleavage of repeat RNA (left panel). The secondary structure of the sense repeat RNA with cleavage site and estimated minimum free energy (ΔG° = −9.05 kcal/mol) is shown below the PAA gel image. The cleavage site within repeat RNA is pointed by a scissor and dashed line. Synthetic 5' fluorescent-labeled antisense repeat RNA (ΔG° = −7.97 kcal/mol) incubated with rLinCas6 (4 μM) was resolved onto denaturing urea gel. In contrast to the cleavage of sense RI-B RNA, distinct cleavage at multiple positions and degradation of antisense repeat RNA substrate were observed (right panel). (B) The rLinCas6 activity was not observed on synthetic 5' fluorescent-labeled sense repeat RNA consensus of non-cognate CRISPRs; LIC_Cr[3] (ΔG° = −8.02 kcal/mol) (left panel) and LA_Cr[1] (ΔG° = −2.66 kcal/mol) (right panel). Gel images are spliced for labeling purposes. All incubation steps were carried out in the nuclease activity buffer for 1 h at 37°C.

Two classes of CRISPR repeats, canonical and non-canonical, have been described previously (Sefcikova, Roth et al. 2017). Many repeats display a clear palindromic feature that would result in a stable stem-loop structure immediately preceding the cleavage site and is categorized in canonical class (−19.0 kcal/mol < ΔG° < 4.8 kcal/mol) (Sefcikova, Roth et al. 2017), as observed for LIC_Cr[2] repeat RNA **(Figure 4.2A, bottom panel)**. The non-canonical class

includes relaxed repeats (unstructured, $\Delta G^\circ > 0$ kcal/mol) that lack the characteristic stem-loop structure or form short stem-loops at variable locations relative to the predicted site of cleavage (Sefcikova, Roth et al. 2017). Most of the subtype CRISPR I-A and -B repeats were reported to be relaxed, and therefore, Cas6 proteins of these subtypes are supposed to remodel the repeat to form the required hairpin structure to reposition the cleavage site (Wang, Preamplume et al. 2011; Shao and Li 2013; Shao, Richter et al. 2016; Sefcikova, Roth et al. 2017). Notably, the non-canonical class of repeats appears more often in archaea or thermophiles, suggesting an impact of the environment on the Cas6-mediated CRISPR RNA processing (Sefcikova, Roth et al. 2017). Thus, here reported CRISPR I-B repeat of *Leptospira* does not seem to depend on remodeling by LinCas6 and can be further ascertained on structural analysis. Interestingly, the antisense repeat RNA substrate was cleaved distinctly and less efficiently by the rLinCas6 (**Figure 4.2A, right panel**). These results agree with the proposed orientation of the CRISPR array by RT-PCR of the I-B locus in *Leptospira*.

To assess the specificity of the LinCas6 for CRISPR repeat RNA, additional CRISPR repeats (sense direction) of the two computationally predicted arrays (LIC_Cr^3 and LA_Cr^1) were analyzed for cleavage by LinCas6. No cleavage of the consensus CRISPR repeat substrates of the array LIC_Cr^3 **(Figure 4.2B, left panel)** or LA_Cr^1 **(Figure 4.2B, right panel)** was detected. It agrees with a previous study ascertaining all repeat RNAs may not be embraced as substrates by the non-cognate Cas6 (Reimann, Alkhnbashi et al. 2017). It was suggested that the Cas6 proteins have co-evolved with the CRISPR repeat sequences. To interact with specific crRNAs, Cas6 proteins have to adapt to their respective repeat sequences (Wang, Zheng et al. 2012; Reimann, Alkhnbashi et al. 2017).

4.2.1.3 LinCas6 cleaves cognate repeat RNA in a single turnover mode

The Cas6 protein acts as a single or multiple turnover enzyme depending on the high or low affinity of the protein, respectively, toward cleaved CRISPR repeat RNA (Behler and Hess 2020). To study the functioning mode of the LinCas6, a time-dependent assay (0-30 min) of rLinCas6 (500 nM) was performed to investigate the cleavage kinetics on the CRISPR repeat RNA substrate (250 nM). Denaturing PAGE analysis of the reactions suggested time-dependent (1-30 min) cleavage of the given amount of CRISPR repeat RNA substrate and the endoribonuclease rLinCas6 **(Figure 4.3A)**. The fraction of CRISPR repeat RNA substrate (250 nM) cleaved by the rLinCas6 (500 nM) against a given time (1-10 min) illustrated an exponential curve **(Figure 4.3B)**. The exponential decay of the CRISPR repeat RNA substrate

indicates that LinCas6 functions in single turnover mode, as also reported previously for the other Cas6 orthologs (Haurwitz, Jinek et al. 2010; Sashital, Jinek et al. 2011; Sternberg, Haurwitz et al. 2012; Niewoehner, Jinek et al. 2014; Jesser, Behler et al. 2019). The rate constant (k_{cat}) of the LinCas6, calculated using the graph of the natural logarithm of fraction uncleaved as a function of time (0-5 min), was 0.514±0.15 min^{-1} **(Figure 4.3C)**.

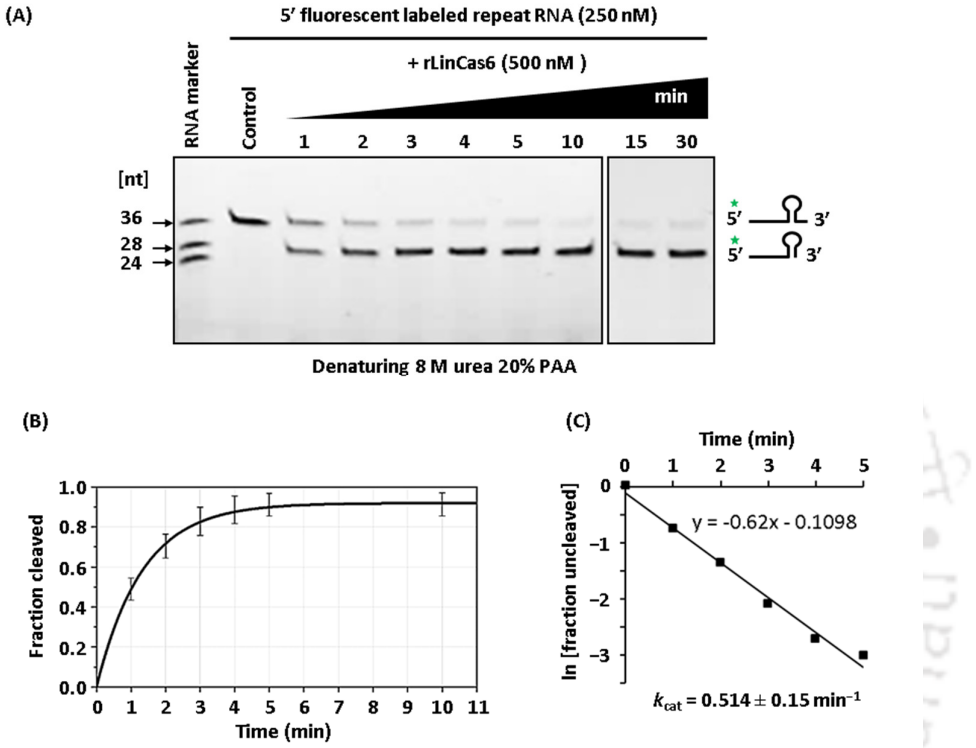


Figure 4.3. The rLinCas6 mediated cleavage of repeat RNA and pre-crRNA of LIC_Cr². (A) The rLinCas6 is a single turnover endoribonuclease on its cognate RNA repeat substrate. The 5' fluorescent-labeled repeat RNA substrate (250 nM) was incubated with rLinCas6 (500 nM) for 1-30 min. The reaction product was electrophoresed onto denaturing 8M urea 20% PAA (polyacrylamide) gel and visualized directly under UV light. Gel images are spliced for labeling purposes. All incubation steps were carried out in the nuclease activity buffer for indicated time points at 37°C. (B) The plot of substrate fraction cleaved as a function of the time graph follows the first-order kinetics where an exponential decay of the substrate was observed, suggesting a single turnover mode of cleavage of the cognate repeat RNA by rLinCas6. Error bars are indicative of two independent experiments (C) The rate constant of the cleavage assay (0.514 ± 0.15 min^{-1}) was estimated from the slope of the line obtained after plotting the natural logarithm of the fraction of uncleaved repeat RNA as a function of time (1–5 min) until saturation was observed. A graph of one of the duplicate experiments is shown.

This k_{cat} value of LinCas6 is approximately one order of magnitude lower than that for TthCas6A (3.2 min^{-1}), TthCas6B (3.7 min^{-1}), TthCas6e/Cse3 (4.9 min^{-1}) of *T. thermophilus* (Sashital, Jinek et al. 2011; Niewoehner, Jinek et al. 2014), PaeCas6f/Csy4 (~3 min^{-1}) of *P.*

aeruginosa (Haurwitz, Sternberg et al. 2012), and SsoCas6-1 (3.69 min^{-1}) of *S. solfataricus* (Sokolowski, Graham et al. 2014). However, the k_{cat} of LinCas6 is approximately one order of magnitude higher than that of the Cas6-1 (0.04-0.06 min^{-1}) of *Synechocystis* species (Jesser, Behler et al. 2019). Nevertheless, these values are several orders of magnitude lower than the RNases, such as RNase A (910 to 40500 min^{-1}) (Katoh, Yoshinaga et al. 1986). Although the exact reasons are not clear, it was suggested that CRISPR maturation enzymes such as Csy4 (PaeCas6f) of *P. aeruginosa* and Cyanobacterial Cas6-1 evolved as RNA binding proteins exhibiting highly accurate substrate selection while retaining only modest cleavage kinetics due to the lack of selection for rapid cleavage kinetics (Haurwitz, Sternberg et al. 2012; Jesser, Behler et al. 2019). Hence, here the reported LinCas6 with a moderate rate constant comes off as pertinent to the crRNA processing.

4.2.1.4 LinCas6 binds to the cleaved repeat RNA

The single turnover kinetics of the rLinCas6 endoribonuclease advocate that the enzyme conceivably remains bound to the cleaved CRISPR repeat fragment. To endorse this, an increasing concentration of rLinCas6 (50-1000 nM) was incubated with a fixed amount of the CRISPR repeat RNA (250 nM). Analysis of the reactions on denaturing urea PAGE indicated an increase in cleaved repeat RNA fragment (28 nt) in proportion to rLinCas6 concentration, and concurrently, there was a perpetual decline in band intensity of intact repeat RNA (36 nt) **(Figure 4.4A)**. A comparison of the same reaction in the equivalent amount on the native PAGE confirmed the rLinCas6 concentration-dependent (300-1000 nM) shift in the migration of cleaved $R^{I\text{-}B}$ RNA **(Figure 4.4B)**. An additional faint band of increasing intensity at 300-1000 nM of rLinCas6 was also observed **(Figure 4.4B)**, suggesting stepwise oligomerization of rLinCas6 bound to the cleaved repeat RNA. Cooperative oligomerization of MmaCas6b (Shao, Richter et al. 2016) and SsoCas6-1 (Sokolowski, Graham et al. 2014) in cleavage of their respective repeat RNAs have been illustrated previously. However, the observed oligomerization of rLinCas6 with cleaved repeat RNA suggested its role even after cleavage of the repeat RNA. Moreover, the shifting of the uncleaved RNA substrate was not detected on the native gel, perhaps due to the immediate cleavage of the substrate by rLinCas6 or due to the higher affinity of the protein towards the cleaved fragment. The rLinCas6, at a limiting concentration (100 nM), cleaved less than half the amount of repeat RNA (250 nM) in an hour **(Figure 4.4A),** and a further increase in incubation time (2-5 h) could not increase the cleaved fragment (data not shown). It suggests that rLinCas6 remains bound to the processed crRNA

and rationalizes the single turnover mode operation. In most of the I-B systems with the unstructured repeat RNA, the processed crRNA encounters additional trimming at the 3′ end, thus removing the hairpin results in dissociation of the Cas6 from crRNA (Richter, Zoephel et al. 2012; Hille, Richter et al. 2018). On the other hand, the Cas6 (I-B system) of *Haloferax volcanii* remains bound to the crRNA and participates in Cascade formation for downstream targeting (Brendel, Stoll et al. 2014; Hochstrasser and Doudna 2015). Herein, the stable association of the cleaved CRISPR repeat RNA fragment with rLinCas6 leads us to speculate that LinCas6 may form the component of Cascade. However, further studies are required to determine the role of LinCas6 in forming the interference complex in the *Leptospira*.

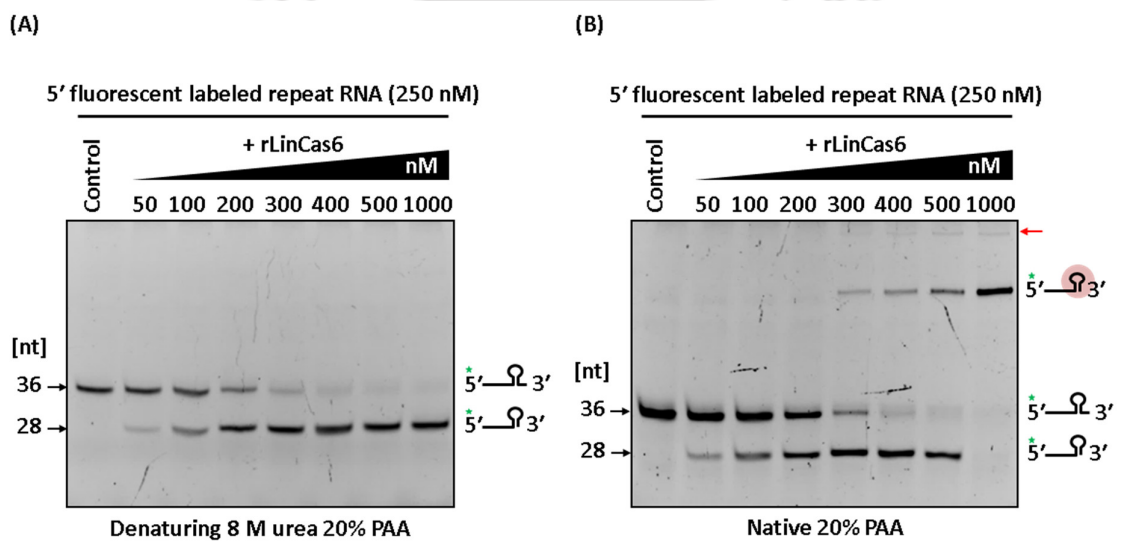


Figure 4.4. The rLinCas6 concentration-dependent cleavage of repeat RNA. (A) Urea PAGE analysis of rLinCas6 concentration-dependent cleavage of repeat RNA. The 5' fluorescent-labeled repeat RNA substrate (250 nM) was incubated with an increasing concentration of rLinCas6 (50-1000 nM). On denaturing urea PAGE, the resolved reaction products indicated an increase in cleaved repeat RNA fragment (28 nt) in proportion to rLinCas6 concentration, and concurrently, there was a perpetual decline in band intensity of intact repeat RNA (36 nt). (B) Native PAGE analysis of rLinCas6 concentration-dependent cleavage of repeat RNA. On native PAGE, the same reaction products from (A) showed a shift in migration of cleaved repeat RNA (28 nt) in proportion to the amount of rLinCas6 (300-1000 nM). An additional faint band with increasing intensity was also observed (shown by a red arrowhead) at 300-1000 nM of rLinCas6. Such a shift in the migration of cleaved products suggests that rLinCas6 remains bound to the processed product.

4.2.1.5 LinCas6 processes pre-crRNA into mature crRNAs

The generation of mature crRNAs from pre-crRNA using the Cas endoribonuclease is central to the functionality of RNA-directed CRISPR-Cas immunity. To endorse rLinCas6 mediated processing of LIC_Cr^2 transcript, the total possible transcript fragments (n=14) that could be

derived from pre-crRNA cleavage at the repeat sequences using rLinCas6 have been mapped **(Figure 4.5A)**. These transcript fragments have been categorized as two types of cleavage products; fragments (n=9; 256-78 nt) with intact repeat, and fragments (n=5; 75-8 nt) with cleaved repeat, here demarcated as incompletely processed (IP) fragments and as completely processed (CP) fragments, respectively **(Figure 4.5A)**. The CP fragments include three mature crRNAs (70-75 nt) with unique spacer RNA segments.

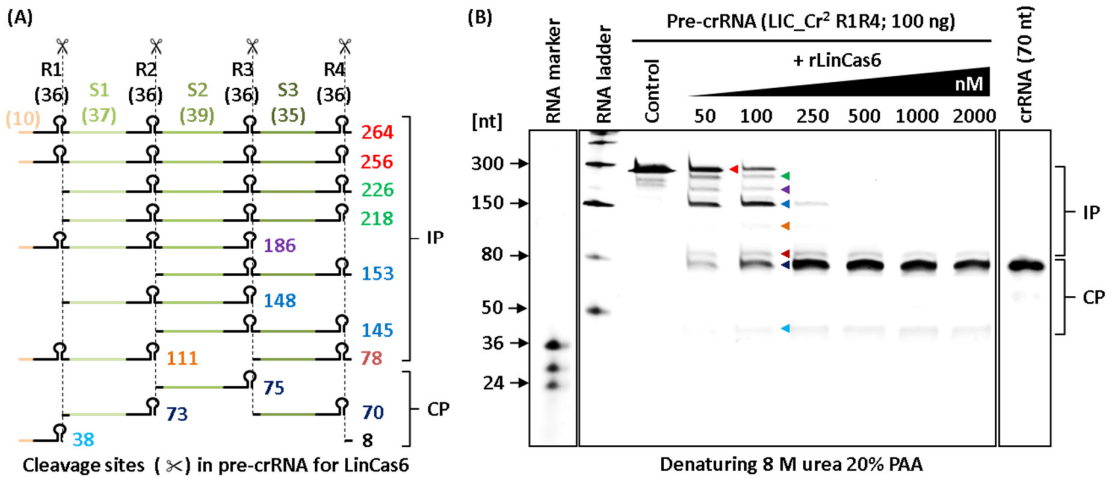


Figure 4.5. The rLinCas6-mediated canonical processing of pre-crRNA. (A) Mapping of cleavage sites for rLinCas6 within pre-crRNA. The possible pre-crRNA-derived cleaved fragments (n=14) by rLinCas6 are illustrated as incompletely processed (IP) and completely processed (CP) fragments. Scissors and vertical dashed black lines demarcate cleavage sites within repeat segments of pre-crRNA. (B) The rLinCas6 concentration-dependent processing of pre-crRNA. *In vitro* synthesized full-length LIC_Cr2 transcript (264 nt), including vector-derived additional 10 nt at its 5' end (demarcated by light orange color), was incubated (1h, 37°C) with increasing concentrations of rLinCas6 (50-2000 nM). The reaction products were analyzed on denaturing urea polyacrylamide gel after staining with SYBR Gold. At 100 nM of rLinCas6, 8 bands of different lengths were detected. Each of these bands is marked with unique color and is correlated with individual RNA fragments illustrated graphically in (A). On urea-polyacrylamide gel, a fluorescent-labeled mature crRNA (70 nt) resolved at the position of crRNAs (70-75 nt) of CP products.

After mapping the cleavage sites and processed RNA fragments from pre-crRNA, an *in vitro* cleavage assay of rLinCas6 was set up with the unlabeled full-length pre-crRNA (LIC_Cr2 R1R4; 100 ng). Denaturing urea-PAGE analysis of reactions, where pre-crRNA (100 ng) was incubated with increasing concentrations of rLinCas6, indicated a total of 8 processed RNA

fragments of different molecular lengths **(Figure 4.5B)**. Larger cleaved IP fragments (256-78 nt) of pre-crRNA were observed at an initial concentration (50-100 nM) of rLinCas6, and their intensity was found to reduce with increasing concentration of rLinCas6 (250-2000 nM). Concurrently, an increase in the CP fragments (75-38 nt) intensity was detected with an increase in rLinCas6 (50-250 nM) **(Figure 4.5B)**. The CP fragments were the most prominent at a higher concentration of rLinCas6 (250-2000 nM). Further, mature crRNA (70 nt) having terminal spacer (34 nt) of LIC_Cr^2 resolved on urea-polyacrylamide gel at the position where mature crRNAs (70-75 nt) of CP fragments were expected **(Figure 4.5B)**. This result embarks upon canonical processing of cognate pre-crRNA (LIC_Cr^2 transcript) into mature crRNAs using rLinCas6 under *in vitro* conditions.

4.2.1.6 LinCas6 and crRNA form a ribonucleoprotein complex

Before examining the binding of rLinCas6 to crRNA, the nuclease activity of rLinCas6 was investigated on crRNA. No cleavage of 5′ fluorescent-labeled crRNA (250 nM) was observed when incubated with rLinCas6 (500 nM) for 1 h at 37°C **(Figure 4.6A)**. To examine whether crRNA can bind to rLinCas6, EMSA was performed. The 5′ fluorescent-labeled crRNA (250 nM) incubated with an increasing concentration of rLinCas6 (500-4000 nM) was resolved on native PAA gel, where it displayed retardation of nucleoprotein complex versus the control (no protein) **(Figure 4.6B)**. Such migration shift suggested that the rLinCas6 binds to the mature crRNA. The mobility of crRNA on the gel was more restricted at higher concentrations (2000-4000 nM) of rLinCas6 **(Figure 4.6B)**. This sequential binding of additional rLinCas6 with crRNA suggests the cooperative oligomerization of rLinCas6.

To elucidate the number of rLinCas6 that remains bound to crRNA (70-75 nucleotides), size exclusion chromatography (SEC) of the rLinCas6 incubated with or without pre-crRNA (264 nucleotides) was performed. Based on the elution profile of standard protein markers, the pure rLinCas6 (25.5 kDa) got eluted at ~25 kDa **(Figure 4.6C),** whereas rLinCas6 incubated with pre-crRNA eluted at ~73 kDa. The molecular weight of mature crRNA (70-75 nt) was computationally (OligoCalc (Kibbe 2007)) estimated to be 23-24 kDa. The elution profiles of the pure rLinCas6 and the rLinCas6 bound crRNA suggested that two rLinCas6 (25.5×2) molecules remain bound to crRNAs (23-24 kDa), adding up the total molecular weight of the complex to 74-75 kDa. This indicated that the addition migration shift of crRNA observed at 2000-4000 nM of rLinCas6 **(Figure 4.6B)** has a stoichiometry of 2:1 (rLinCas6: crRNA).

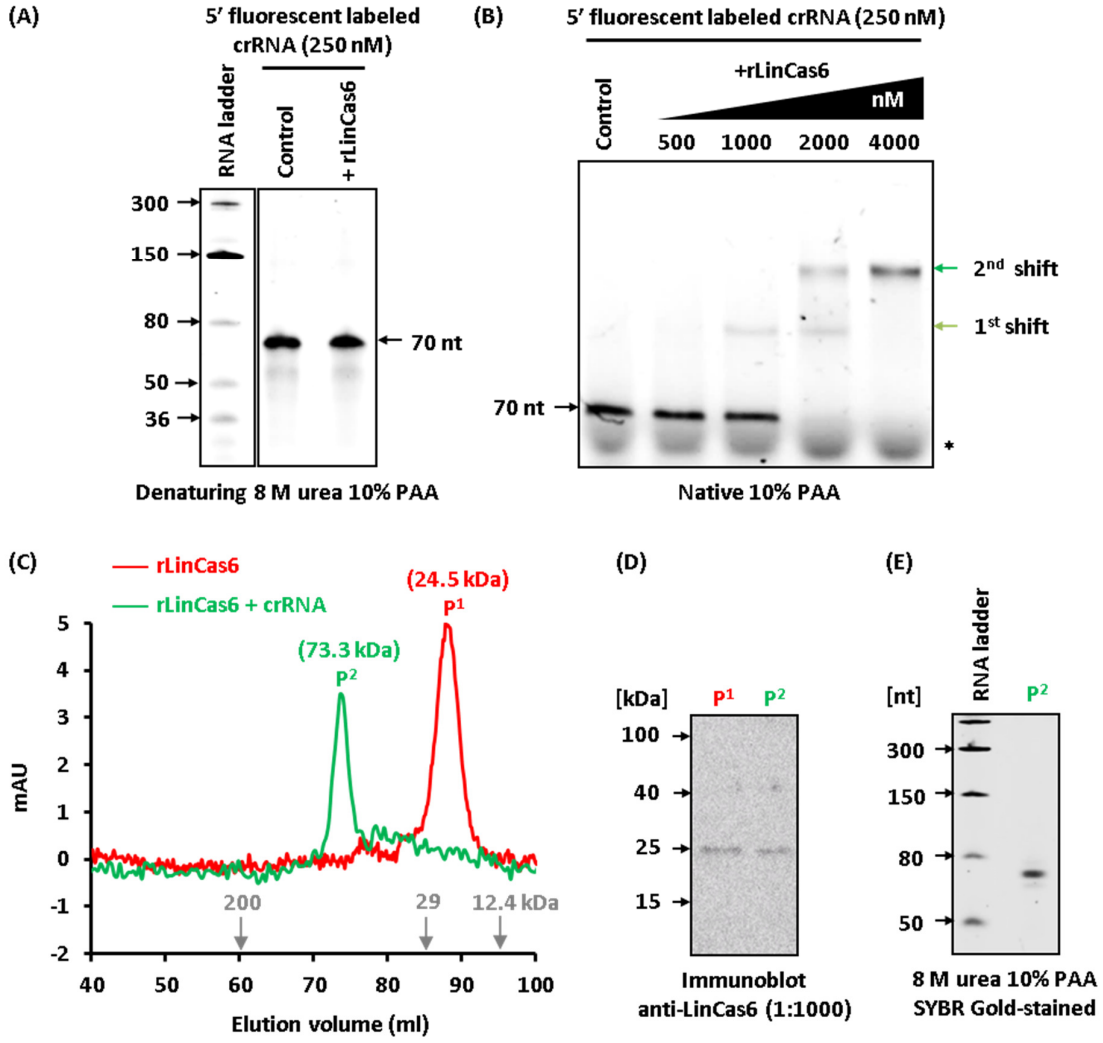


Figure 4.6. Binding of rLinCas6 to mature crRNA, SEC of the ribonucleoprotein complex, and detection of individual components in the eluted fraction. (A) Nuclease activity of rLinCas6 on mature crRNA. The rLinCas6 (500 nM) was incubated (1 h, 37°C) with 5' fluorescent-labeled crRNA (250 nM) and resolved on denaturing 10% PAA. (B) The rLinCas6 forms a stable nucleoprotein complex with the cognate mature crRNA. The electrophoretic mobility shift of mature crRNA with increasing concentrations of rLinCas6 (500-4000 nM) was detected on resolving onto native PAA gel. In contrast to the control (no protein), the mobility shift of mature crRNA (light green arrow) was detected at 1000 nM of rLinCas6. An increase in the concentration of rLinCas6 (2000-4000 nM) resulted in the additional shift of mature crRNA (dark green arrow). Asterisk (*) shows the migrated Bromophenol blue of loading dye on the native gel. (C) SEC of rLinCas6 and rLinCas6-crRNA. The rLinCas6 was incubated with or without pre-crRNA and was run in a chromatography column. Running profiles (mAU vs elution volume) of rLinCas6 (red) and rLinCas6-crRNA (green) indicated their elution at peaks; P^1 (24.5 kDa) and P^2 (73.3 kDa), respectively. (D) Immunoblotting of the collected fraction. The rLinCas6 protein was detected in each fraction collected at P^1 and P^2 through immunoblot using anti-rLinCas6 (1:1000). (E) Urea-PAGE of the SEC elute. The

concentrated (10×) fraction (P^2) was resolved onto denaturing urea PAA and stained in SYBR Gold. Detection of an RNA fragment below 80 nt of ladder confirmed the presence of crRNA in fraction eluted at P^2. All reactions, including controls, were incubated for 1 h at 37°C.

Furthermore, the immunoblot of the SEC elutes using anti-LinCas6 confirmed the presence of rLinCas6 at both peaks (~25 and 73 kDa) **(Figure 4.6D)**. In the SEC elute at ~73 kDa, the mature crRNAs at their expected size (70-75 nt) were also detected on denaturing urea PAGE **(Figure 4.6E)**. Altogether, these data illustrate that two subunits of rLinCas6 remain bound to mature crRNA and is in line with the findings of the assay where additional retardation in the mobility of mature crRNA in the presence of rLinCas6 was detected **(Figure 4.6B)**. It is known that Cas6 binds to the stem-loop of crRNA towards its 3′ end. Thus, in 1:1 stoichiometry of the rLinCas6-crRNA complex, the rLinCas6 molecule is more likely to bind to the stem-loop of crRNA. However, the binding site of additional rLinCas6 molecule in 2:1 stoichiometry of the rLinCas6-crRNA complex is unclear. Further study on Cascade-forming Cas proteins is warranted to unveil the stoichiometry of the effector I-B complex in *Leptospira*.

In most subtypes of CRISPR-Cas type I, the biogenesis of crRNAs from pre-crRNA requires Cas6 endoribonucleases, whereas subtype I-C relies on the functional variant of Cas5 for the processing of pre-crRNAs (Charpentier, Richter et al. 2015; Hochstrasser and Doudna 2015). In most I-B systems with unstructured repeat RNAs, the processed crRNAs come across further trimming at the 3′ end. It thus causes the removal of the hairpin, resulting in the dissociation of Cas6 from crRNA (Richter, Zoephel et al. 2012; Hille, Richter et al. 2018). On the contrary, Cas6 of *Haloferax volcanii* (I-B system) remains associated with the crRNA and participates in forming a Cascade for downstream targeting (Brendel, Stoll et al. 2014; Hochstrasser and Doudna 2015). In this study, stable binding of the rLinCas6 to the cleaved CRISPR repeat RNA fragment led to our speculation that LinCas6 is an integral part of Cascade. However, a study on the interference complex in *Leptospira* is required to determine the fate of LinCas6 after crRNA biogenesis.

4.2.1.7 Comparison of LinCas6 and its orthologs

The remarkable sequence diversity among the known CRISPR array processing endoribonucleases prompted us to compare the LinCas6 to its available and reported orthologs. The UniProtKB BLAST analysis of the LinCas6 amino acid sequence revealed 38 hits within the genus *Leptospira*. Out of these hits, LinCas6 orthologs within various serovars of *L.*

interrogans (Linhai, Canicola, Icterohaemorrhagiae, Bataviae, Hardjo, Lora, Lai, Grippotyphosa, Australis) demonstrated a high level of sequence conservation (94-100% sequence identity and 100% query coverage). Less conserved orthologs of LinCas6 (36-46% sequence identity and 46-92% query coverage) were found in the few pathogenic (*L. weilii*, *L. santarosai, L. alstonii, and L. kirschneri*) and in an intermediate-pathogenic *Leptospira* species (*L. broomii*). Notably, one of the LinCas6 orthologs from *L. interrogans* sv. Linhai (UniProtKB entry: A0A0C5XAU5) was annotated ATP-dependent RNA helicase and was found to be fused with Cas3, a signature Cas protein of type-I that is recruited by Cascade for progressive degradation of the target DNA (Westra, van Erp et al. 2012). To our belief, the natural fusion partner of Cas6 has not been documented to date. However, Cas3 fusion with Cse1 (CasA; homolog of Cas8), a large subunit protein of Cascade (crRNP complex) that function in target DNA selection, is recorded in *Streptomyces* sp. SPB78, *Streptomyces griseus*, and *Catenulispora acidiphila* DSM 44928 (Westra, van Erp et al. 2012). Based on this fusion of Cas3 and Cse1, it was proposed that the Cse1 subunit may furnish a docking site for the association of standalone Cas3 known to interact with the Cascade directly (Westra, van Erp et al. 2012). We speculate that LinCas6 may be an integral part of Cascade I-B. This may be an interesting subject for future study to investigate the docking site in LinCas6 bound crRNA for the Cascade interacting standalone LinCas3.

Next, a multiple sequence alignment (MSA) was performed to compare the LinCas6 protein with the Cas6 orthologs and a Cas5d from a previous study (Jesser, Behler et al. 2019). These endoribonucleases belong to different archaea and bacteria like *Sulfolobus solfataricus* (SsoCas6-1A, SsoCas6-1B, and SsoCas6-3) (Lintner, Kerou et al. 2011; Shao and Li 2013; Sokolowski, Graham et al. 2014); *Thermus thermophiles* (TthCas6A, TthCas6B, and TthCas6e) (Sashital, Jinek et al. 2011; Niewoehner, Jinek et al. 2014); *Methanococcus maripaludis* (MmpCas6b) (Richter, Zoephel et al. 2012); *Methanosarcina mazei* (MmzCas6b-IB) (Nickel, Ulbricht et al. 2019); *Pyrococcus furiosus* (PfuCas6) (Wang, Preamplume et al. 2011); *Pseudomonas aeruginosa* (PaeCas6f) (Haurwitz, Jinek et al. 2010; Haurwitz, Sternberg et al. 2012); and *Bacillus halodurans* (BhaCas5d) (Nam, Haitjema et al. 2012). The enlisted Cas6 orthologs are associated with the CRISPR-Cas subtypes I-A, -B, -C, -E, and -F, as well as III-B. The identity matrix of the alignment suggests that the percentage of identical residues in LinCas6 with its characterized Cas6 orthologs is in the range of 13-25% **(Figure 4.7A).**

(A)

1	LinCas6 (210 residues)	100										
2	SsoCas6-1A (289 residues)	18.07	100									
3	SsoCas6-1B (287 residues)	18.90	89.20	100								
4	SsoCas6-3 (261 residues)	13.25	27.73	29.41	100							
5	TthCas6A (239, residues)	18.9	16.58	16.67	19.21	100						
6	TthCas6B (264 residues)	25.42	19.53	19.16	21.5	32.77	100					
7	TthCas6e (211 residues)	19.47	10.56	10.56	14.29	24.2	24.12	100				
8	MmpCas6b (218 residues)	17.31	18.79	19.51	16.78	13.18	13.97	13.04	100			
9	MmzCas6b-IB (216 residues)	18.27	20	21.34	18.79	20.31	17.78	17.65	43.06	100		
10	PfuCas6-1 (264 residues)	18.24	16.02	16.67	13.41	18.86	20.6	17.42	16.18	13.97	100	
11	PaeCas6f (187 residues)	21.49	11.72	9.45	15.25	21.21	23.03	28.09	10.64	13.83	16.54	100
12	BhaCas5d (236 residues)	19.13	15.07	15.28	12.78	15.79	20.69	15.17	13.56	14.41	15.04	16.67

(B)

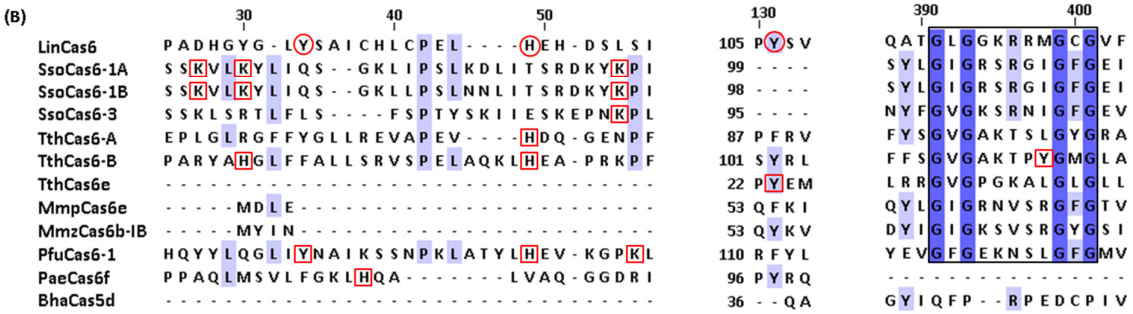


Figure 4.7. Multiple sequence alignment (MSA) of LinCas6 and other known CRISPR maturation proteins. Protein sequences of the Cas6/Cas5d orthologs used in this alignment are available in the UniProtKB database with the following entry names: Q72TS8 (LinCas6), Q97WV8 (SsoCas6-1A), Q97Y96 (SsoCas6-1B), A0A0E3GTV5 (SsoCas6-3), Q5SM65 (TthCas6A), Q53VU8 (TthCas6B), Q53WG9 (TthCas6e), A4FXZ3 (MmpCas6b), Q8PZD4 (MmzCas6b-IB), Q8U1S4 (PfuCas6-1), Q02MM2 (PaeCas6f), and A0A4Y7WTU1 (BhaCas5d). (A) The identity matrix of the alignment revealed tremendous diversity of LinCas6 among the selected Cas6 orthologs and Cas5d. The aforementioned Cas6 orthologs and the LinCas6 were aligned using Clustal Omega, and the identity matrix was generated. Values given in the cells of the matrix are the percent identity of respective proteins with LinCas6 (B) MSA of LinCas6 and its other orthologs. The N-terminal of the Cas6 proteins displays different degrees of conservation, as highlighted by multiple blue shading. The potential active site residues (Lys, His, Tyr) of Cas6 orthologs are boxed in the red square. The predicted conserved active site residues (Tyr, His) in LinCas6 after alignment are red-circled. The motif (GxGxxxxGxG) is a glycine-rich loop (G-loop) boxed in a black rectangle and is found to be conserved in the LinCas6.

In MSA **(Figure 4.7A)**, the N-terminal of the LinCas6 showed three residues (Tyr27, His38, and Tyr106) that aligned with either the active site residues of at least one or multiple Cas6 orthologs. The Tyr27 residue of LinCas6 showed conservation with catalytic Tyr of the known PfuCas6-1 (Wang, Preamplume et al. 2011). Similarly, His38 of LinCas6 aligned with catalytic His of TthCas6-A (Niewoehner, Jinek et al. 2014), TthCas6-B (Niewoehner, Jinek et al. 2014),

and PfuCas6-1 (Wang, Preamplume et al. 2011), whereas Tyr106 of LinCas6 aligned with catalytic Tyr of TthCas6e (Sashital, Jinek et al. 2011) **(Figure 4.7B)**. Based on MSA, the three residues (Tyr27, His38, and Tyr106) in LinCas6 may be the potential active site for endonuclease activity. Among others, the conserved arginine, a positively charged amino acid in the N-terminal part of the aligned Cas6 proteins **(Figure 4.7B)**, may be involved in RNA-protein interactions or functions in the catalytic center, as proposed previously for the cyanobacterial Cas6 protein (Jesser, Behler et al. 2019). In the C-terminal region of the majority of the aligned ribonucleases including LinCas6, a previously described glycine-rich-loop (G-loop) (Sefcikova, Roth et al. 2017) was also found to be conserved. The G-loop of Cas6 is crucial for its folding (Wang, Preamplume et al. 2011) and RNA recognition (Wang, Zheng et al. 2012).

4.2.1.8 Recombinant LinCas6 of serovar Copenhageni processes the CRISPR I-B transcripts of serovar Lai

Using the endoribonuclease activity of rLinCas6 of sv. Copenhageni, generation of mature crRNAs from pre-crRNA (LIC_Cr^2) of sv. Copenhageni has been demonstrated **(Figure 4.5B)**. LinCas6 protein (210 residues) in sv. Copenhageni and Lai (LIC10939 and LA3189, respectively) share 96.2% sequence identity with 100% query coverage. Pairwise alignment of LIC10939 (LinCas6) and LA3189 proteins revealed eight amino acid residues mismatches in LA3189 (T36 and 124, Q65 and 146, I143, K149, V184, and S202). These mismatches were not observed at the potential catalytic triad and in the G-loop of LinCas6. Therefore, whether the CRISPR transcripts in sv. Lai can be processed by rLinCas6 of sv. Copenhageni to yield mature crRNAs was investigated. Previously, a miniature version of pre-crRNA has been employed in the RNase assay of Cas6 protein to demonstrate the crRNA biogenesis (Reimann, Alkhnbashi et al. 2017). Similarly, we set up an *in vitro* cleavage assay where a miniature version of the LA_Cr^6 RNA (LA_Cr^6 R2R4 RNA) was incubated with increasing concentrations of rLinCas6. For analysis of cleavage products, the total feasible RNA fragments (n=9) obtained after processing of pre-crRNA (188 nt)) with rLinCas6 were mapped **(Figure 4.8A, top panel)**. These RNA fragments have been grouped into incompletely processed (IP; n=5, 79-180 nt) and completely processed (CP; n=4, 8-71 nt) RNA fragments, as described previously for the LIC_Cr^2 transcript **(Figure 4.5A)**. On denaturing urea-PAGE, the miniature pre-crRNA cleavage by rLinCas6 revealed six bands of different molecular weights **(Figure 4.8A, bottom panel)**. Larger IP fragments were identified when rLinCas6 was

employed in reaction at a lower range of concentrations (50-250 nM). An increase in the band intensity of CP fragments was observed at a higher concentration of rLinCas6 (500-2000 nM). In CP products, RNA fragments of 71 nt indicate the rLinCas6-mediated crRNA biogenesis from the miniature pre-crRNA of LA_Cr^{6}. Similarly, the generation of mature crRNAs from the miniature LA_Cr^{12} transcript was also investigated. Alike the processed fragments from miniature LA_Cr^{6}, RNA fragments of 71 nt in CP products could be detected following the processing of miniature pre-crRNA of LA_Cr^{12} (**Figure 4.8B**). These *in vitro* RNase assays of rLinCas6 on the miniature LA_Cr^{6} and LA_Cr^{12} transcripts suggest that rLinCas6 may also process the transcripts corresponding to the sequence of other five arrays (LA_Cr$^{7\text{-}11}$) of sv. Lai to yield mature crRNAs.

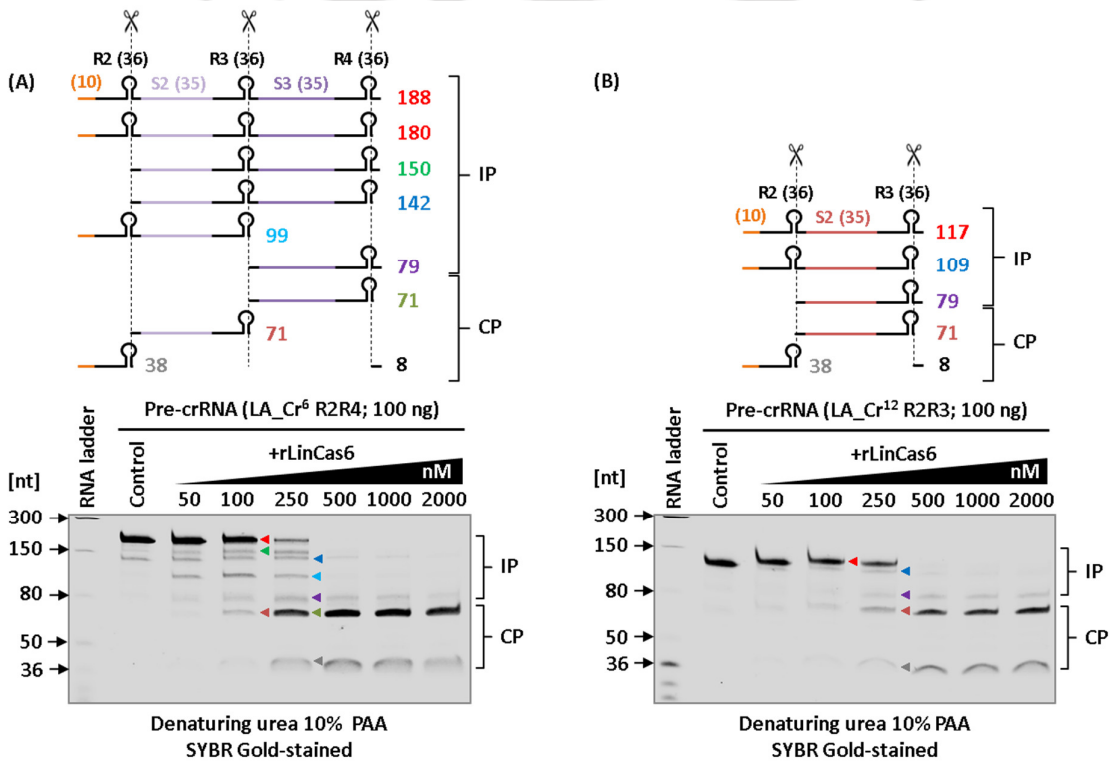


Figure 4.8. The rLinCas6 (of sv. Copenhageni)-mediated processing of miniature pre-crRNAs of sv. Lai. (A) RNase activity of rLinCas6 on LA_Cr^{6} R2R4 RNA. *In vitro* synthesized LA_Cr^{6} R2R4 transcript (188 nt) was incubated with increasing concentrations of rLinCas6 (50–2000 nM). The incompletely processed and completely processed [(IP; n=5), (CP; n=4)] fragments derived from the processing of LA_Cr^{6} R2R4 transcript by rLinCas6 are illustrated in the top panel. The reactions were analyzed on denaturing urea polyacrylamide gel after staining with SYBR Gold. Six bands of different molecular sizes, each marked with unique color, were detected at 100-250 nM of rLinCas6 (bottom panel). (B) RNase activity of rLinCas6 on LA_Cr^{12} R2R3 RNA. IP and CP fragments on the gel were mapped and indicated on the right of the gel images. Repeats and spacers are shown in black and unique colors in the pre-

crRNA outline. The orange line denotes a vector-derived additional 10 nt at its 5' end of each pre-crRNA. All reactions, including controls (no protein), were incubated for 1 h at 37 °C.

4.2.1.9 Biochemical analyses of the mutant variant of LinCas6 (LinCas6^{H38A})

In the MSA of LinCas6 and its orthologs, His38 of LinCas6 appeared to be moderately conserved with other known orthologs **(Figure 4.7B)**. Thus, the role of His38 of LinCas6 in RNA cleavage activity was investigated. The cleavage activity of the mutant endoribonuclease LinCas6 (rLinCas6^{H38A}) on the CRISPR repeat RNA substrate (sense) was compared with the rLinCas6 activity after resolving the reaction products over denaturing urea gel. The cleavage of the repeat RNA substrate by the rLinCas6^{H38A} (1 µM) was reduced compared to that by the LinCas6 **(Figure 4.9A, left and middle panel)**. Similarly, a reduction in cleavage activity was also observed for MmaCas6b of *M. maripaludis* when active site histidine (His38 or 40) was substituted by alanine (Richter, Zoephel et al. 2012). Interestingly, the supplementation of 500 mM imidazole (a histidine mimic) in the cleavage buffer restored the activity of rLinCas6^{H38A} to the basal level **(Figure 4.9A, right panel)**. Such compensation by imidazole for the substituted active site histidine residue is documented for TthCas6 (Niewoehner, Jinek et al. 2014) and PaeCas6 (Lee, Haurwitz et al. 2013). In contrast, such reparation was not observed for the Cas6 of *Synechocystis* species (Jesser, Behler et al. 2019), signifying the difference in the currently existing CRISPR-associated endoribonucleases active residues and possibly many others that remain to be explored.

To determine the fold reduction in cleavage activity of rLinCas6^{H38A}, increasing concentrations (0.05-5 µM) of rLinCas6 and rLinCas6^{H38A} were incubated separately with 5′ fluorescent-labeled sense repeat RNA. A faint band of uncleaved repeat RNA was visible at the highest concentration (5 µM) of rLinCas6^{H38A} **(Figure 4.9B, left panel)** and was of similar intensity to that at 0.5 µM of rLinCas6 **(Figure 4.9B, right panel)**. Thus, a 10-fold reduction in cleavage activity of rLinCas6^{H38A} was detected on cognate repeat RNA substrate. The cleavage activity of rLinCas6^{H38A} (0.05-5 µM) was also explored on 5′ fluorescent-labeled antisense repeat RNA, where a slight reduction in cleavage activity was observed **(Figure 4.9C, left panel)**, as compared to that in an equimolar concentration of rLinCas6 **(Figure 4.9C, right panel)**. At a higher concentration of rLinCas6 (1-2 µM), a prominent degradation of antisense repeat RNA substrate was evident **(Figure 4.9C, right panel)** than rLinCas6^{H38A} **(Figure 4.9C, left panel)**. Similarly, the processing efficiency of pre-crRNA by rLinCas6^{H38A} was also reduced (Figure 4.9D) compared with that by rLinCas6 **(Figure 4.5B)**. The IP fragments of processed pre-

crRNA were still visible when cleaved by 2 µM of rLinCas6^{H38A} **(Figure 4.9D)** compared to the equimolar rLinCas6 **(Figure 4.5B)**. Thus, these RNase assays suggest that the cleavage efficiency of rLinCas6H38A, compared to rLinCas6, was reduced on repeat RNA and pre-crRNA substrates.

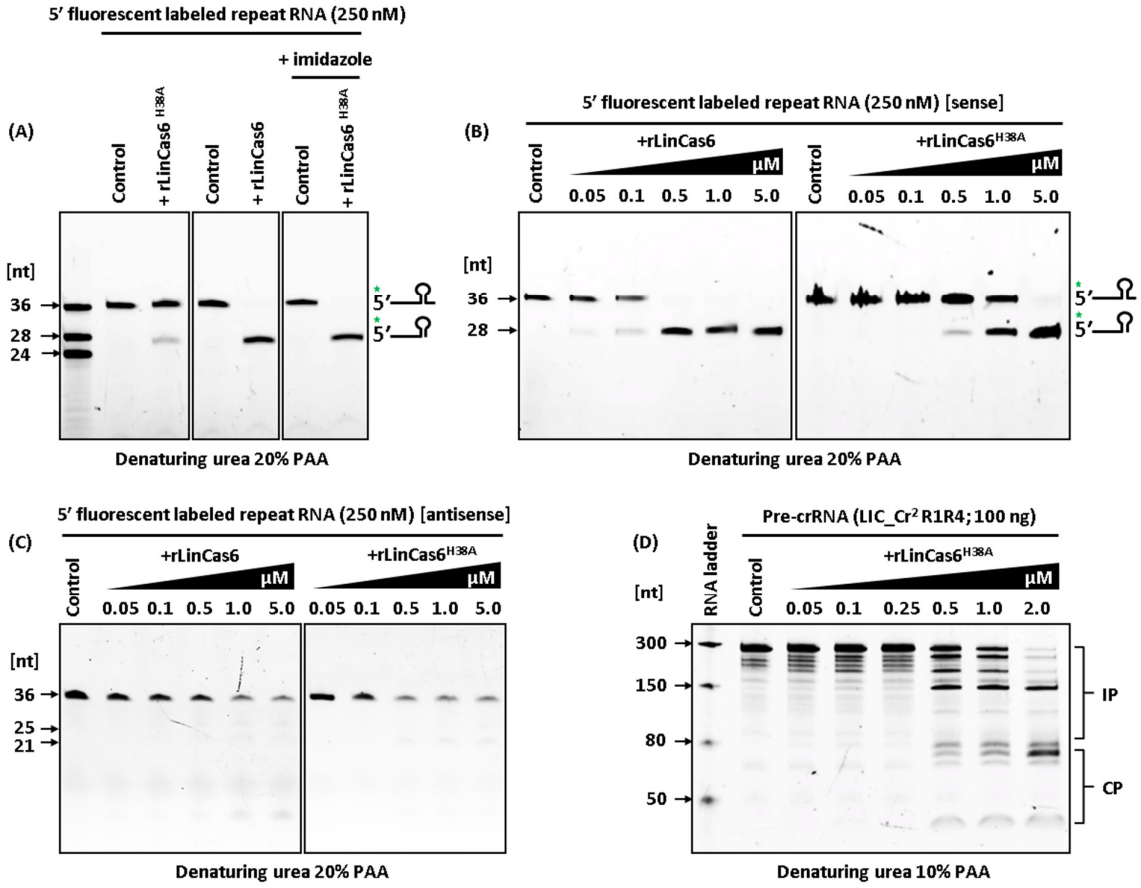


Figure 4.9. Comparison of nuclease activities between rLinCas6 and rLinCas6^{H38A}. (A) The cleavage activity rLinCas6^{H38A} on its cognate repeat RNA (sense) is compromised. The 5' fluorescent-labeled sense repeat RNA (250 nM) was incubated with 1 µM of rLinCas6^{H38A} or rLinCas6 and resolved on denaturing urea gel. Partial cleavage of repeat RNA by the rLinCas6^{H38A} was detected (left panel), whereas the substrate was entirely cleaved by an equimolar concentration of LinCas6 (middle panel). The activity of the mutant variant, rLinCas^{H38A,} was retained at the basal level after the supplementation of imidazole (500 mM) in cleavage buffer (right panel). Increasing concentration (0.05-5 µM) of rLinCas6 and rLinCas6^{H38A} was incubated separately with 5' fluorescent-labeled (B) sense repeat RNA and (C) antisense repeat RNA substrates (250 nM each). Reactions were resolved on denaturing urea gel and visualized directly. (D) Increasing concentration (0.05-2 µM) of rLinCas6^{H38A} was incubated with unlabeled pre-crRNA (LIC_Cr^2 transcript). Reactions were resolved on urea gel and visualized after staining with SYBR-Gold. All reactions, including controls, were incubated for 1 h at 37°C. Gel images are spliced for labeling purposes.

4.2.2 Characterization of rLinCas5

4.2.2.1 The rLinCas5 is catalytically inactive on nucleic acids

Cas5 proteins from type I systems (except the I-C subtype) have been reported to be catalytically inactive (Hochstrasser and Doudna 2015). To investigate the catalytic nature of Cas5 of the I-B system of *Leptospira*, the activity of rLinCas5 was examined on non-specific nucleic acids. The rLinCas5 (2 µM) was incubated (1 h, 37°C) with circular and linear forms of plasmid DNA (100 ng each), and reactions were resolved on an agarose gel. The rLinCas5 was found inactive on circular **(Figure 4.10A, left panel)** as well as linear plasmid DNA **(Figure 4.10A, left panel)**, inferred by no cleavage activity of rLinCas5 on the DNA substrates.

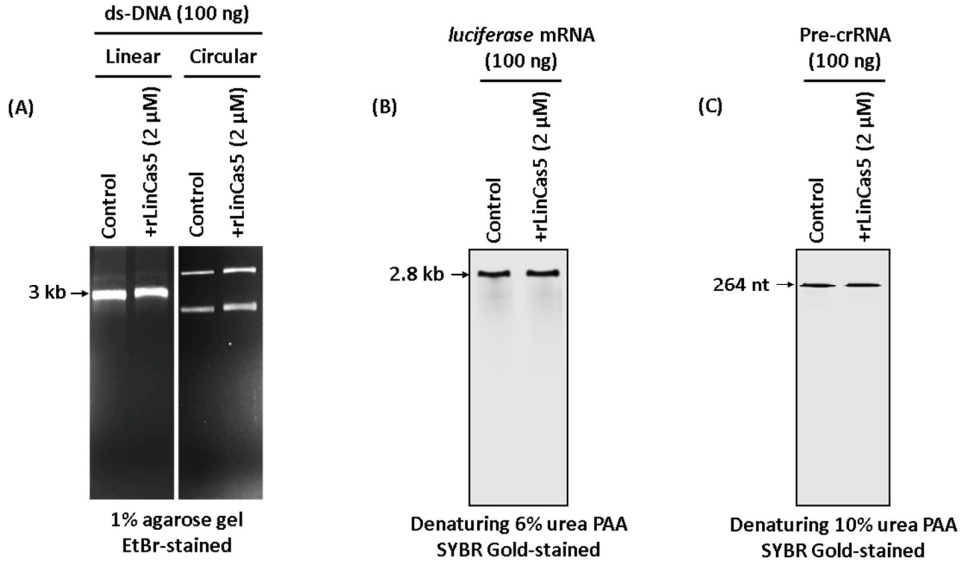


Figure 4.10. The activity of rLinCas5 on nucleic acid substrates. (A) The activity of rLinCas5 on non-specific plasmid DNA. A circular or linear plasmid DNA (pTZ57R/T, 100 ng each) was incubated with or without (control) rLinCas5 (2 µM) and resolved on 1% agarose gel. (B) The activity of rLinCas5 on non-specific RNA. A non-specific luciferase mRNA substrate (100 ng) was incubated with or without rLinCas5 (2 µM) and resolved on denaturing urea 6% PAA. (C) The activity of rLinCas5 on specific RNA. A specific pre-crRNA substrate of LIC_Cr2 (100 ng) was incubated with or without rLinCas5 (2 µM) and resolved on denaturing urea 10% PAA. All reactions, including controls, were incubated for 1 h at 37°C.

The activity of rLinCas5 (2 µM) was also investigated on non-specific *luciferase* mRNA (100 ng) under similar reaction conditions. Urea PAGE analysis of the reaction suggested that rLinCas5 exhibits no cleavage activity on *luciferase* mRNA **(Figure 4.10B)**. Similarly, the

activity of rLinCas5 was examined on a system-specific pre-crRNA (LIC_Cr^{2} transcript) substrate. Urea PAGE of reaction showed no cleavage of pre-crRNA. Altogether, these results implied that rLinCas5, under *in vitro* conditions, is inactive on non-specific and specific nucleic acid substrates.

4.2.2.2 RNA binding analysis of rLinCas5

The family of Cas5 proteins possesses the RNA recognition motif (RRM) required for RNA binding (Zheng, Li et al. 2020). To study the binding ability of rLinCas5, EMSA analysis of the reaction containing pre-crRNA and rLinCas5 was performed. In reactions, the unlabeled pre-crRNA (LA_Cr^{12} R3R4 transcript; 100 ng) was incubated with increasing concentration of rLinCas5 (1-4 µM) and then resolved on a native polyacrylamide gel. With increasing concentration of rLinCas5 in reactions, a concurrent decrease in the intensity of pre-crRNA (117 nt) and a simultaneous increase in smear formation above 117 nt was observed **(Figure 4.11A)**. However, a clear shift in migration of pre-RNA could not be attained on the native gel.

Since Cas5 is known to bind with the 5′ handle of mature crRNA in Cascade (Zheng, Li et al. 2020), the binding of rLinCas5 with the unlabeled mature crRNA was examined. For this purpose, the mature crRNA was generated via the treatment of pre-crRNA (LA_Cr^{12} R3R4 transcript) with rLinCas6 and purified through phenol-chloroform extraction. Afterward, the purified mature crRNA was incubated with an increasing concentration of rLinCas5 (1-4 µM). On native PAGE analysis, one faint IP fragment (78 nt) and two prominent CP fragments containing mature crRNA (71 nt), and a cleavage product of 38 nt were observed in control **(Figure 4.11B)**, as illustrated previously **(Figure 4.5)**. The intensity of each of these fragments was diminished concurrently with the increase in the concentration of rLinCas5 **(Figure 4.11B)**, as observed in the case of pre-crRNA **(Figure 4.11A)**. Similarly, a simultaneous increase in smearing above RNA bands was also observed; however, a clear migration shift of mature crRNA could not be attained. After such observation, we also used a synthetic 5′ fluorescent-labeled crRNA (70 nt), instead of purified mature crRNA, in the EMSA experiment. With increasing concentration of rLinCas5 in reaction, concurrent smearing of 5′ fluorescent-labeled crRNA above 70 nt was observed on the native gel **(Figure 4.11C)**. These results suggested that rLinCas5 may have a weak binding affinity towards pre-crRNA and mature crRNA.

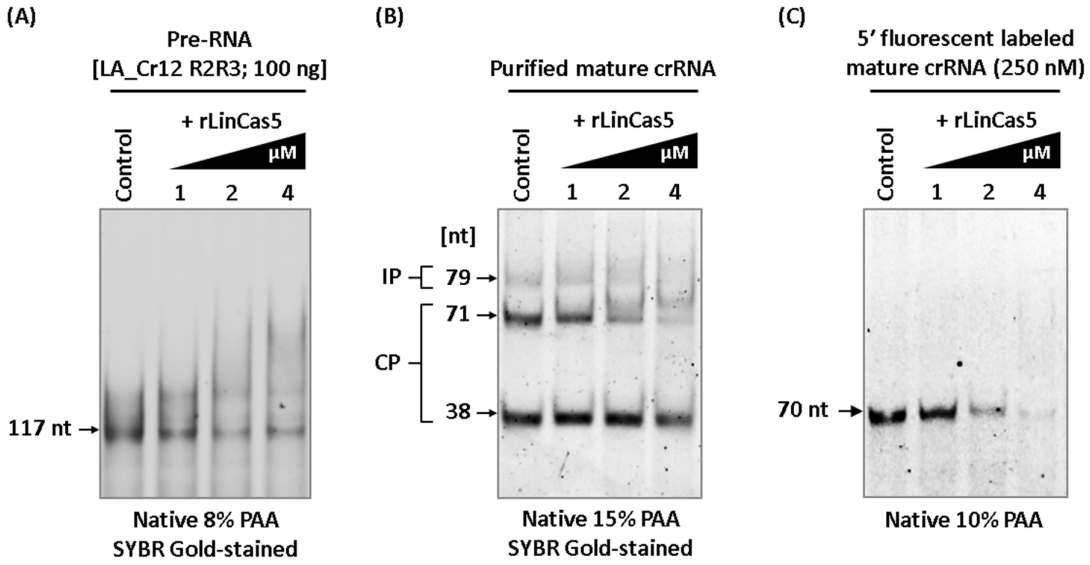


Figure 4.11. EMSA of pre-crRNA and mature crRNA incubated with rLinCas5. (A) Native PAGE of pre-crRNA incubated with rLinCas5. The unlabeled pre-crRNA (LA_Cr^{12} R2R3 transcript; 100 ng) was incubated with increasing concentration of rLinCas5 (1-4 µM) and resolved on 8% native polyacrylamide gel. (B) Native PAGE of purified mature-crRNA incubated with rLinCas5. The mature crRNA, purified via rLinCas6 treatment of pre-crRNA, was incubated with increasing concentration of rLinCas5 (1-4 µM) and resolved on 15% native polyacrylamide gel. (C) Native PAGE of 5' fluorescent-labeled mature crRNA incubated with rLinCas5. All reactions, including controls, were incubated for 1 hour at 37°C.

4.2.2.3 Binding of rLinCas5 to the rLinCas6 bound crRNA

The binding of Cas6 to the crRNA stem-loop after the cleavage of pre-crRNA provides a nucleation point to assemble the remaining Cascade subunits that include Cas5 towards the 5′ end of crRNA (Sashital, Jinek et al. 2011; Sternberg, Haurwitz et al. 2012; Hochstrasser and Doudna 2015). Therefore, rLinCas6 was first allowed to bind to the crRNA, and then the binding of rLinCas5 to the rLinCas6-crRNA complex was examined. To achieve this, 5′ fluorescent-labeled crRNA (250 nM) was incubated with increasing concentration of rLinCas6 (0.1-1 µM) for 30 min at 37°C. Native PAGE analysis of reactions showed rLinCas6-crRNA complex formation in 1:1 and 2:1 stoichiometry at 0.5 and 1 µM of rLinCas6, respectively **(Figure 4.12A)**, as observed previously **(Figure 4.6B)**. Further, to the reaction containing rLinCas6-crRNA (1:1), rLinCas5 (0.5-2 µM) was added and incubated for additional 30 min at 37°C. On native polyacrylamide gel, rLinCas5 concentration-dependent migration shift of rLinCas6-crRNA complex (1:1) was observed **(Figure 4.12B)**. This result indicated the formation of a higher molecular weight complex containing rLinCas5, rLinCas6, and crRNA.

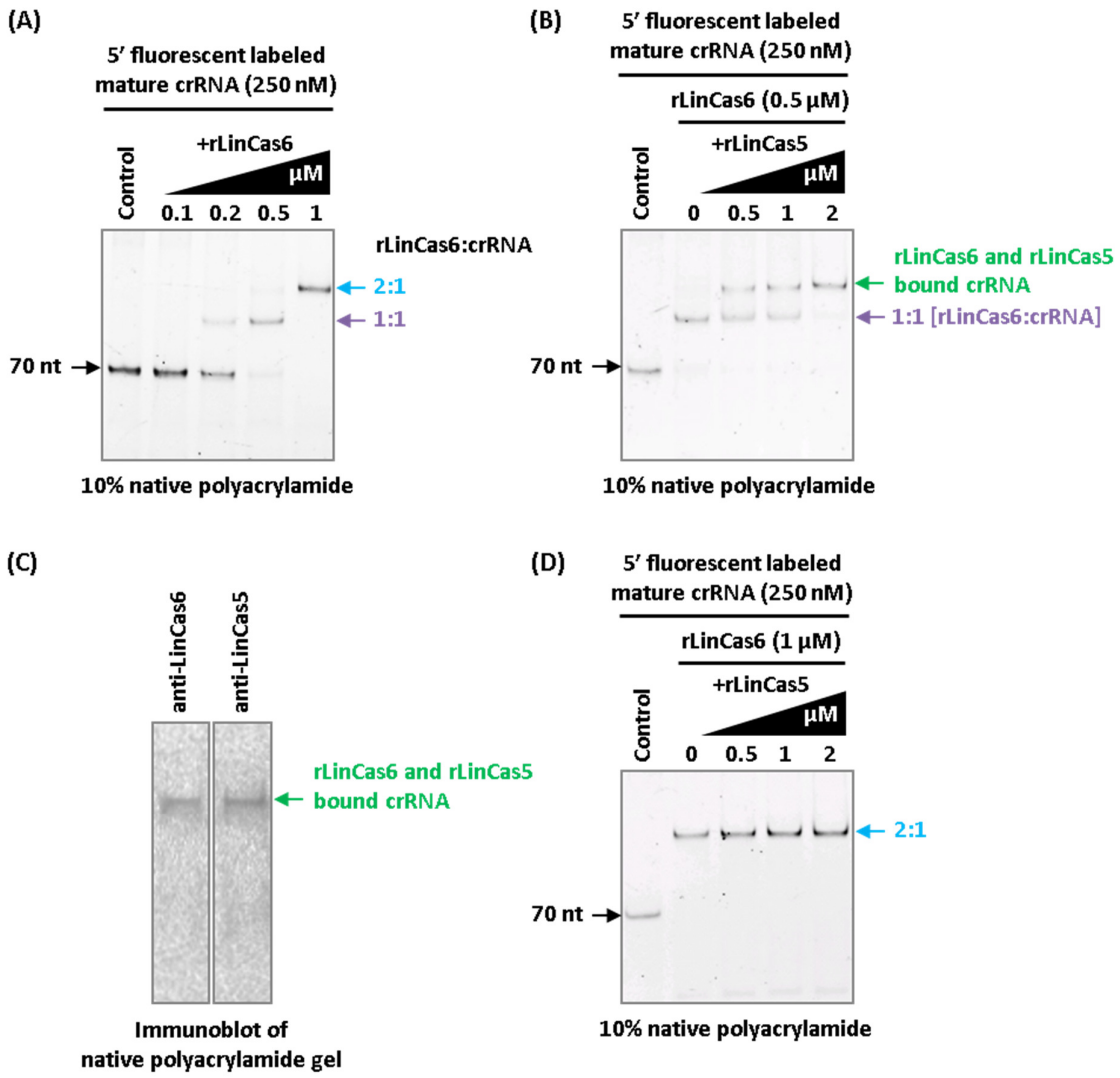


Figure 4.12. EMSA of crRNA bound with rLinCas6 and rLinCas5. (A) Native PAGE of crRNA incubated with rLinCas6. The 5' fluorescent-labeled crRNA (250 nM) was incubated (30 min at 37°C) with increasing concentrations of rLinCas6 (0.1–1 μM) and resolved on native PAA gel. Purple and blue marked bands on the gel correspond to the rLinCas6-crRNA complex in 1:1 and 2:1 stoichiometry, respectively. (B) Native PAGE of rLinCas6-crRNA complex incubated with rLinCas5. The reaction containing rLinCas6-crRNA complex in 1:1 stoichiometry was incubated (30 min at 37°C) with increasing concentration (0.5-2 μM) of rLinCas5 and resolved on a native gel. The green arrow represents crRNA bound with both rLinCas6 and rLinCas5. (C) Immunoblotting of native gel from (B) using anti-LinCas6 and anti-LinCas5 antibodies for detecting rLinCas6 and rLinCas5 in the shifted band (green arrow). (D) Native PAGE of rLinCas6-crRNA complex (2:1 stoichiometry) incubated with rLinCas5 (0.5-2 μM).

To confirm the presence of rLinCas5 and rLinCas6 in the ribonucleoprotein complex, the reaction in which a complete migration shift of crRNA was observed (at 2 μM rLinCas5) **(Figure 4.12B)** was immunoblotted using the anti-LinCas5/LinCas6. In the immunoblot, the

detection of rLinCas5 and rLinCas6 justified the formation of a ribonucleoprotein complex containing rLinCas5, rLinCas6, and crRNA **(Figure 4.12C)**. Interestingly, the migration shift of the rLinCas6-crRNA complex (2:1) was not observed when incubated with an equimolar concentration range (0.5-2 µM) of rLinCas5 **(Figure 4.12D)**. This result suggested that in the rLinCas6-crRNA complex (2:1 stoichiometry), one rLinCas6 protein might have bound towards the 5′ end of crRNA; thus, circumventing the interaction of rLinCas5 with the crRNA.

4.2.3 Characterization of rLinCas3

4.2.3.1 LinCas3 is a fusion of nuclease and helicase

The two enzymatic domains of Cas3 are a histidine-aspartate (HD) nuclease domain and a Superfamily 2 (SF2) helicase domain (Jackson, Lavin et al. 2014). These two domains may be fused as a single polypeptide or expressed separately as Cas3′ (SF2 helicase) and Cas3″ (HD nuclease). In the I-B system of sv. Copenhageni, a single copy of gene *cas3* (*LIC10938*) encoding LinCas3, was identified. Thus, LinCas3 seems to be a fusion of HD nuclease and SF2 helicase domains. As anticipated, analysis of the LinCas3 sequence (750 amino acids) through the ScanProsite tool revealed fusion of LinCas3 HD nuclease [HD Cas3-type (19-198); 180 residues)] and LinCas3 SF2 helicase domains **(Figure 4.13)**. Further, the two domains within LinCas3 SF2 helicase, “helicase ATP-binding” domain (278-463; 186 residues) and “helicase C-terminal” domain (487-658; 172 residues) were identified in the central and C-terminal region of LinCas3, respectively **(Figure 4.13)**.

For Cas3 HD nuclease activity, variation in substrate specificities and cofactor (metal ion) requirements has been reported (Han and Krauss 2009; Beloglazova, Petit et al. 2011; Mulepati and Bailey 2011; Sinkunas, Gasiunas et al. 2011; Sinkunas, Gasiunas et al. 2013; Gong, Shin et al. 2014; Huo, Nam et al. 2014). To determine the metal ion supporting the LinCas3 HD nuclease activity, nuclease activity assays of rLinCas3 on DNA and RNA substrates were performed.

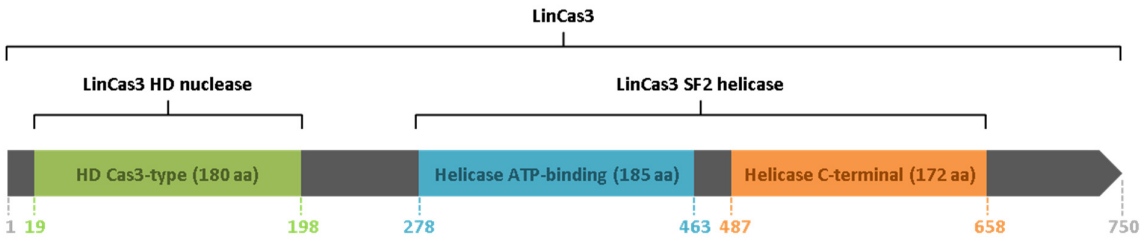


Figure 4.13. Schematic representation of domains identified in LinCas3. Analysis of LinCas3 through ScanProsite tool identified LinCas3 HD (histidine-aspartate) nuclease (green) and LinCas3 SF2 helicase containing helicase ATP-binding domain (blue) and helicase C-terminal domain (orange). The position and length of each domain in LinCas3 are shown by labeled vertical dashed lines and under brackets in the schematic. Small case "aa" stands for amino acid residues.

4.2.3.2 LinCas3 is a metal-dependent ss-DNase

Biochemical analysis of most of the Cas3 or Cas3′′ proteins showed ss-DNase activity supported by various metal ions. Thus, the nuclease activity of rLinCas3 (apo form; 250 nM) was first evaluated on circular ss-DNA (ΦX174 virion; 100 ng) in reaction containing no or specific metal ions [Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, and Zn^{2+} (10 mM each)]. Reactions were resolved on agarose gel after 1 h of incubation at 37°C. The optimum nuclease activity of rLinCas3 on the ss-DNA was observed in the presence of the Ni^{2+} ion, inferred by the detection of degraded product at around 100 bp on the agarose gel **(Figure 4.14A)**. Other divalent metal ions such as Mn^{2+}, Co^{2+}, Mg^{2+}, and Ca^{2+} also supported the ss-DNase activity of rLinCas3. However, minimal or no rLinCas3 nuclease activity on ss-DNA was evident in the presence of Cu^{2+} and Zn^{2+} ions. Diffused band pattern in "Cu^{2+} and Zn^{2+}" lanes of agarose gel could be due to *in vitro* artifacts. Nevertheless, the ability of rLinCas3 to degrade DNA substrate only with selective metal ions **(Figure 4.14A)** suggested that LinCas3 is a metal ion-dependent ss-DNase. This result is consistence with previous studies where Mg^{2+}, Mn^{2+}, Ni^{2+}, and Co^{2+} were found to be as most common activators of Cas3 nuclease activities (Han and Krauss 2009; Beloglazova, Petit et al. 2011; Mulepati and Bailey 2011; Sinkunas, Gasiunas et al. 2011; Sinkunas, Gasiunas et al. 2013; Gong, Shin et al. 2014; Huo, Nam et al. 2014). To determine the concentration of metal ion in reaction for optimal nuclease activity of rLinCas3, 1 µM of rLinCas3 was incubated (1 h at 37°C) with circular ss-DNA (100 ng) in buffer containing Ni^{2+} or Mg^{2+} ions with concentration ranging 1 µM-100 mM. As evident from the agarose gel electrophoresis of reactions, the degradation of ss-DNA increased with a concurrent increase in the Ni^{2+} **(Figure 4.14B, left panel)** or Mg^{2+} ion concentration (1 µM-10 mM) **(Figure 4.14B, right panel)**. Notably, the nuclease activity of rLinCas3 was optimum at 1-10 mM of metal ions. Interestingly, at a specific metal ions concentration (100 mM), inhibition of rLinCas3 nuclease activity was observed **(Figure 4.14B)**.

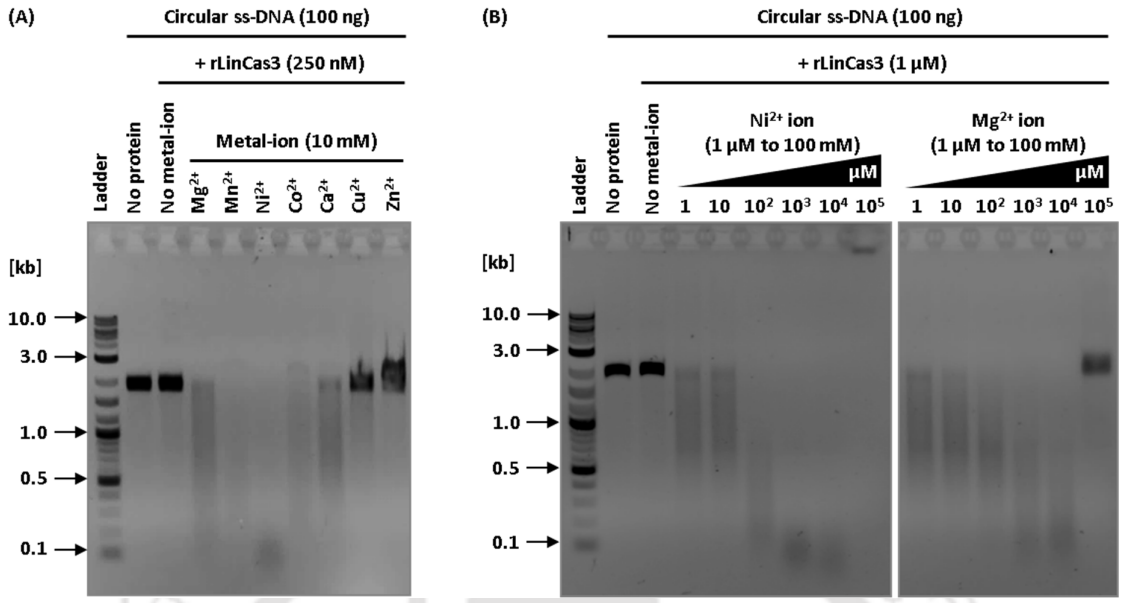


Figure 4.14. Nuclease activity of rLinCas3 on circular ss-DNA (ΦX174 virion). (A) The activity of rLinCas3 on circular ss-DNA in the presence of different metal ions. The rLinCas3 (250 nM) was incubated with circular ss-DNA (100 ng) in the absence of metal ions or the presence of Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} (10 mM each) (B) The activity of rLinCas3 on circular ss-DNA in buffer containing varying concentration of the metal ion. The rLinCas3 (1 µM) was incubated with circular ss-DNA (100 ng) in buffer containing Ni^{2+} (left panel) or Mg^{2+} (right panel) at increasing concentration (1 µM-100 mM). Reactions, including controls, were incubated for 1 h at 37°C and resolved onto 1% agarose gels.

4.2.3.3 ATP in reaction intervenes with the ss-DNase activity of rLinCas3

Nuclease activity in the HD domain of Cas3 from *Streptococcus thermophiles* (SthCas3) degrades circular ss-DNA in a metal-dependent manner. The rate of SthCas3 nuclease activity on circular ss-DNA was similar in the absence or presence of ATP (Sinkunas, Gasiunas et al. 2011). To test whether the ATP affects the ss-DNase activity of rLinCas3, the metal-dependent nuclease activity of rLinCas3 was examined on circular ss-DNA in the absence or presence of ATP. The circular ss-DNA (ΦX174 virion; 100 ng) was incubated with an increasing concentration (50-500 nM) of rLinCas3 in a buffer lacking or containing ATP (2 mM). Agarose gel electrophoresis of reactions suggested that 50 nM (or less) of rLinCas3 is sufficient to degrade 100 ng of the circular ss-DNA substrate **(Figure 4.15, left panel)**. The reaction product (around 100 bp) could not be degraded further by increasing the concentration of rLinCas3 up to 500 nM **(Figure 4.15, left panel)**. Interestingly, with the addition of ATP in the reaction, the rLinCas3 efficacy to degrade circular ss-DNA was relatively reduced **(Figure 4.15, middle panel)**. Such reduction in nuclease activity of rLinCas3 on circular ss-DNA was also evident

when ATP in the reaction was replaced by a non-hydrolyzable derivative of ATP (ATP-γ-S; 2 mM) **(Figure 4.15, right panel)**. These results suggested that the presence of ATP in the reaction intervenes with the ss-DNase activity of rLinCas3, and such intervention is probably not due to the hydrolysis of ATP by rLinCas3. Moreover, in the presence and absence of ATP, the equimolar concentration of untagged rLinCas3 (50-500 nM) showed degradation of circular ss-DNA (data not shown), as observed in Figure 4.15 where SUMO-tagged rLinCas3 was used. It suggests that the presence of the SUMO tag in the rLinCas3 did not change its biochemical properties. Hence, for further downstream analysis, we used the SUMO-tagged version of rLinCas3.

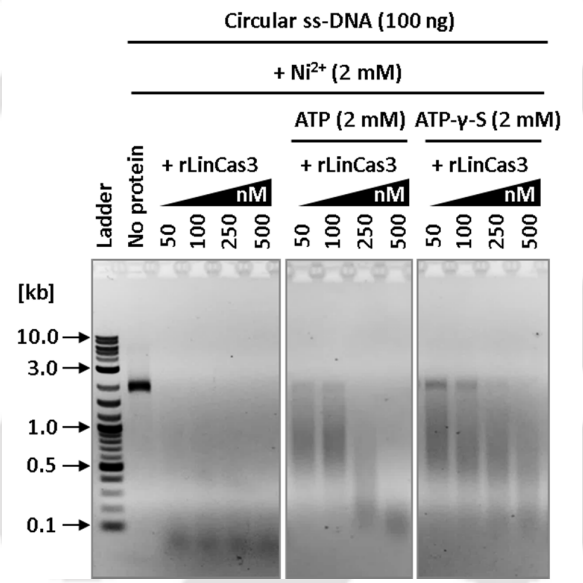


Figure 4.15. Effect of ATP on nuclease activity of rLinCas3 on circular ss-DNA. The activity of rLinCas3 (50-500 nM) on circular ss-DNA (ΦX174 virion; 100 ng) in the absence of nucleotide (left panel) or in the presence of ATP (middle panel) or ATP-γ-S (right panel). All reactions were performed in the buffer containing 2 mM of Ni^{2+} ion. Reactions, including controls, were incubated for 1 h at 37°C and resolved onto 1% agarose gels.

4.2.3.4 LinCas3 is a metal-dependent ds-DNase

After exploring the activity of rLinCas3 on circular ss-DNA, the nuclease activity of rLinCas3 on ds-DNA was evaluated. The rLinCas3 (250 nM) was incubated (1h, 37°C) with circular or linear form of ds-DNA (pTZ57R/T; 100 ng) in the absence or presence of metal ion [Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} (10 mM each)]. Agarose gel electrophoresis of reactions after incubation suggested degradation of circular ds-DNA exclusively in the presence of Mg^{2+} ion **(Figure 4.16A)**. In control (without rLinCas3), bands of supercoiled and open circular (nicked) forms of plasmid DNA were evident. In the presence of rLinCas3 and Mn^{2+}, Ni^{2+},

Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} ion, an increase in the band intensity corresponding to open circular plasmid DNA was observed. This result suggested rLinCas3-mediated nicking of the supercoiled form of plasmid DNA resulting in an open circular form. In addition, linearization of plasmid DNA by rLinCas3 was also observed in the presence of Mg^{2+}, Mn^{2+}, Ni^{2+}, or Co^{2} **(Figure 4.16A)**. Diffused band pattern in "Mn^{2+}" lane of agarose gel could be due to *in vitro* artifacts. Nevertheless, this result suggested that LinCas3 is Mg^{2+} ion-dependent ds-DNase. In consistence with this result, degradation of a linear ds-DNA by rLinCas3 was evident exclusively in the presence of Mg^{2+} ion **(Figure 4.16B)**.

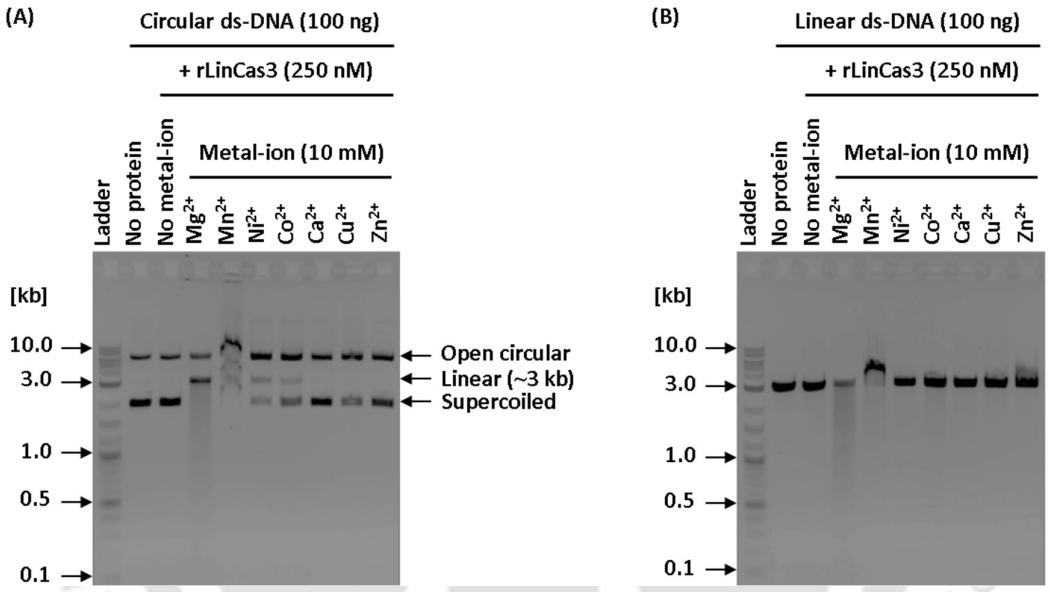


Figure 4.16. Nuclease activity of rLinCas3 on ds-DNA (pTZ57R/T). The rLinCas3 (250 nM) was incubated with (A) circular ds-DNA or (B) linear ds-DNA (100 ng each) in the absence of metal ions or the presence of Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} ion (10 mM each). Reactions, including controls, were incubated for 1 h at 37°C and resolved onto 1% agarose gels. Three forms of plasmid DNA (open circular, linear, and supercoiled) are demarcated right to the gel image (A).

In line with this finding, the Cas3 HD domain protein from *S. solfataricus* has been shown to degrade ds-DNA. In contrast, ds-DNase activity on plasmid DNA (without Cascade or R-loop) by standalone Cas3 or Cas″ proteins from *S. thermophilus*, *T. thermophilus*, *E. coli*, and *M. jannaschii* was not observed (Beloglazova, Petit et al. 2011; Mulepati and Bailey 2011; Sinkunas, Gasiunas et al. 2011; Mulepati and Bailey 2013).

Since ATP intervened with the ss-DNase activity of rLinCas3 **(Figure 4.15)**, the effect of ATP was also studied on its ds-DNase activity. The circular or linear ds-DNA (100 ng each) was

incubated with an increasing concentration (50-1000 nM) of rLinCas3 in the presence of Mg^{2+} (2 mM) in a buffer lacking or containing ATP (2 mM). Agarose gel electrophoresis of reactions suggested rLinCas3 concentration-dependent degradation of circular **(Figure 4.17A)** as well as linear **(Figure 4.17B)** ds-DNA in both the presence and absence of ATP. With the addition of ATP in the reaction, the rLinCas3 efficacy to degrade circular ds-DNA was relatively reduced **(Figure 4.17A, right panel)**. This observation was consistent with the reduced rLinCas3 efficacy in degrading circular ss-DNA in the presence of ATP. Interestingly, rLinCas3 efficacy in degrading linear ds-DNA was similar in the absence **(Figure 4.17B, left panel)** and presence **(Figure 4.17B, right panel)** of ATP.

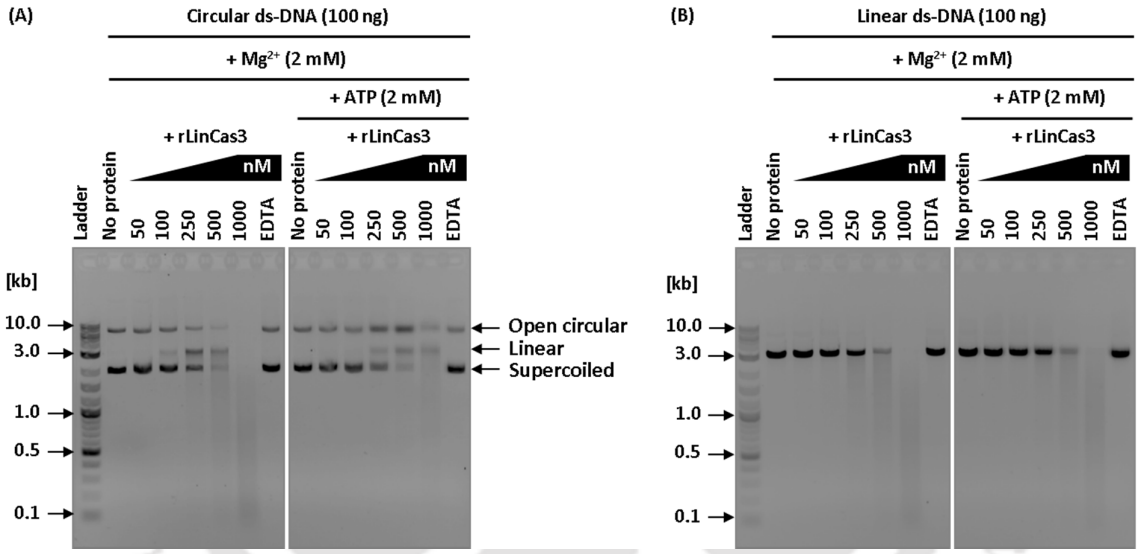


Figure 4.17. Effect of ATP on nuclease activity of rLinCas3 on ds-DNA (pTZ57R/T). (A) Activity of rLinCas3 (50-1000 nM) on circular ds-DNA (100 ng) in absence of nucleotide (left panel) or in presence of ATP (right panel). Three forms of plasmid DNA (open circular, linear, and supercoiled) are demarcated right to the gel image. (B) Activity of rLinCas3 (50-1000 nM) on linear ds-DNA (100 ng) in absence of nucleotide (left panel) or in presence of ATP (right panel). In an additional reaction containing rLinCas3 (1000 nM), EDTA (10 mM) was used to chelate metal-ion. All reactions were performed in the buffer containing 2 mM of Mg^{2+} ion. Reactions, including controls, were incubated for 1 h at 37°C and resolved onto 1% agarose gels.

4.2.3.5 LinCas3 is a metal-dependent ss-RNase

In addition to the metal-dependent ss-DNase activity, metal-dependent ss-RNase activity has also been commonly reported in Cas3 or Cas3″. To investigate the RNase activity of rLinCas3, an unlabeled luciferase mRNA substrate (100 ng) was incubated with rLinCas3 (250 nM) in the absence or in the presence of metal ion [Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} (10

mM each)]. Agarose gel electrophoresis of reactions showed no RNase activity in rLinCas3 without any metal ion **(Figure 4.18A)**. In contrast, in the presence of Mg^{2+} ion, degradation of the RNA substrate by rLinCas3 was observed on the agarose gel. Migration shift of smeared bands observed on the agarose gel suggested that rLinCas3 binds to the degraded or undegraded RNA in the presence of other metal ions (Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, and Zn^{2+}) **(Figure 4.18A)**. Nevertheless, this result suggested that LinCas3 is metal ion-dependent RNase.

To examine the effect of ATP on the RNase activity of rLinCas3, *luciferase* mRNA (100 ng) was incubated (1 h, 37°C) with increasing concentration of rLinCas3 in the absence or presence of ATP in a buffer containing Mg^{2+} ion (2 mM). Agarose gel electrophoresis of reactions indicated rLinCas3 concentration-dependent degradation of *luciferase* mRNA **(Figure 4.18B)**. A similar amount of degradation was evident at an equimolar concentration of rLinCas3 in both absence **(Figure 4.18B, left panel)** and presence **(Figure 4.18B, right panel)** of ATP. Thus, ATP could not intervene with the RNase activity of rLinCas3.

The nuclease activity of rLinCas3 on various nucleic acid substrates in the presence of ATP suggested that nucleotide only intervened with rLinCas3 mediated degradation of circular nucleic acid substrates (circular ss- and ds-DNA). Thus, it can be concluded that in the presence of nucleotide, the nicking activity of rLinCas3 is reduced and results in the slow degradation of circular nucleic acid substrates. It hinted that nucleotide might regulate the nuclease activity in LinCas3 during target degradation in *Leptospira*.

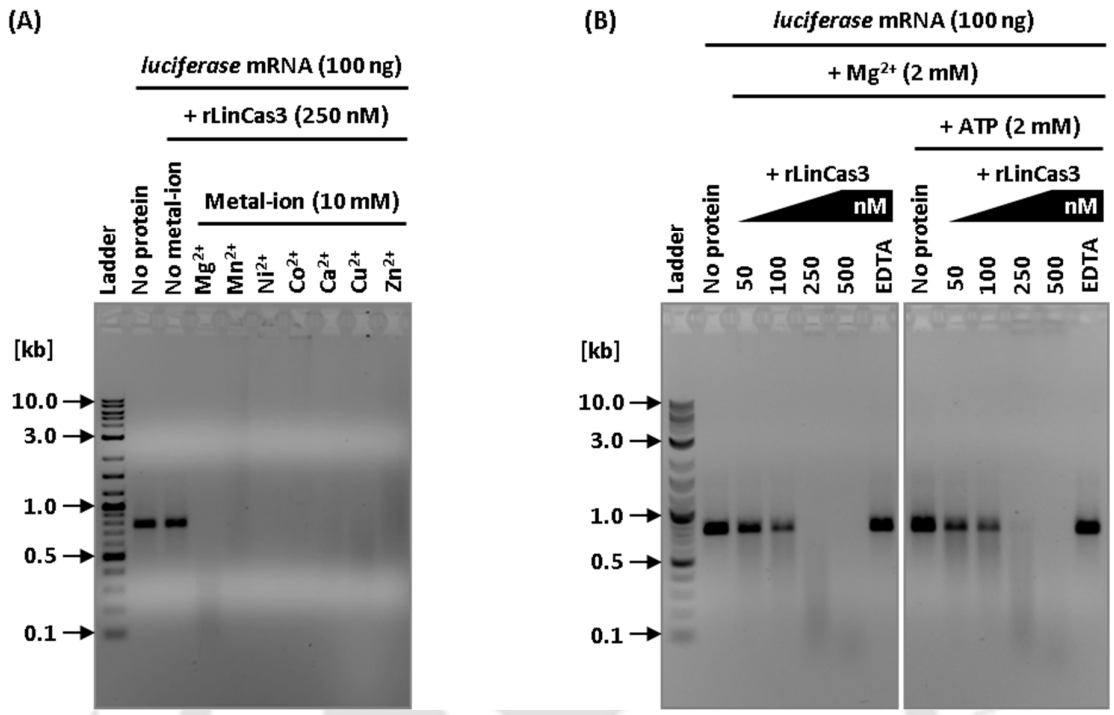


Figure 4.18. Nuclease activity of rLinCas3 on linear ss-RNA (*luciferase* mRNA). (A) The activity of rLinCas3 on RNA in the presence of different metal ions. The rLinCas3 (250 nM) was incubated with RNA (100 ng) in the absence of metal ion or in the presence of Mg^{2+}, Mn^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+}, Cu^{2+}, or Zn^{2+} (10 mM each). (B) The activity of rLinCas3 (50-500 nM) on RNA (100 ng) in the absence of nucleotide (left panel) or in the presence of ATP (right panel). In an additional reaction containing rLinCas3 (500 nM), EDTA (10 mM) was used to chelate metal-ion. Reactions, including controls, were incubated for 1 h at 37°C and resolved onto 1% agarose gels.

4.2.3.6 LinCas3 is inactive on DNA oligonucleotide but cleaves RNA oligonucleotide

After investigating the nuclease activity of rLinCas3 on longer-size nucleic acid substrates (more than 1.5 kb), the nuclease activity of rLinCas3 was examined on the nucleic acid substrate of short-length (DNA and RNA oligonucleotides; 36 nucleotides). According to previously observed results, nuclease activities of rLinCas3 on ss-DNA and ss-RNA oligos were performed in a reaction buffer containing Ni^{2+} and Mg^{2+} metal ions, respectively. The rLinCas3 (50 or 100-1000 nM) was incubated with 5′ fluorescent-labeled repeat DNA or RNA (antisense; 250 nM) for 1 h at 37°C. Denaturing urea PAGE of reaction products showed no activity of rLinCas3 on ss-DNA oligo **(Figure 4.19A)**. However, rLinCas3 cleaves the ss-RNA oligo in a concentration and metal-dependent manner **(Fig. 4.19B)**.

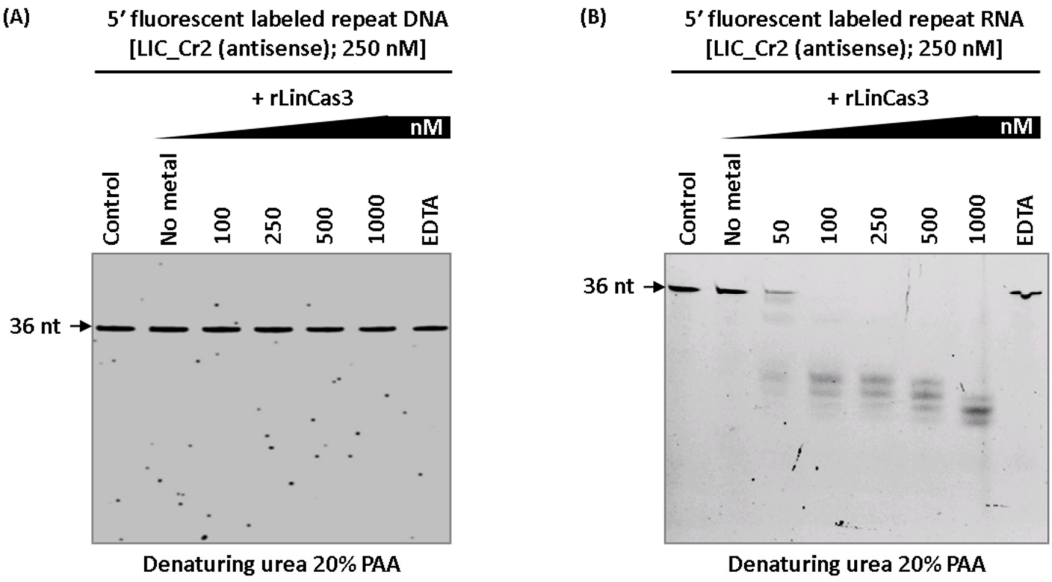


Figure 4.19. Nuclease activity of rLinCas3 on 5' fluorescent-labeled ss-DNA and ss-RNA oligos. (A) The activity of rLinCas3 (100-1000 nM) on ss-DNA oligo in the presence of Ni^{2+} ion. (B) The activity of rLinCas3 (50-1000 nM) on ss-RNA oligo in the presence of Mg^{2+} ion. In an additional reaction containing rLinCas3 (1000 nM), EDTA (10 mM) was used to chelate metal ion. Reactions, including controls, were incubated for 1 h at 37°C, resolved on denaturing 20% polyacrylamide gels, and visualized without staining.

4.2.3.7 Multiple sequence alignment of LinCas3 and its orthologs

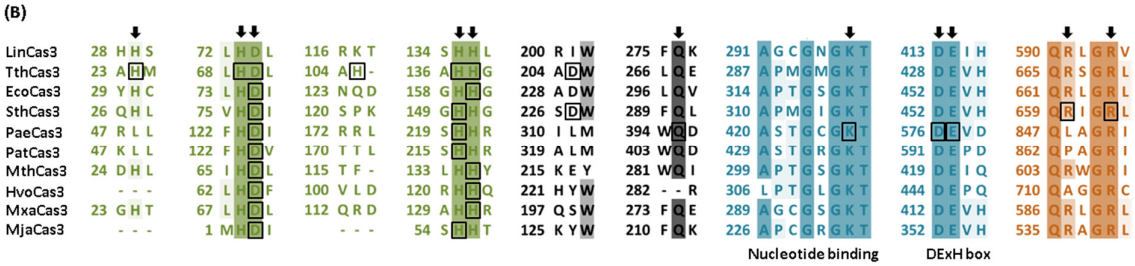
The nuclease function of Cas3 or Cas3″ is activated via metal coordinating residues. To predict the metal ion binding residues in LinCas3, MSA of LinCas3 and several known Cas3 orthologs were performed. According to reviewed Cas3 entries (Q53VY2, P38036, F2XG53, Q02ML8, Q6D0W9, O27158, D4GQN8, Q1CW46, and Q57821) in the UniProtKB database, known Cas3 orthologs [*Thermus thermophilus* (TthCas3), *Escherichia coli* (EcoCas3), *Streptococcus thermophiles* (SthCas3), *Pseudomonas aeruginosa* (PaeCas3), *Pectobacterium atrosepticum* (PatCas3), *Methanothermobacter thermautotrophicus* (MthCas3), *Haloferax volcanii* (HvoCas3), *Myxococcus xanthus* (MxaCas3), and *Methanocaldococcus jannaschii* (MjaCas3), respectively] were identified and used in the alignment with LinCas3. In the aforementioned Cas3 proteins, domains arrangement, metal-ion binding residues, nucleotide-binding region, ATPase motif (DExH/D box), and critical residues for helicase function were assigned by the UniProtKB database based on automatic annotation, sequence or structural similarities or a previous report (Mulepati and Bailey 2011). MSA generated an identity matrix that revealed sequence diversity between LinCas3 and Cas3 orthologs, with an identity of ~16-33%. In

MSA, known metal ion binding residues in Cas3 orthologs aligned with H29, 73, 135, 136, and D74 in LinCas3. Thus, these five residues are potential metal co-ordinating residues identified in the HD nuclease domain of LinCas3.

(A)

	LinCas3	TthCas3	EcoCas3	SthCas3	PaeCas3	PatCas3	MthCas3	HvoCas3	MxaCas3	MjaCas3
LinCas3 (750 residues)	100.00	22.09	19.75	19.85	16.90	16.26	21.21	17.55	32.97	21.91
TthCas3 (920 residues)	22.09	100.00	28.50	30.85	20.29	20.39	22.74	21.16	27.34	20.48
EcoCas3 (888 residues)	19.75	28.50	100.00	24.18	18.53	16.36	18.98	17.93	22.67	24.61
SthCas3 (926 residues)	19.85	30.85	24.18	100.00	17.73	18.30	20.77	18.62	21.51	22.36
PaeCas3 (1076 residues)	16.90	20.29	18.53	17.73	100.00	48.64	18.33	19.00	18.17	17.91
PatCas3 (1098 residues)	16.26	20.39	16.36	18.30	48.64	100.00	18.84	18.24	17.82	17.82
MthCas3 (830 residues)	21.21	22.74	18.98	20.77	18.33	18.84	100.00	18.99	19.38	21.14
HvoCas3 (937 residues)	17.55	21.16	17.93	18.62	19.00	18.24	18.99	100.00	18.64	18.09
MxaCas3 (746 residues)	32.97	27.34	22.67	21.51	18.17	17.82	19.38	18.64	100.00	20.83
MjaCas3 (728 residues)	21.91	20.48	24.61	22.36	17.91	17.82	21.14	18.09	20.83	100.00

(B)



(C)

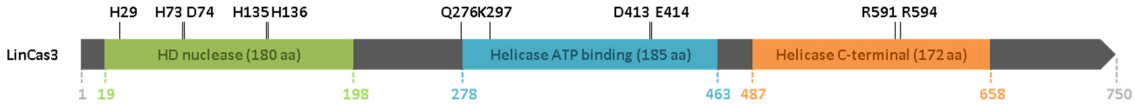


Figure 4.20. MSA of LinCas3 and its orthologs. Protein sequences of the Cas3 orthologs used in this alignment are available in the UniProtKB database with the following entry names: Q72TS9 (LinCas3), Q53VY2 (TthCas3), P38036 (EcoCas3), F2XG53 (SthCas3), Q02ML8 (PaeCas3), Q6D0W9 (PatCas3), O27158 (MthCas3), D4GQN8 (HvoCas3), Q1CW46 (MxaCas3), and Q57821 (MjaCas3). (A) The identity matrix of LinCas3 and its orthologs. The Cas3 orthologs and the LinCas3 were aligned using Clustal Omega, and the identity matrix was generated. Values given in the matrix cells are percent identity between respective proteins. (B) MSA of LinCas3 and its orthologs. Parts of aligned regions around known metal ion binding or catalytic residues for ATPase/helicase (black square-shaped boxes) in Cas3 orthologs are shown. Different degrees of conservation are highlighted by multiple shading of green (HD domain), grey (the region between HD and ATP binding domain), blue (ATP binding domain), and orange (helicase C-terminal domain) colors. The potential metal binding (H29, 73, 135, 136, and D74) and catalytic residues [ATPase (Q276, K297, D413, and E414) and helicase (R591 and 594)] of LinCas3 are indicated by vertical downward arrows. (C). Schematic of potentially critical residues of LinCas3. Potential metal binding and catalytic (ATPase/helicase) residues of LinCas3 are indicated over the architecture of LinCas3 drawn to the scale. Green, blue and orange colors represent HD nuclease, helicase ATP-binding, and helicase C-terminal domains of LinCas3.

Additional fused SF2 helicase domain of Cas3 contains catalytic residues for ATPase and helicase functions. In addition to metal binding residues in the HD domain of LinCas3, potentially critical residues for ATPase (Q276, K297, D413, and E414) and helicase (R591 and 594) activities were revealed in LinCas3 thorough alignment with known Cas3 orthologs. Characterization of ATPase/helicase activity of LinCas3, including validation of predicted metal binding and catalytic residues, is required to understand the biochemical properties of standalone LinCas3.

The mature crRNA generated after processing is supplied to an effector complex containing multiple protein subunits, forming a Cascade that delivers crRNA to the target (He, St John James et al. 2020). Recruited by Cascade at the target, the signature Cas3 protein (type I) forms the interference complex that degrades the target DNA (He, St John James et al. 2020). In general, the signature Cas3 proteins exhibit a C-terminal superfamily 2 (SF2) helicase domain which contains a DEAD/DEAH box region (Jackson, Lavin et al. 2014; He, St John James et al. 2020). Many proteins from the DEAD-box family (SF2) of RNA helicases have been shown *in vitro* to unfold RNA hairpins and resolve RNA secondary structures. (Rogers, Richter et al. 1999; Marsden, Nardelli et al. 2006; König, Liyanage et al. 2013). Since Cas3 is found naturally fused with Cas6 in *L. interrogans* sv. Linhai, Cas3 may thus have a role in the expression stage, at least in *Leptospira*, if not in all cases. The presence of single lengthy pre-crRNA arrays in *L. interrogans* sv. Lai could have an evolutionary role in the natural fusion of LinCas6 and LinCas3. As previously observed (König, Liyanage et al. 2013), the lengthy pre-crRNA is expected to form a compact and stable structure. Thus cleavage sites within compact pre-crRNA may not be uniformly accessible to LinCas6. It is possible that the processing of such pre-crRNA requires another helper protein(s) to sustain the cleavage sites within pre-crRNA accessible to LinCas6. We speculate that LinCas3 helicase may unwind the compact pre-crRNA, allowing LinCas6 to access the cleavage sites within repeat RNA regions. However, further research is needed to confirm this notion.

4.3 Conclusion

In this study, the functional orientation of CRISPR I-B arrays was endorsed using LinCas6. *In vitro* RNase assay using the recombinant *Leptospira* Cas6 (rLinCas6; N-terminal 6×his tagged) demonstrated the conventional cleavage (upstream to 8 nt from the 3′ end) of cognate repeat RNA substrate in the proposed sense orientation. It thus agreed with the RT-PCR analysis,

which demonstrated a co-directional arrangement of CRISPR-Cas I-B in *Leptospira*. In the rLinCas6 RNase assay, exponential decay of repeat RNA substrate suggested that LinCas6 follows single turnover kinetics. This cleavage mode means that rLinCas6 remains bound with the cleaved repeat RNA product. It was advocated by a polyacrylamide gel-based EMSA experiment where rLinCas6 concentration-dependent migration shift of cleaved repeat RNA product was observed. Similar to the cleavage of repeat RNA alone, rLinCas6 cleaved repeat RNA segments within pre-crRNAs to generate mature crRNAs. It demonstrates LinCas6-mediated crRNA biogenesis associated with the CRISPR-Cas I-B system in *Leptospira*. In agreement with the binding of rLinCas6 to the cleaved repeat RNA, the binding of rLinCas6 to mature crRNAs was also evident in the polyacrylamide gel-based EMSA. Migration shifts (EMSA) and elution profile (SEC) of rLinCas6-crRNA complex suggested that under *in vitro* conditions, rLinCas6 can bind to the mature crRNA in two stoichiometries (1:1 or 2:1). For recognition and binding of crRNA, a conserved G-loop of Cas6 is known to be crucial. This G-loop was also evident in LinCas6 in MSA with known Cas6 orthologs. This MSA also suggested a potential catalytic triad (Y27, H38, and Y106) in LinCas6 that might be employed during its acid-base catalysis on RNA substrates. Moreover, the substitution of one of the predicted active site residues (H38) in LinCas6 (LinCas6^{H38A}) resulted in reduced cleavage activity on cognate repeat RNA and pre-crRNA substrates.

Besides LinCas6, LinCas5 was also characterized in this study as a first step towards understanding the Cascade I-B assembly in *Leptospira*. The rLinCas5 (N-terminal 6×his-SUMO-tagged) exhibited no nuclease activities on DNA and RNA, suggesting it to be catalytically inactive on nucleic acids. Polyacrylamide gel-based EMSAs suggest that the rLinCas5 may have a weak binding affinity towards pre-crRNA and mature crRNA. However, a clear migration shift of the crRNA-rLinCas6 complex (1:1 stoichiometry) in the presence of rLinCas5 suggested the formation of a higher molecular weight complex containing rLinCas6, rLinCas5 and crRNA.

Furthermore, the nuclease activity of the signature Cas protein (LinCas3) was characterized in this study to understand the physiological requirements of *Leptospira* CRISPR-Cas I-B in the degradation of DNA targets during interference. In nuclease assays, rLinCas3 (N-terminal 6×his-SUMO-tagged) can effectively degrade both ss- and ds-DNA in the presence of Ni^{2+} and Mg^{2+} ions, respectively. The presence of ATP (or non-hydrolyzable ATP derivative) in the reaction intervened with the nuclease activity of rLinCas3 (tagged or untagged version) on

circular DNA substrates. It suggested that such intervention by nucleotide in DNase activity on circular DNA substrates is the inherent characteristic of LinCas3 and is not due to the hydrolysis of ATP by LinCas3 ATPase activity. It hinted that nucleotide might regulate the nuclease activity in LinCas3 during target degradation in *Leptospira*. The rLinCas3 also exhibited divalent metal ion-dependent RNase activity on mRNA and RNA oligonucleotide. However, no nuclease activity of rLinCas3 was observed on DNA oligonucleotides. The present study on rLinCas3 indicated that LinCas3 is an active protein with a biologically relevant nuclease function.

The present study highlights the comprehensive understanding of the *Leptospira* CRISPR array transcription and its processing by LinCas6, which is central to RNA-mediated CRISPR-Cas I-B adaptive immunity. This study also underlines the formation of the nucleoprotein complex (crRNA, rLinCas6, and rLinCas5), which is essential for constructing a functional Cascade molecule. In addition, the biochemical analysis of LinCas3 nuclease presented in this study signifies the similarity and differences in the currently known Cas3 nuclease.

CHAPTER 5

Conclusion and Future Prospects

5.1 Conclusion

The presence of CRISPR-Cas system/s mainly in infectious *Leptospira* species is thought to be one of the virulence factors and leptospiral refractoriness against its genetic manipulation. It emphasizes how crucial it is to understand the endogenous CRISPR-Cas system of *Leptospira* so that it may be utilized to create a tool to investigate gene function via reverse genetics. Strains of pathogenic *L. interrogans* commonly harbor subtypes I-B and I-C of the type I system. However, the CRISPR locus was identified only in the vicinity of the I-B system; thus, the CRISPR-Cas I-B system of *Leptospira* is more likely to function in its immunity against MGEs. The main focus of the current research is to understand the expression stage of the CRISPR-Cas defense system in *Leptospira,* which is essential for RNA-mediated interference of invading nucleic acids.

Across the serovars of *L. interrogans*, two independent *cas* I-B operons spanned a region of variable lengths, hence referred to as the hypervariable regions. These hypervariable regions accommodate the CRISPR locus that contains either a single array or multiple arrays. To identify the CRISPR arrays at the I-B locus of two reference serovars of *L. interrogans* (sv. Copenhageni str. Fiocruz L1-130 and sv. Lai str. 56601), we utilized the CRISPRCasdb database that is a repository of computationally predicted CRISPR-Cas loci through the upgraded version of CRISPRFinder program (CRISPRCasFinder). At the I-B locus of *L. interrogans* sv. Copenhageni, the CRISPRCasdb database revealed a single CRISPR array with unknown orientation. Using RT-PCR, the CRISPR I-B array was found transcriptionally active in sv. Copenhageni. In addition, transcription of this array was identified in the direction of *cas*-operons. Thus, a co-directional arrangement of cas genes and CRISPR array was ascertained at the I-B locus in *Leptospira*. It also indicated the location of the leader [towards *cas2* of the *cas*-operon I (*cas4-cas1-cas2*)] that drives the transcription of the CRISPR I-B array in *Leptospira*.

At the I-B locus of sv. Lai, the CRISPRCasdb database revealed seven CRISPR arrays between the two *cas*-operons. The orientation of each of these seven arrays was projected in the opposite

direction of *cas* I-B operons, which disagreed with the proposed co-directional arrangement of CRISPR-Cas I-B in *Leptospira*. Analysis of spacer sequences associated with CRISPR I-B arrays of sv. Lai suggested a mishap in the database-annotated repeat-spacer boundaries. Moreover, analysis of inter-array sequences revealed three repeat-spacer units that were not apparent in the database. Thus, we failed to apply the CRISPR direction and boundaries projected by the CRISPRCasdb in the genome of *L. interrogans* sv. Lai. Therefore, based on sequence analyses, we redefined the CRISPR arrays at the I-B locus of sv. Lai.

Alignment of repeats composing CRISPR I-B arrays in svs. Copenhageni and Lai revealed sequence variations, mainly in first and terminal repeats. Apart from repeats identical to the repeat consensus, 10 repeat variants were found to compose the CRISPR arrays (I-B) of svs. Copenhageni and Lai. Comparative analysis of hypervariable regions between svs. Copenhageni and Lai indicated that the sequence between cas-operon I (3′ end of *cas2*) to its proximal repeat is more conserved than that of cas-operon II (5′ end of *cas6*) to its proximal repeat. In addition, in both serovars, around 77 bp regions downstream of every CRISPR I-B array are highly conserved. Sequence variation in the hypervariable region among the serovars of *L. interrogans* appears mainly due to CRISPR-associated spacers. Using the spacer sequences of CRISPR I-B arrays in sv. Lai, a conserved trinucleotide 5′-ATG-3′ immediately adjacent to the 5′ ends of targets (protospacers), was predicted. It suggests that sequence 5′-ATG-3′ as a PAM could be employed during the interference process in *Leptospira*.

Like the active transcription of CRISPR I-B array in sv. Copenhageni, transcript expression from the CRISPR I-B arrays of sv. Lai was also evident in RT-PCR. Moreover, in sv. Lai, a continuous transcription of all seven CRISPR I-B arrays, was demonstrated through RT-PCR. Thus, we concluded that a single leader drives the transcription of multiple CRISPR arrays identified in the hypervariable region of sv. Lai. Such continuous transcription of CRISPR arrays in sv. Lai would generate a long precursor transcript corresponding to multiple CRISPR arrays and inter-array sequences. Through q-PCR analysis, around 8 copies of these precursor transcripts (per 10^6 copies of *16S rRNA* transcripts) were recorded in sv. Lai. Moreover, a similar number of precursor CRISPR I-B array transcripts were also estimated in sv. Copenhageni. These q-PCR data suggested an active but basal-level transcription of CRISPR I-B arrays in *Leptospira*.

The functional orientation of CRISPR I-B arrays was endorsed using the Cas6 endoribonuclease of *Leptospira* (LinCas6). *In vitro* RNase assay using the recombinant

Leptospira Cas6 (rLinCas6; N-terminal 6×his tagged) demonstrated the conventional cleavage (upstream to 8 nt from the 3′ end) of cognate repeat RNA substrate in the proposed sense orientation. It thus agreed with the RT-PCR analysis, which demonstrated a co-directional arrangement of CRISPR-Cas I-B in *Leptospira*. In the rLinCas6 RNase assay, exponential decay of repeat RNA substrate suggested that LinCas6 follows single turnover kinetics. This cleavage mode means that rLinCas6 remains bound with the cleaved repeat RNA product. It was advocated by a polyacrylamide gel-based EMSA experiment where rLinCas6 concentration-dependent migration shift of cleaved repeat RNA product was observed. Similar to the cleavage of repeat RNA alone, rLinCas6 cleaved repeat RNA segments within pre-crRNAs to generate mature crRNAs. It demonstrates LinCas6-mediated crRNA biogenesis associated with the CRISPR-Cas I-B system in *Leptospira*. In agreement with the binding of rLinCas6 to the cleaved repeat RNA, the binding of rLinCas6 to mature crRNAs was also evident in the polyacrylamide gel-based EMSA. Migration shifts (EMSA) and elution profile (SEC) of rLinCas6-crRNA complex suggested that under *in vitro* conditions, rLinCas6 can bind to the mature crRNA in two stoichiometries (1:1 or 2:1). For recognition and binding of crRNA, a conserved G-loop of Cas6 is known to be crucial. This G-loop was also evident in LinCas6 in MSA with known Cas6 orthologs. This MSA also suggested a potential catalytic triad (Y27, H38, and Y106) in LinCas6 that might be employed during its acid-base catalysis on RNA substrates. Moreover, the substitution of one of the predicted active site residues (H38) in LinCas6 (LinCas6^{H38A}) resulted in reduced cleavage activity on cognate repeat RNA and pre-crRNA substrates.

Besides LinCas6, LinCas5 was also characterized in this study as a first step towards understanding the Cascade I-B assembly in *Leptospira*. Both 6×his and 6×his-SUMO tagged (N-terminal) LinCas5 were expressed in inclusion bodies of *E. coli* BL21. However, Only 6×his-SUMO-tagged LinCas5 could be dialyzed while remaining in the soluble form. Nevertheless, removing the tag from 6×his-SUMO-tagged LinCas5 resulted in the precipitation of untagged LinCas5. Hence, 6×his-SUMO tagged LinCas5 (rLinCas5) was employed in this study to characterize LinCas5. The rLinCas5 exhibited no nuclease activities on DNA and RNA, suggesting it to be catalytically inactive on nucleic acids. Polyacrylamide gel-based EMSAs suggest that the rLinCas5 may have a weak binding affinity towards pre-crRNA and mature crRNA. However, a clear migration shift of the crRNA-rLinCas6 complex (1:1 stoichiometry) in the presence of rLinCas5 suggested the formation of a higher molecular weight complex containing rLinCas6, rLinCas5 and crRNA.

Furthermore, the nuclease activity of the signature Cas protein (LinCas3) was characterized in this study to understand the physiological requirements of *Leptospira* CRISPR-Cas I-B in the degradation of DNA targets during interference. We could express N-terminal 6×his-SUMO tagged LinCas3, rather than N-terminal 6×his tagged LinCas3, in *E. coli* BL21 cells. Hence, 6×his-SUMO tagged LinCas3 was employed in this study to characterize LinCas3 nuclease. In nuclease assays, rLinCas3 can effectively degrade both ss- and ds-DNA in the presence of Ni^{2+} and Mg^{2+} ions, respectively. The presence of ATP (or non-hydrolyzable ATP derivative) in the reaction intervened with the nuclease activity of rLinCas3 (tagged or untagged version) on circular DNA substrates. It suggested that such intervention by nucleotide in DNase activity on circular DNA substrates is the inherent characteristic of LinCas3 and is not due to the hydrolysis of ATP by LinCas3 ATPase activity. It hinted that nucleotide might regulate the nuclease activity in LinCas3 during target degradation in *Leptospira*. The rLinCas3 also exhibited divalent metal ion-dependent RNase activity on mRNA and RNA oligonucleotide. However, no nuclease activity of rLinCas3 was observed on DNA oligonucleotides. The present study on rLinCas3 indicated that LinCas3 is an active protein with a biologically relevant nuclease function.

The present study highlights the comprehensive understanding of the *Leptospira* CRISPR array transcription and its processing by LinCas6, which is central to RNA-mediated CRISPR-Cas I-B adaptive immunity. This study also underlines the formation of the nucleoprotein complex (crRNA, rLinCas6, and rLinCas5), which is essential for constructing a functional Cascade molecule. In addition, the biochemical analysis of LinCas3 nuclease presented in this study signifies the similarity and differences in the currently known Cas3 nuclease.

5.2 Future Prospects

The current work indicates LinCas6-mediated canonical processing of individual pre-crRNA (*Leptospira* I-B) that resulted in the generation of mature crRNA. This study suggested that in serovars of *L. interrogans*, multiple CRISPR I-B arrays with inter-array sequences transcribe together as a long precursor transcript unit. Further research is needed to understand the crRNA biogenesis from these long precursor transcripts. In addition, the biological relevance of these inter-array regions needs to be elucidated. The present study suggested that rLinCas6 remains bound with mature crRNA under *in vitro* conditions. The fate of Cas6 after crRNA biogenesis has an important impact on the subsequent step in the CRISPR immunity pathway that supplies

the mature crRNA to an effector complex for interference. In some I-B systems, further trimming of mature crRNAs by unknown host nuclease releases the Cas6 from Cascade. Hence, analysis of *Leptospira* Cascade I-B assembled under *in vivo* conditions is needed to determine whether LinCas6 is an integral part of the effector complex.

We relied entirely on CRISPR databases to detect CRISPR arrays in *Leptospira*. While the database located CRISPR arrays in the genomes of *Leptospira* serovars, this study is limited to the CRISPR loci observed in the hypervariable region. CRISPR arrays outside the hypervariable region can be associated with a remotely located *cas* locus in the same genome or can be non-functional. Moreover, not all predicted arrays qualify as true CRISPR arrays. Therefore, the biological relevance of such arrays in *Leptospira* is still unknown and needs to be elucidated for a complete annotation. In the present work, we studied CRISPR arrays in sv. Lai genome as a representative of *Leptospira* strains harboring multiple CRISPR I-B arrays. However, studies on CRISPR loci in other remaining *Leptospira* strains are required to understand their evolution.

This study identified multiple variants of repeat sequences in the CRISPR I-B loci of *L. interrogans* sv. Lai. Although our *in vitro* cleavage assays demonstrated that rLinCas6 could process all these repeat variants within pre-crRNA to generate the mature crRNA, the significance of such variations in the CRISPR-Cas biology of *Leptospira* is still unclear. Knowledge regarding how these nucleotide variations affect the kinetics of rLinCas6-mediated cleavage is restricted in the present work. Besides, sequence analysis of CRISPR loci in sv. Lai showed additional repeat variants that were not annotated in the database CRISPRCasdb. Since our study is limited to database-defined CRISPR loci in *Leptospira*, genome-wide sequence analysis is needed to rule out the existence of additional repeat variants outside the CRISPR loci.

Our RT-PCR analysis revealed transcriptionally active CRISPR I-B arrays, while q-PCR analysis showed only a basal level of pre-crRNA transcripts in the serovars of *L. interrogans*. These analyses were performed using total RNA isolated from *Leptospira* at their mid-log phase. Therefore, our RT-PCR and q-PCR analyses of CRISPR arrays are limited to a particular growth phase of *Leptospira* under the laboratory condition (29 °C). Further studies are needed to investigate whether the CRISPR transcripts are modulated at different time points of *Leptospira* growth or upon any environmental cues.

Besides the biochemical properties of LinCas3 nuclease, the present work indicates the formation of incomplete Cascade (crRNA, rLinCas6 and rLinCas5) under *in vitro* conditions. Further research on *Leptospira* Cascade I-B and LinCas3 is warranted to understand the interference process of the *Leptospira* CRISPR-Cas I-B system. In addition, validation of the predicted consensus PAM sequence in the interference of DNA targets awaits further research. The findings of the present work, such as CRISPR array orientation, location of the leader, and appropriate repeat-spacer boundaries, will prove beneficial in harnessing the endogenous CRISPR-Cas I-B system of *Leptospira* for genome editing or gene silencing applications. It will go a long way in understanding the *Leptospira* pathogenesis, which is still considered to be confined due to the lack of efficient genetic manipulation tools.

Appendix A- Supplementary data to chapter 2

Table S2.1. Primer pair used in the study

Primer names	Sequences (5′-3′)	Purpose
LIC_Cr^{2} S1 forward (^{C2}S1$_{f}$)	AAAGGATCCTTTGATCAAAAGAATT	Detection of CRISPR I-B arrays of svs. Copenhageni and Lai through PCR or RT-PCR
LIC_Cr^{2} S3 reverse (^{C2}S3$_{r}$)	AAGTTTTTCACGGGGTGACG	
LA_Cr^{6} S1 forward (^{6}S1$_{f}$)	CCGTTCTGATTTTTTCTTTTCCT	

LA_Cr^{6} S3 reverse (^{6}S3$_{r}$)	GCGAGCATCGGTAGTTTTACC	
LA_Cr^{7} S1 forward (^{7}S1$_{f}$) LA_Cr^{7} S4 reverse (^{7}S4$_{r}$)	TTGATTGGTGCAGTTGTGCTT TACGCCGGTTCCTCTTTTTTG	
LA_Cr^{9} S1 forward (^{9}S1$_{f}$) LA_Cr^{9} S4 reverse (^{9}S4$_{r}$)	AAAGACAAATCGGTTCATTTGA TTTACGTTTTGAGGATACCTCA	
LA_Cr^{10} S1 forward (^{10}S1$_{f}$) LA_Cr^{10} S4 reverse (^{10}S4$_{r}$)	GAATAACTCGTTCGGAAAGCGT GCAAAGAGAATTGTATTCCGTGT	
LA_Cr^{11} S1 forward (^{11}S1$_{f}$) LA_Cr^{11} S2 reverse (^{11}S2$_{r}$)	CACAACCGTGACAAATATTTGCA CAATGTCGCGGATAAACTTAAGG	
LA_ Cr^{12} S1 forward (^{12}S1$_{f}$) LA_Cr^{12} S2 reverse (^{12}S2$_{r}$)	ACCCGGTTTGCATTTACCGAAG ACAATCCCTCTAAATCTAGCCTCC	
LA_Cr^{6} S1 forward (^{6}S1$_{f}$) LA_Cr^{7} S4 reverse (^{7}S4$_{r}$)	CCGTTCTGATTTTTTCTTTTCCT TACGCCGGTTCCTCTTTTTTG	Detection of partially overlapping CRISPR transcripts (LA_Cr^{6} S1 to Cr^{7} S4, LA_Cr^{7} S2 to Cr^{9} S4, and LA_Cr^{9} S2 to Cr^{10} S4) through RT-PCR
LA_Cr^{7} S2 forward (^{7}S2$_{f}$) LA_Cr^{9} S4 reverse (^{9}S4$_{r}$)	CCGTTCTGATTTTTTCTTTTCCT TTTACGTTTTGAGGATACCTCA	
LA_Cr^{9} S2 forward (^{9}S2$_{f}$) LA_Cr^{10} S4 reverse (^{10}S4$_{r}$)	AAAGACAAATCGGTTCATTTGA TTTACGTTTTGAGGATACCTCA	

Table S2.2. Details of LA_Cr$^{6\text{-}12}$-associated repeats and spacers provided by the database CRISPRCasdb

CRISPR arrays	Repeats and spacers	Coordinate start…end	Sequence (5′-3′)	Length (bp)
LA_Cr^{6}	R1	3163254…3163281	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S1	3163282…3163324	TTGAGCACCCGTTCTGATTTTTTCTTTTCCTTCCTTTTGTTAT	43
	R2	3163325…3163352	CTGAATATAACTTTGATGCCGTTAGGCG	28

	S2	3163353…3163395	TTGAGCACACCCCACGATACTACCTGTCAGACCGTGCCCGGAT	43
	R3	3163396…3163432	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S3	3163433…3163466	TTGAGCACGCACTCCTCGAACTGGTAAAACTACCGATGCTCGC	43
	R4	3163467…3163494	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[7]	R1	3163731…3163758	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S1	3163759…3163801	TTGAGCACCCGTTCTGATTTTTTCTTTTCCTTCCTTTTGTTAT	43
	R2	3163802…3163829	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S2	3163830…3163873	TTGAGCACCAGTAGATTTTGGATACACAAACCCGTTTGTGTTTC	44
	R3	3163874…3163901	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S3	3163902…3163944	TTGAGCACGAATACAACCTCTTCAAAAAAGAGGAACCGGCGTA	43
	R4	3163945…3163972	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[8]	R1	3164138…3164165	CTTACAAAAATCGGGATGCCGGTAGGCG	28
	S1	3164166…3164210	TTGAGCACAAGGAAACAGGTCTCTTAAGTTTAGCACCTGCTGCAG	45
	R2	3164211…3164238	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[9]	R1	3164476…3164503	CTGAATATAACTTTGATGCCGTTAGGTG	28
	S1	3164504…3164546	TTGAGCACAAAGACAAATCGGTTCATTTGATTTTTTCGGATCT	43
	R2	3164547…3164574	CTGAATATAACTTTGATGCCGGTAGGCG	28
	S2	3164575…3164618	TTGAGCACGAGTTAAATTCTGCCCACTCCATGGCCTAAAATCAG	44
	R3	3164619…3164646	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S3	3164647…3164690	TTGAGCACAGGGGCTATAAAATTGAGGTATCCTCAAAACGTAAA	44
	R4	3164691…3164718	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[10]	R1	3164839…3164876	CTTACAAAAATCGGGATGCCGGTAGGCG	28
	S1	3164877…3164908	TTGAGCACGAATAACTCGTTCGGAAAGCGTTCTGCGGATCTT	42
	R2	3164909…3164936	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S2	3164937…3164981	TTGAGCACTCGTAAAGATCGTCTGCGTGTTCGTCGTGATACGTGT	45
	R3	3164982…3165009	CTGAATATAACTTTGATGCCGGTAGGCG	28
	S3	3165010…3165053	TTGAGCACAGCATAGCGGACGTGTTCTTTGTTCGTTTTATCGAA	44
	R4	3165054…3165081	CTGAATATAACTTTGATGCCGTTAGGCG	28
	S4	3165082…3165125	TTGAGCACAGCGTGCAACAAAACACGGAATACAATTCTCTTTGC	44
	R5	3165126…3165153	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[11]	R1	3165387…3165413	CTGAATATAACTTTGATGCCATTAGGCG	28
	S1	3165414…3165458	TTGAGCACGTGCCTTGAGAGACCTTAAGTTTATCCGCGACATTG	44
	R2	3165459…3165486	CTGAATATAACTTTGATGCCGTTAGGCG	28
LA_Cr[12]	R1	3165651…3165678	CTTACAAAAATCGGGATGCCGGTAGGCG	28
	S1	3165679…3165722	TTGAGCACACCCGGTTTGCATTTACCGAAGTCCAAATCAATTTC	44
	R2	3165723…3165750	CTGAATATAACTTTGATGCCGGTAGGCG	28
	S2	3165751…3165793	TTGAGCACTAAGAGTAAGTGGAGGCTAGATTTAGAGGGATTGT	43
	R3	3165794…3165821	CTGAATATAACTTTGATGCCGTTAGGCG	28

Table S2.3. Repeats and spacers of redefined arrays LA_Cr[6-12]

CRISPR arrays	Repeats (5′-3′)	Spacers (5′-3′)
LA_Cr[6]	R1: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R2: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R3: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R4: CTGAATATAACTTTGATGCCGTTAGGCGGTTTAGAA	S1: CCGTTCTGATTTTTTCTTTTCCTTCCTTTTGTTAT S2: ACCCCACGATACTACCTGTCAGACCGTGCCCGGAT S3: GCACTCCTCGAACTGGTAAAACTACCGATGCTCGC
LA_Cr[7]	R1: CTTACAAAAATCGGGTGCCGGTAGGCGTTGAGCAC	S1: TTGATTGGTGCAGTTGTGCTTGTTTTGATTGGTGT

	R2: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R3: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R4: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R5: CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAT	S2: CCGTTCTGATTTTTTCTTTTCCTTCCTTTTGTTAT S3: CAGTAGATTTTGGATACACAAACCCGTTTGTGTTTC S4: GAATACAACCTCTTCAAAAAAGAGGAACCGGCGTA
LA_Cr[8]	R1: CTTACAAAAATCGGGATGCCGGTAGGCGTTGAGCAC R2: CTGAATATAACTTTGATGCCGTTAGGCGGTTTAGAA	S1: AAGGAAACAGGTCTCTTAAGTTTAGCACCTGCTGCAG
LA_Cr[9]	R1: CTTACAAAAATCGGGATGTCGGTAGGCGTTGAGTAC R2: CTGAATATAACTTTGATGCCGTTAGGTGTTGAGCAC R3: CTGAATATAACTTTGATGCCGGTAGGCGTTGAGCAC R4: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R5: CTGAATATAACTTTGATGCCGTTAGGCGGTTTAGAA	S1: TGAGTATGCAAATGAGCTTCGGCTTCGAATTCCCT S2: AAAGACAAATCGGTTCATTTGATTTTTTCGGATCT S3: GAGTTAAATTCTGCCCACTCCATGGCCTAAAATCAG S4: AGGGGCTATAAAATTGAGGTATCCTCAAAACGTAAA
LA_Cr[10]	R1: CTTACAAAAATCGGGATGCCGGTAGGCGTTGAGCAC R2: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R3: CTGAATATAACTTTGATGCCGGTAGGCGTTGAGCAC R4: CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC R5: CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAC	S1: GAATAACTCGTTCGGAAAGCGTTCTGCGGATCTT S2: TCGTAAAGATCGTCTGCGTGTTCGTCGTGATACGTGT S3: AGCATAGCGGACGTGTTCTTTGTTCGTTTTATCGAA S4: AGCGTGCAACAAAACACGGAATACAATTCTCTTTGC
LA_Cr[11]	R1: CTTACAAAAATCGGGTGCCGGTAGGCGTTGAGCAC R2: CTGAATATAACTTTGATGCCATTAGGCGTTGAGCAC R3: CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAC	S1: CACAACCGTGACAAATATTTGCAAATCGTTCGACT S2: GTGCCTTGAGAGACCTTAAGTTTATCCGCGACATTG
LA_Cr[12]	R1: CTTACAAAAATCGGGATGCCGGTAGGCGTTGAGCAC R2: CTGAATATAACTTTGATGCCGGTAGGCGTTGAGCAC R3: CTGAATATAACTTTGATGCCGTTAGGCGATTTAGAC	S1: ACCCGGTTTGCATTTACCGAAGTCCAAATCAATTTC S2: TAAGAGTAAGTGGAGGCTAGATTTAGAGGGATTGT

List of publications

Publications from thesis work

1. **Prakash A**, Kumar M. Characterizing the transcripts of *Leptospira* CRISPR IB array and its processing with endoribonuclease LinCas6. *International Journal of Biological Macromolecules.* 2021;182:785-795

2. **Prakash A**, Kumar M. Transcriptional analysis of CRISPR IB arrays of *Leptospira interrogans* serovar Lai and its processing by Cas6. *Frontiers in Microbiology.* 2022;13

Publications from other collaborative research work

1. Dixit B, **Prakash A**, Kumar P et al. The core Cas1 protein of CRISPR-Cas IB in *Leptospira* shows metal-tunable nuclease activity. *Current Research in Microbial Sciences.* 2021;2:100059

2. Ghosh KK, **Prakash A**, Dhara A et al. Role of supramolecule ErpY-like lipoprotein of *Leptospira* in thrombin-catalyzed fibrin clot inhibition and binding to complement factors H and I, and its diagnostic potential. *Infection and immunity.* 2019;87:e00536-00519

3. Ghosh KK, **Prakash A**, Balamurugan V et al. Catecholamine-modulated novel surface-exposed adhesin LIC20035 of *Leptospira* spp. binds host extracellular matrix components and is recognized by the host during infection. *Applied and environmental microbiology.* 2018;84:e02360-02317

4. Ghosh KK, **Prakash A**, Shrivastav P et al. Evaluation of a novel outer membrane surface-exposed protein, LIC13341 of *Leptospira*, as an adhesin and serodiagnostic candidate marker for leptospirosis. *Microbiology.* 2018;164:1023-1037

Presentations in conferences

1. **Aman Prakash**, and Manish Kumar (2016) “Cloning and expression of novel Lon Protease of *Leptospira interrogans* Copenhageni strain Fiocruz L1-130.”, Global Symposium on “Animal Health: Newer Technologies and their Applications” in Veterinary College Khanapara.

2. **Aman Prakash**, and Manish Kumar (2016) "Characterization of a novel Cas5 protein of CRISPR-Cas type I-B in pathogenic *Leptospira interrogans*.", 57th Annual Conference of Association of Microbiologists of India and International symposium on "Microbes and Biosphere" in Gauhati University, Assam.

3. **Aman Prakash**, Karukriti Kaushik Ghosh, and Manish Kumar (2018) "Characterization of *Leptospira* putative lipoprotein LIC11966 and its serological diagnostic application in diverse hosts including humans", National Conference VIBCON-2018 on "Innovative Biotechnological Approaches for Improving Animal Health and Productivity" in Nagaland.

4. **Aman Prakash**, and Manish Kumar (2018) "Identification and characterization of essential elements involved in CRISPR expression and maturation in pathogenic *Leptospira interrogans*", International Conference on Recent Research in Biomedical Engineering, Cancer Biology, Stem Cells, Bioinformatics and Applied Biotechnology (BECBAB-2018) in JNU, New Delhi.

5. **Aman Prakash**, and Manish Kumar (2019) "Potential application of supramolecule ErpY-Like lipoprotein of *Leptospira* as anticoagulant, complement regulation, and leptospirosis diagnostic kit development", Global Bio-India 2019 in New Delhi.

Workshop attended

1. CDC Sponsored "Workshop on Laboratory Capacity Building for Leptospirosis" organized by ICAR-NIVEDI, Yelahanka, Bengaluru (2017).

REFERENCES

Adler, B. and A. de la Peña Moctezuma (2010). "Leptospira and leptospirosis." Veterinary microbiology **140**(3): 287-296.

Agari, Y., K. Sakamoto, et al. (2010). "Transcription profile of Thermus thermophilus CRISPR systems after phage infection." Journal of molecular biology **395**(2): 270-281.

Amitai, G. and R. Sorek (2016). "CRISPR–Cas adaptation: insights into the mechanism of action." Nature Reviews Microbiology **14**(2): 67-76.

Backstedt, B. T., O. Buyuktanir, et al. (2015). "Efficient detection of pathogenic leptospires using 16S ribosomal RNA." PLoS One **10**(6): e0128913.

Barrangou, R., C. Fremaux, et al. (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-1712.

Barrangou, R. and P. Horvath (2017). "A decade of discovery: CRISPR functions and applications." Nature microbiology **2**(7): 1-9.

Behler, J. and W. R. Hess (2020). "Approaches to study CRISPR RNA biogenesis and the key players involved." Methods **172**: 12-26.

Beloglazova, N., P. Petit, et al. (2011). "Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference." The EMBO journal **30**(22): 4616-4627.

Bernhart, S. H., I. L. Hofacker, et al. (2008). "RNAalifold: improved consensus structure prediction for RNA alignments." BMC bioinformatics **9**(1): 474.

Bharti, A. R., J. E. Nally, et al. (2003). "Leptospirosis: a zoonotic disease of global importance." The Lancet infectious diseases **3**(12): 757-771.

Bharti, A. R., J. E. Nally, et al. (2003). "Leptospirosis: a zoonotic disease of global importance." Lancet Infect. Dis **3**.

Biswas, A., J. N. Gagnon, et al. (2013). "CRISPRTarget: bioinformatic prediction and analysis of crRNA targets." RNA biology **10**(5): 817-827.

Blow, M. J., T. A. Clark, et al. (2016). "The epigenomic landscape of prokaryotes." PLoS genetics **12**(2): e1005854.

Brendel, J., B. Stoll, et al. (2014). "A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in Haloferax volcanii." Journal of Biological Chemistry **289**(10): 7164-7177.

Brouns, S. J., M. M. Jore, et al. (2008). "Small CRISPR RNAs guide antiviral defense in prokaryotes." Science **321**(5891): 960-964.

Cameron, C. E. (2015). Leptospiral structure, physiology, and metabolism. Leptospira and Leptospirosis, Springer**:** 21-41.

Carte, J., R. T. Christopher, et al. (2014). "The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in S treptococcus thermophilus." Molecular microbiology **93**(1): 98-112.

Carte, J., N. T. Pfister, et al. (2010). "Binding and cleavage of CRISPR RNA by Cas6." Rna **16**(11): 2181-2188.

Carte, J., R. Wang, et al. (2008). "Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes." Genes & development **22**(24): 3489-3496.

Charpentier, E., H. Richter, et al. (2015). "Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity." FEMS microbiology reviews: fuv023.

Cheng, F., L. Gong, et al. (2017). "Harnessing the native type IB CRISPR-Cas for genome editing in a polyploid archaeon." Journal of Genetics and Genomics **44**(11): 541-548.

Cochrane, J. C. and S. A. Strobel (2008). "Catalytic strategies of self-cleaving ribozymes." Accounts of chemical research **41**(8): 1027-1035.

Colavecchio, A., B. Cadieux, et al. (2017). "Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the Enterobacteriaceae family–a review." Frontiers in Microbiology **8**: 1108.

Costa, F., J. E. Hagan, et al. (2015). "Global morbidity and mortality of leptospirosis: a systematic review." PLoS neglected tropical diseases **9**(9): e0003898.

Costa, S., A. Almeida, et al. (2014). "Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system." Frontiers in Microbiology **5**: 63.

Croda, J., C. P. Figueira, et al. (2008). "Targeted mutagenesis in pathogenic Leptospira species: disruption of the LigB gene does not affect virulence in animal models of leptospirosis." Infection and immunity **76**(12): 5826-5833.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome research **14**(6): 1188-1190.

Cullen, P. A., D. A. Haake, et al. (2003). "LipL21 is a novel surface-exposed lipoprotein of pathogenic Leptospira species." Infection and immunity **71**(5): 2414-2421.

Deltcheva, E., K. Chylinski, et al. (2011). "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III." Nature **471**(7340): 602.

Deveau, H., R. Barrangou, et al. (2008). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus." Journal of bacteriology **190**(4): 1390-1400.

Dixit, B., K. K. Ghosh, et al. (2016). "Dual Nuclease Activity of a Cas2 protein in CRISPR-Cas subtype I-B of Leptospira interrogans." FEBS letters.

Dixit, B., K. K. Ghosh, et al. (2016). "Dual nuclease activity of a Cas2 protein in CRISPR–Cas subtype I-B of Leptospira interrogans." FEBS Letters **590**(7): 1002-1016.

Drabavicius, G., T. Sinkunas, et al. (2018). "DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system." EMBO reports **19**(7): e45543.

Dupureur, C. M. (2008). "Roles of metal ions in nucleases." Current opinion in chemical biology **12**(2): 250-255.

East-Seletsky, A., M. R. O'Connell, et al. (2016). "Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection." Nature **538**(7624): 270-273.

Faine S, A. B., Bolin C and Perolat P (1999). "Leptospira and Leptospirosis." Medisci(2nd ed.).

Fernandes, L. and A. Nascimento (2022). "A Novel Breakthrough in Leptospira spp. Mutagenesis: Knockout by Combination of CRISPR/Cas9 and Non-homologous End-Joining Systems." Frontiers in Microbiology **13**.

Fernandes, L. G. V., L. Guaman, et al. (2019). "Gene silencing based on RNA-guided catalytically inactive Cas9 (dCas9): a new tool for genetic engineering in Leptospira." Scientific reports **9**(1): 1-14.

Fernandes, L. G. V., R. Hornsby, et al. (2021). "Genetic manipulation of pathogenic Leptospira: CRISPR interference (CRISPRi)-mediated gene silencing and rapid mutant recovery at 37 C." Scientific reports **11**(1): 1-12.

Fonfara, I., H. Richter, et al. (2016). "The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA." Nature **532**(7600): 517-521.

Fouts, D. E., M. A. Matthias, et al. (2016). "What makes a bacterial species pathogenic?: Comparative genomic analysis of the genus Leptospira." PLoS Negl Trop Dis **10**(2): e0004403.

García-Aljaro, C., E. Ballesté, et al. (2017). "Beyond the canonical strategies of horizontal gene transfer in prokaryotes." Current opinion in microbiology **38**: 95-105.

Garneau, J. E., M.-È. Dupuis, et al. (2010). "The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA." Nature **468**(7320): 67.

Garside, E. L., M. J. Schellenberg, et al. (2012). "Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases." Rna **18**(11): 2020-2028.

Gerdes, K., S. K. Christensen, et al. (2005). "Prokaryotic toxin–antitoxin stress response loci." Nature Reviews Microbiology **3**(5): 371-382.

Gesner, E. M., M. J. Schellenberg, et al. (2011). "Recognition and maturation of effector RNAs in a CRISPR interference pathway." Nature structural & molecular biology **18**(6): 688-692.

Ghosh, K. K., A. Prakash, et al. (2018). "Catecholamine-modulated novel surface-exposed adhesin LIC20035 of Leptospira spp. binds host extracellular matrix components and is recognized by the host during infection." Applied and environmental microbiology **84**(6): e02360-02317.

Ghosh, K. K., A. Prakash, et al. (2018). "Evaluation of a novel outer membrane surface-exposed protein, LIC13341 of Leptospira, as an adhesin and serodiagnostic candidate marker for leptospirosis." Microbiology **164**(8): 1023-1037.

Goldfarb, T., H. Sberro, et al. (2015). "BREX is a novel phage resistance system widespread in microbial genomes." The EMBO journal **34**(2): 169-183.

Gong, B., M. Shin, et al. (2014). "Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3." Proceedings of the National Academy of Sciences **111**(46): 16359-16364.

Grainy, J., S. Garrett, et al. (2019). "CRISPR repeat sequences and relative spacing specify DNA integration by Pyrococcus furiosus Cas1 and Cas2." Nucleic acids research **47**(14): 7518-7531.

Grissa, I., G. Vergnaud, et al. (2007). "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats." Nucleic acids research **35**(suppl 2): W52-W57.

Gudbergsdottir, S., L. Deng, et al. (2011). "Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers." Molecular microbiology **79**(1): 35-49.

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

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

Harrington, L. B., E. Ma, et al. (2020). "A scoutRNA is required for some type V CRISPR-Cas systems." Molecular cell **79**(3): 416-424. e415.

Hartskeerl, R., M. Collares-Pereira, et al. (2011). "Emergence, control and re-emerging leptospirosis: dynamics of infection in the changing world." Clinical Microbiology and Infection **17**(4): 494-501.

Hatoum-Aslan, A., I. Maniv, et al. (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 **108**(52): 21218-21222.

Haurwitz, R. E., M. Jinek, et al. (2010). "Sequence-and structure-specific RNA processing by a CRISPR endonuclease." Science **329**(5997): 1355-1358.

Haurwitz, R. E., S. H. Sternberg, et al. (2012). "Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA." The EMBO journal **31**(12): 2824-2832.

He, L., M. St John James, et al. (2020). "Cas3 protein—a review of a multi-tasking machine." Genes **11**(2): 208.

Hille, F., H. Richter, et al. (2018). "The biology of CRISPR-Cas: backward and forward." Cell **172**(6): 1239-1259.

Hochstrasser, M. L. and J. A. Doudna (2015). "Cutting it close: CRISPR-associated endoribonuclease structure and function." Trends in biochemical sciences **40**(1): 58-66.

Hochstrasser, M. L., D. W. Taylor, et al. (2014). "CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference." Proceedings of the National Academy of Sciences **111**(18): 6618-6623.

Horvath, P., D. A. Romero, et al. (2008). "Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus." Journal of bacteriology **190**(4): 1401-1412.

Howard, J. A., S. Delmas, et al. (2011). "Helicase dissociation and annealing of RNA-DNA hybrids by Escherichia coli Cas3 protein." Biochemical Journal **439**(1): 85-95.

Huo, Y., K. H. Nam, et al. (2014). "Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation." Nature structural & molecular biology **21**(9): 771-777.

Jackson, R. N., S. M. Golden, et al. (2014). "Crystal structure of the CRISPR RNA–guided surveillance complex from Escherichia coli." Science **345**(6203): 1473-1479.

Jackson, R. N., M. Lavin, et al. (2014). "Fitting CRISPR-associated Cas3 into the helicase family tree." Current opinion in structural biology **24**: 106-114.

Jansen, R., J. Embden, et al. (2002). "Identification of genes that are associated with DNA repeats in prokaryotes." Molecular microbiology **43**(6): 1565-1575.

Jesser, R., J. Behler, et al. (2019). "Biochemical analysis of the Cas6-1 RNA endonuclease associated with the subtype ID CRISPR-Cas system in Synechocystis sp. PCC 6803." RNA biology **16**(4): 481-491.

Jinek, M., K. Chylinski, et al. (2012). "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity." Science **337**(6096): 816-821.

Jore, M. M., S. J. Brouns, et al. (2012). "RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements." Cold Spring Harbor perspectives in biology **4**(6): a003657.

Jore, M. M., M. Lundgren, et al. (2011). "Structural basis for CRISPR RNA-guided DNA recognition by Cascade." Nature structural & molecular biology **18**(5): 529.

Karpagam, K. B. and B. Ganesh (2020). "Leptospirosis: a neglected tropical zoonotic infection of public health importance—an updated review." European Journal of Clinical Microbiology & Infectious Diseases **39**(5): 835-846.

Katoh, H., M. Yoshinaga, et al. (1986). "Kinetic studies on turtle pancreatic ribonuclease: a comparative study of the base specificities of the B2 and P0 sites of bovine pancreatic ribonuclease A and turtle pancreatic ribonuclease." Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology **873**(3): 367-371.

Kibbe, W. A. (2007). "OligoCalc: an online oligonucleotide properties calculator." Nucleic acids research **35**(suppl_2): W43-W46.

Kim, S., L. Loeff, et al. (2020). "Selective loading and processing of prespacers for precise CRISPR adaptation." Nature **579**(7797): 141-145.

Ko, A. I., C. Goarant, et al. (2009). "Leptospira: the dawn of the molecular genetics era for an emerging zoonotic pathogen." Nature Reviews Microbiology **7**(10): 736-747.

König, S. L., P. S. Liyanage, et al. (2013). "Helicase-mediated changes in RNA structure at the single-molecule level." RNA biology **10**(1): 133-148.

Koo, Y., D. Ka, et al. (2013). "Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system." Journal of molecular biology **425**(20): 3799-3810.

Koonin, E. V. and M. Krupovic (2015). "Evolution of adaptive immunity from transposable elements combined with innate immune systems." Nature Reviews Genetics **16**(3): 184-192.

Koonin, E. V., K. S. Makarova, et al. (2017). "Evolutionary genomics of defense systems in archaea and bacteria." Annual review of microbiology **71**: 233.

Koonin, E. V., K. S. Makarova, et al. (2017). "Diversity, classification and evolution of CRISPR-Cas systems." Current opinion in microbiology **37**: 67-78.

Kumar, M., X. Yang, et al. (2010). "BBA52 facilitates Borrelia burgdorferi transmission from feeding ticks to murine hosts." The Journal of infectious diseases **201**(7): 1084-1095.

Kunin, V., R. Sorek, et al. (2007). "Evolutionary conservation of sequence and secondary structures in CRISPR repeats." Genome biology **8**(4): 1-7.

Labrie, S. J., J. E. Samson, et al. (2010). "Bacteriophage resistance mechanisms." Nature Reviews Microbiology **8**(5): 317-327.

Lee, H., Y. Zhou, et al. (2018). "Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays." Molecular cell **70**(1): 48-59. e45.

Lee, H. Y., R. E. Haurwitz, et al. (2013). "RNA–protein analysis using a conditional CRISPR nuclease." Proceedings of the National Academy of Sciences **110**(14): 5416-5421.

LeFebvre, R. (2004). "Spiral-Curved Organisms: Leptospira." Veterinary microbiology: 148-152.

Li, Y., S. Pan, et al. (2016). "Harnessing Type I and Type III CRISPR-Cas systems for genome editing." Nucleic acids research **44**(4): e34-e34.

Lieber, M. R. (2010). "The mechanism of double-strand DNA break repair by the nonhomologous DNA end joining pathway." Annual review of biochemistry **79**: 181.

Lin, M., O. Surujballi, et al. (1997). "Identification of a 35-kilodalton serovar-cross-reactive flagellar protein, FlaB, from Leptospira interrogans by N-terminal sequencing, gene cloning, and sequence analysis." Infection and immunity **65**(10): 4355-4359.

Lintner, N. G., M. Kerou, et al. (2011). "Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE)." Journal of Biological Chemistry **286**(24): 21643-21656.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method." Methods **25**(4): 402-408.

Madeira, F., Y. M. Park, et al. (2019). "The EMBL-EBI search and sequence analysis tools APIs in 2019." Nucleic acids research **47**(W1): W636-W641.

Maier, L.-K., A.-E. Stachler, et al. (2019). "The nuts and bolts of the Haloferax CRISPR-Cas system IB." RNA biology **16**(4): 469-480.

Maikova, A., V. Kreis, et al. (2019). "Using an endogenous CRISPR-Cas system for genome editing in the human pathogen Clostridium difficile." Applied and environmental microbiology **85**(20): e01416-01419.

Makarova, K. S., L. Aravind, et al. (2002). "A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis." Nucleic acids research **30**(2): 482-496.

Makarova, K. S., L. Aravind, et al. (2011). "Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems." Biology direct **6**(1): 1-27.

Makarova, K. S., N. V. Grishin, et al. (2006). "A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action." Biology direct **1**(1): 1-26.

Makarova, K. S., D. H. Haft, et al. (2011). "Evolution and classification of the CRISPR–Cas systems." Nature Reviews Microbiology **9**(6): 467-477.

Makarova, K. S., S. Karamycheva, et al. (2019). "Predicted highly derived class 1 CRISPR-Cas system in Haloarchaea containing diverged Cas5 and Cas7 homologs but no CRISPR array." FEMS microbiology letters **366**(7): fnz079.

Makarova, K. S., Y. I. Wolf, et al. (2015). "An updated evolutionary classification of CRISPR-Cas systems." Nature Reviews Microbiology.

Makarova, K. S., Y. I. Wolf, et al. (2015). "An updated evolutionary classification of CRISPR–Cas systems." Nature Reviews Microbiology **13**(11): 722.

Makarova, K. S., Y. I. Wolf, et al. (2013). "Comparative genomics of defense systems in archaea and bacteria." Nucleic acids research **41**(8): 4360-4377.

Makarova, K. S., Y. I. Wolf, et al. (2020). "Unprecedented Diversity of Unique CRISPR-Cas-Related Systems and Cas1 Homologs in Asgard Archaea." The CRISPR Journal **3**(3): 156-163.

Marraffini, L. A. and E. J. Sontheimer (2010). "CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea." Nature Reviews Genetics **11**(3): 181-190.

Marsden, S., M. Nardelli, et al. (2006). "Unwinding single RNA molecules using helicases involved in eukaryotic translation initiation." Journal of molecular biology **361**(2): 327-335.

Mohamadi, S., S. Z. Bostanabad, et al. (2020). "CRISPR arrays: a review on its mechanism." Journal of Applied Biotechnology Reports **7**(2): 81-86.

Mojica, F. J., C. Díez-Villaseñor, et al. (2009). "Short motif sequences determine the targets of the prokaryotic CRISPR defence system." Microbiology **155**(3): 733-740.

Mojica, F. J., J. García-Martínez, et al. (2005). "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements." Journal of molecular evolution **60**(2): 174-182.

Mosterd, C., G. M. Rousseau, et al. (2021). "A short overview of the CRISPR-Cas adaptation stage." Canadian journal of microbiology **67**(1): 1-12.

Mulepati, S. and S. Bailey (2011). "Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3)." Journal of Biological Chemistry **286**(36): 31896-31903.

Mulepati, S. and S. Bailey (2013). "In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target." Journal of Biological Chemistry **288**(31): 22184-22192.

Mulepati, S., A. Héroux, et al. (2014). "Crystal structure of a CRISPR RNA–guided surveillance complex bound to a ssDNA target." Science **345**(6203): 1479-1484.

Murray, G. L. (2013). "The lipoprotein LipL32, an enigma of leptospiral biology." Veterinary microbiology **162**(2-4): 305-314.

Murray, P., K. Rosenthal, et al. (2009). "Medical microbiology: Mosby Inc."

Musso, D. and B. La Scola (2013). "Laboratory diagnosis of leptospirosis: a challenge." Journal of Microbiology, Immunology and Infection **46**(4): 245-252.

Nam, K. H., C. Haitjema, et al. (2012). "Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype IC/Dvulg CRISPR-Cas system." Structure **20**(9): 1574-1584.

Nickel, L., A. Ulbricht, et al. (2019). "Cross-cleavage activity of Cas6b in crRNA processing of two different CRISPR-Cas systems in Methanosarcina mazei Gö1." RNA biology **16**(4): 492-503.

Niewoehner, O., M. Jinek, et al. (2014). "Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases." Nucleic acids research **42**(2): 1341-1353.

Pappas, C. J., N. Benaroudj, et al. (2015). "A replicative plasmid vector allows efficient complementation of pathogenic Leptospira strains." Appl. Environ. Microbiol. **81**(9): 3176-3181.

Pappas, C. J., N. Benaroudj, et al. (2015). "A replicative plasmid vector allows efficient complementation of pathogenic Leptospira strains." Applied and environmental microbiology **81**(9): 3176-3181.

Parma, D. H., M. Snyder, et al. (1992). "The Rex system of bacteriophage lambda: tolerance and altruistic cell death." Genes & development **6**(3): 497-510.

Peng, W., M. Feng, et al. (2015). "An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference." Nucleic acids research **43**(1): 406-417.

Perolat, P., R. J. Chappel, et al. (1998). "Leptospira fainei sp. nov., isolated from pigs in Australia." International journal of systematic and evolutionary microbiology **48**(3): 851-858.

Picardeau, M. (2017). "Virulence of the zoonotic agent of leptospirosis: still terra incognita?" Nature Reviews Microbiology **15**(5): 297-307.

Picardeau, M., D. M. Bulach, et al. (2008). "Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of Leptospira and the pathogenesis of leptospirosis." PLoS One **3**(2): e1607.

Pinilla-Redondo, R., D. Mayo-Muñoz, et al. (2020). "Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids." Nucleic acids research **48**(4): 2000-2012.

Plagens, A., V. Tripp, et al. (2014). "In vitro assembly and activity of an archaeal CRISPR-Cas type IA Cascade interference complex." Nucleic acids research **42**(8): 5125-5138.

Pourcel, C., G. Salvignol, et al. (2005). "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies." Microbiology **151**(3): 653-663.

Pourcel, C., M. Touchon, et al. (2020). "CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers." Nucleic acids research **48**(D1): D535-D544.

Pyne, M. E., M. R. Bruder, et al. (2016). "Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in Clostridium." Scientific reports **6**(1): 1-15.

Redding, S., S. H. Sternberg, et al. (2015). "Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system." Cell **163**(4): 854-865.

Reeks, J., J. H. Naismith, et al. (2013). "CRISPR interference: a structural perspective." Biochemical Journal **453**(2): 155-166.

Reeks, J., R. D. Sokolowski, et al. (2013). "Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing." Biochemical Journal **452**(2): 223-230.

Reimann, V., O. S. Alkhnbashi, et al. (2017). "Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs." Nucleic acids research **45**(2): 915-925.

Richter, H., S. J. Lange, et al. (2013). "Comparative analysis of Cas6b processing and CRISPR RNA stability." RNA biology **10**(5): 700-707.

Richter, H., J. Zoephel, et al. (2012). "Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis." Nucleic acids research **40**(19): 9887-9896.

Ristow, P., P. Bourhy, et al. (2007). "The OmpA-like protein Loa22 is essential for leptospiral virulence." PLoS pathogens **3**(7): e97.

Robert, X. and P. Gouet (2014). "Deciphering key features in protein structures with the new ENDscript server." Nucleic acids research **42**(W1): W320-W324.

Roelofs, J., A. Suppahia, et al. (2018). "Native gel approaches in studying proteasome assembly and chaperones." The Ubiquitin Proteasome System: 237-260.

Rogers, G. W., N. J. Richter, et al. (1999). "Biochemical and kinetic characterization of the RNA helicase activity of eukaryotic initiation factor 4A." Journal of Biological Chemistry **274**(18): 12236-12244.

Rollie, C., S. Graham, et al. (2018). "Prespacer processing and specific integration in a Type IA CRISPR system." Nucleic acids research **46**(3): 1007-1020.

Sahoo, N., V. Cuello, et al. (2020). CRISPR-Cas9 Genome Editing in Human Cell Lines with Donor Vector Made by Gibson Assembly. RNA Interference and CRISPR Technologies, Springer**:** 365-383.

Sampson, T. R., S. D. Saroj, et al. (2013). "A CRISPR/Cas system mediates bacterial innate immune evasion and virulence." Nature **497**(7448): 254-257.

Sampson, T. R. and D. S. Weiss (2014). "CRISPR-Cas systems: new players in gene regulation and bacterial physiology." Frontiers in cellular and infection microbiology **4**: 37.

Sashital, D. G., M. Jinek, et al. (2011). "An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3." Nature structural & molecular biology **18**(6): 680-687.

Sashital, D. G., B. Wiedenheft, et al. (2012). "Mechanism of foreign DNA selection in a bacterial adaptive immune system." Molecular cell **46**(5): 606-615.

Sasnauskas, G., B. A. Connolly, et al. (2007). "Site-specific DNA transesterification catalyzed by a restriction enzyme." Proceedings of the National Academy of Sciences **104**(7): 2115-2120.

Sefcikova, J., M. Roth, et al. (2017). "Cas6 processes tight and relaxed repeat RNA via multiple mechanisms: A hypothesis." BioEssays **39**(6): 1700019.

Semenova, E., M. M. Jore, et al. (2011). "Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence." Proceedings of the National Academy of Sciences **108**(25): 10098-10103.

Shang, E. S., T. A. Summers, et al. (1996). "Molecular cloning and sequence analysis of the gene encoding LipL41, a surface-exposed lipoprotein of pathogenic Leptospira species." Infection and immunity **64**(6): 2322-2330.

Shao, Y. and H. Li (2013). "Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6." Structure **21**(3): 385-393.

Shao, Y., H. Richter, et al. (2016). "A non-stem-loop CRISPR RNA is processed by dual binding Cas6." Structure **24**(4): 547-554.

Shapiro, R. S., A. Chavez, et al. (2018). "CRISPR-based genomic tools for the manipulation of genetically intractable microorganisms." Nature Reviews Microbiology **16**(6): 333-339.

Sharma, M. and A. Yadav (2008). "Leptospirosis: epidemiology, diagnosis, and control." J Infect Dis Antimicrob Agents **25**(2): 93-103.

Shmakov, S., O. O. Abudayyeh, et al. (2015). "Discovery and functional characterization of diverse class 2 CRISPR-Cas systems." Molecular cell **60**(3): 385-397.

Sinkunas, T., G. Gasiunas, et al. (2011). "Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system." The EMBO journal **30**(7): 1335-1342.

Sinkunas, T., G. Gasiunas, et al. (2013). "In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus." The EMBO journal **32**(3): 385-394.

Smargon, A. A., D. B. Cox, et al. (2017). "Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28." Molecular cell **65**(4): 618-630. e617.

Snyder, L. (1995). "Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents?" Molecular microbiology **15**(3): 415-420.

Sokolowski, R. D., S. Graham, et al. (2014). "Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system." Nucleic acids research **42**(10): 6532-6541.

Sternberg, S. H., R. E. Haurwitz, et al. (2012). "Mechanism of substrate selection by a highly specific CRISPR endoribonuclease." Rna **18**(4): 661-672.

Taylor, D. W., Y. Zhu, et al. (2015). "Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning." Science **348**(6234): 581-585.

Touchon, M., J. A. M. De Sousa, et al. (2017). "Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer." Current opinion in microbiology **38**: 66-73.

Touchon, M. and E. P. Rocha (2010). "The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella." PLoS One **5**(6): e11126.

Van Der Oost, J., E. R. Westra, et al. (2014). "Unravelling the structural and mechanistic basis of CRISPR–Cas systems." Nature Reviews Microbiology **12**(7): 479-492.

Victoriano, A., L. D. Smythe, et al. (2009). "Leptospirosis in the Asia Pacific region." BMC infectious diseases **9**(1): 1.

Vincent, A. T., O. Schiettekatte, et al. (2019). "Revisiting the taxonomy and evolution of pathogenicity of the genus Leptospira through the prism of genomics." PLoS neglected tropical diseases **13**(5): e0007270.

Vinetz, J. M. (2000). "Ten Common Questions About Leptospirosis." Infectious Diseases in Clinical Practice **9**(2): 59-65.

Vink, J. N., J. H. Baijens, et al. (2021). "PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR-Cas systems." Genome biology **22**(1): 1-25.

Wang, R., G. Preamplume, et al. (2011). "Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage." Structure **19**(2): 257-264.

Wang, R., H. Zheng, et al. (2012). "The impact of CRISPR repeat sequence on structures of a Cas6 protein–RNA complex." Protein Science **21**(3): 405-417.

Waterhouse, A. M., J. B. Procter, et al. (2009). "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." Bioinformatics **25**(9): 1189-1191.

Westra, E. R., P. B. van Erp, et al. (2012). "CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3." Molecular cell **46**(5): 595-605.

Wiedenheft, B., E. van Duijn, et al. (2011). "RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions." Proceedings of the National Academy of Sciences **108**(25): 10092-10097.

Wilkinson, M., G. Drabavicius, et al. (2019). "Structure of the DNA-bound spacer capture complex of a type II CRISPR-Cas system." Molecular cell **75**(1): 90-101. e105.

Xiao, G., Y. Yi, et al. (2019). "Characterization of CRISPR-Cas systems in Leptospira reveals potential application of CRISPR in genotyping of Leptospira interrogans." Apmis **127**(4): 202-216.

Yan, W. X., P. Hunnewell, et al. (2019). "Functionally diverse type V CRISPR-Cas systems." Science **363**(6422): 88-91.

Yang, C. (2007). "Leptospirosis in Taiwan-an underestimated infectious disease." Chang Gung medical journal **30**(2): 109.

Yang, W. (2011). "Nucleases: diversity of structure, function and mechanism." Quarterly reviews of biophysics **44**(1): 1.

Yoganand, K. N., M. Muralidharan, et al. (2019). "Fidelity of prespacer capture and processing is governed by the PAM-mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type IE system." Journal of Biological Chemistry **294**(52): 20039-20053.

Yosef, I., M. G. Goren, et al. (2012). "Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli." Nucleic acids research **40**(12): 5569-5576.

You, L., J. Ma, et al. (2019). "Structure studies of the CRISPR-Csm complex reveal mechanism of co-transcriptional interference." Cell **176**(1-2): 239-253. e216.

Young, J. C., B. D. Dill, et al. (2012). "Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in Streptococcus thermophilus." PLoS One **7**(5): e38077.

Zhang, Y., N. Heidrich, et al. (2013). "Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis." Molecular cell **50**(4): 488-503.

Zhao, C., X. Shu, et al. (2017). "Construction of a gene knockdown system based on catalytically inactive ("dead") Cas9 (dCas9) in Staphylococcus aureus." Applied and environmental microbiology **83**(12): e00291-00217.

Zhao, H., G. Sheng, et al. (2014). "Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli." Nature **515**(7525): 147-150.

Zheng, Y., J. Li, et al. (2020). "Endogenous type I CRISPR-Cas: from foreign DNA defense to prokaryotic engineering." Frontiers in bioengineering and biotechnology **8**: 62.