INVESTIGATIONS ON THE MATURATION OF CRISPR RNA IN TYPE I-C CRISPR-CAS SYSTEM

A THESIS
submitted in partial fulfilment of the requirements
for the award of the degree
of

DOCTOR OF PHILOSOPHY

by

ANKITA PUNETHA

DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

OCTOBER 2016

Dedicated to

my dear parents and brother

for their everlasting love, support and patience....

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING

STATEMENT

I do hereby declare that the matter embodied in this thesis entitled "**Investigations on the maturation of CRISPR RNA in type I-C CRISPR-Cas system**" is the result of investigations carried out by me in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the supervision of **Dr. B. Anand**.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work of other investigators are referred.

Ankita Punetha
Roll No: 10610616

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING

CERTIFICATE

This is to certify that work described in the thesis entitled "**Investigations on the maturation of CRISPR RNA in type I-C CRISPR-Cas system**" by Ms. **Ankita Punetha** (Roll No: 10610616), submitted to the Indian Institute of Technology Guwahati, India for the award of the degree of Doctor of Philosophy, is an authentic record of the results obtained from the research work carried out under my supervision at the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India and this work has not been submitted elsewhere for a degree.

Dr. B. Anand
Assistant Professor
(Thesis Supervisor)

Acknowledgements

PhD was a learning experience not only in the academic grounds but also at personal levels. It taught me how to face adversity and move towards betterment. Throughout these years, I received much support from my supervisor, family and friends. It would be difficult to individually acknowledge all those who deserve it, without any serious omissions. However, I will try to express my gratitude to those who supported me in making this thesis possible.

It is with my deepest sense of appreciation that I express my foremost acknowledgement to my research advisor, Dr. B. Anand, Department of Biosciences and Bioengineering for his continuous care, support and encouragement throughout my research work. He gave me ample freedom to conduct my work smoothly. I am also grateful for his scientific advice and insightful knowledge. He is my primary resource for getting my science questions answered and was instrumental in helping me to shape this thesis.

I am thankful to my doctoral committee members, Prof. R. Swaminathan, Prof. Vikash Kumar Dubey of Department of Biosciences and Bioengineering and Dr. Dipankar Bandyopadhyay of Department of Chemical Engineering, for their constructive criticism and precious suggestions.

I owe my gratitude to the Heads of the Department of Biosciences and Bioengineering, Prof. Arun Goyal, Prof. V. Venkata Dasu and Prof. K. Pakshirajan for extending their support in various ways during their respective tenures and also to other Professors in the department, who imparted the subject knowledge during my course work.

I thank all my group members for their critical comments and suggestions on this work and indispensable help during the due course of PhD. Especially, I would like to acknowledge Siddharth Nimkar, who generously provided substrate for an experimentation.

I am thankful to Dr. Sundar Durai of Department of Biochemical Engineering and Biotechnology, IIT Delhi for providing opportunity to work in various projects and scientific writings which gave me vivid experience and helped in staging my research career. Moreover, he is a source of inspiration not only for his academic soundness and talent, but also for his good human values. I would also take the opportunity to thank Mrs. Vidhya, W/o Dr. Sundar Durai, for providing support and motivation. Her warm and caring nature is worth admiring.

I will always be grateful to Dr. Gitanjali Yadav and Prof. Sushil Kumar of National Institute of Plant Genome Research for having faith in my abilities and providing me the first research platform, which helped me to build my self-confidence and also developed my scientific temperament.

I am also thankful to all the faculty members of Govt. P. G. College Pithoragarh of Kumaon University and faculty members of Banasthali University, Rajasthan who imparted me the subject knowledge and helped to build my basic foundation.

I take this opportunity to express my love and gratitude to my parents – Dr. A. K. Punetha, a senior paediatrician and Dr. Nirmala Punetha, senior gynaecologist and chief medical superintendent in government hospital Pithoragarh and brother Dr. Arpit Punetha, a medical officer, for being near to me all through my career, providing constant support and encouragement. I am also thankful to my cousin Mr. Prateek Oli, whose friendly and jolly nature is worth admiring.

I am thankful and grateful to the two inspiring family members who are a constant source of inspiration. Prof. S. K. Joshi, a renowned physicist, formerly the Director General CSIR and former chairman of UGC, who motivated and encouraged to work towards betterment, skill development and gaining better exposure and experience. Mr. B. D. Dungrakoti, formerly the Director of Himalayan Geology, Geological Survey of India, who constantly followed up my work progress, which further motivated me to stay focused and

strive for excellence. I would also like to convey my regards and gratitude to the departed soul of Mrs. Dungrakoti, whose sweet memories will always remain alive. She was an icon to learn human values. She was always welcoming, warmly and cheerful till her last breath. Her voice never had the ounce of reflection of the sufferings she was undergoing, due to her health issues. Hope the soul of the wonderful person rest in peace.

I would express my humble indebt thanks to my friend, Ms. Senjam Shantirani for her constant encouragement and moral support throughout my PhD work and serving as source of inspiration, which helped me to sail through my adversities. Her positivity helped to revitalize and restored enthusiasm.

I cherish my close association with my friends and my batch mates and other friendly faces and well-wishers always abound in my memories. Their presence made the environment lively and chirpy.

Finally, a whole hearted thanks to Department of Biosciences and Bioengineering, IIT Guwahati for providing me an opportunity to carry out my research and for providing the required platform and infrastructure. I am thankful to all the office staff members and express appreciation for their cordial and cooperative nature. I am grateful to IIT Guwahati for providing me resources and congenial environment to carry out my research work peacefully.

This last paragraph I want to dedicate to thank all the funding agencies which supported my work in form of grants awarded to my supervisor. I acknowledge the support provided by Department of Biotechnology and Board of Research in Nuclear Sciences.

I am also highly thankful to MHRD, GOI for proving the financial assistance during my PhD tenure.

May God Bless all helping hands!

Ankita Punetha
October 2016

Contents

List of Figures

Abbreviations

CRISPR – Clustered regularly interspaced short palindromic repeat

Cas – CRISPR-associated sequence

pre-crRNA – pre-CRISPR RNA

crRNA – CRISPR RNA

crRNP – CRISPR RNA Ribonucleoprotein complex

Cascade – CRISPR-associated complex for antiviral defense

tracrRNA – trans-activating CRISPR RNA

PAM – Protospacer adjacent motif

HGT – Horizontal gene transfer

R-M – Restriction modification

Abi – Abortive infection

TA – Toxin-antitoxin

HR – Homologous repair

NHEJ – Non-homologous end joining

ZFN – Zinc finger nuclease

REase – Restriction endonuclease

MTases – Methyltransferases

1.1. Prokaryotes and their defense systems

Prokaryotes are single celled organisms that are the earliest and most primitive forms of life on earth. They have no true nucleus, *i.e.,* their DNA is not contained within a membrane but is coiled up in a region of the cytoplasm called the nucleoid and all the other intracellular components – proteins and metabolites – are located together in the same volume enclosed by the cell membrane, rather than in separate cellular compartments. However, they do possess protein-based micro compartments, which are thought to act as primitive organelles. Prokaryotic cells are not as complex as eukaryotic cells yet their metabolism is far more varied than that of eukaryotes, leading to many highly distinct prokaryotic types. For example, in addition to using photosynthesis or organic compounds for energy as eukaryotes, the prokaryotes may also obtain energy from inorganic compounds such as hydrogen sulfide. This enables prokaryotes to thrive in various types of environments including extreme habitats such as snow surface of Antarctica, hydrothermal vents, hot springs, swamps, oceans, wetlands, and the guts of animals including humans. Prokaryotic cells can be divided into bacteria and archaea based on the three domain system (Woese et al., 1990). Both bacteria and archaea reproduce through asexual reproduction, usually by binary fission. The genetic exchange and recombination may occur via horizontal gene transfer which involves the transfer of DNA between two cells. Among bacteria these include bacterial virus (bacteriophage) mediated transduction, plasmid-mediated conjugation and natural transformation. The transduction of bacterial genes by bacteriophage appears to reflect an occasional error during intracellular assembly of virus particles, rather than an adaptation of the host bacteria, as the transfer of bacterial DNA is under the control of the bacteriophage's genes rather than the bacterial genes. Among archaea, the DNA transfer

seems to occur by direct contact or by formation of cytoplasmic bridges between cells (Rosenshine et al., 1989). Further, exposure to DNA damaging agents induces cellular aggregation which enhances DNA transfer among cells to facilitate increased repair of damaged DNA via homologous recombination (Frols et al., 2008). All organisms need to continuously adapt to the changes in their environment. Bacteria and archaea can rapidly acquire new traits by horizontal gene transfer in response to their environmental stimulus that may contribute to their survival. However, because new DNA may also cause damage, removal of imported DNA and protection against invading selfish/mobile DNA elements is also important. Hence, there is a delicate balance between DNA uptake and DNA degradation.

In order to survive, all organisms must overcome their invaders or predators. The prokaryotes and their viral predators coexist in natural and man-made environments. It has been estimated that in most environments, phages outnumber their bacterial hosts by approximately tenfold (Bergh et al., 1989; Brussow and Hendrix, 2002; Chibani-Chennoufi et al., 2004) and thus bacteria have a constant threat of getting infected by phages in their surroundings. The pathogen-host relationships and the need of coevolution have been well described by Red Queen hypothesis proposed by Leigh Van Valen in 1973 (Van Valen, 1973) which was later termed as the “evolutionary arms race” hypothesis. The original theory proposed that in competitive environmental interactions, such as those in a prey-predator relationship, changes on only one side may lead to near extinction of the other side. The only way the second side can maintain its fitness is by counter adaptation. This leads to continuous variation and selection towards adaptation of the host and counter adaptations on the side of the pathogen. Thus, species have to constantly evolve to stay at the same fitness level, which can be termed as an evolutionary arms race in host-pathogen relationships. At the microbial level, the phages have a profound impact on the evolution of bacterial and archaeal species,

due to their extremely rapid evolution and turnover that causes acute pressure on microbial communities to evade infection and killing by phages.

Pitted against a hostile environment, organisms seem to have developed multi-layered antiviral defense systems to overcome the predators. In case of eukaryotes, both innate and adaptive immune systems are prevalent. For example, in mammals the infection is combated by apoptosis, clearance of infected cells by natural killer cells and humoral immunity provided by the antibodies (Altfeld et al., 2011) and also by RNA interference (RNAi) (Schnettler et al., 2009). In plants and insects, infection by RNA viruses is mainly controlled by RNAi (Llave, 2010; van Rij and Berezikov, 2009). Plants also employ programmed cell death through the hypersensitive response (Zhang et al., 2010), whereas insects use intracellular protein-mediated antiviral defense (Carre-Mlouka et al., 2007; Wyers et al., 1995). Since long the innate immune systems were known to operate in prokaryotes and the adaptive immunity was not reported until the discovery of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) along with CRISPR associated (Cas) proteins, which form the CRISPR-Cas system (Barrangou et al., 2007). Thus, the existence of the adaptive immunity in prokaryotes was reported for the first time in 2007 (Barrangou et al., 2007). The prokaryote also seems to use a multi-layered antiviral defense, which acts at various stages of virus life cycle (Figure 1.1). The first line of defense can be the obstruction of phage adsorption by reducing the interaction between phage and bacterium. Next, the viral DNA entry can be prevented into the host by superinfection exclusion (Sie), but in case of successful entry, the restriction-modification enzymes (R-M) and R-M like systems can cleave the viral DNA or abortive infection systems (Abi) or CRISPR-Cas mediated acquired immune system can come into play (Hyman and Abedon, 2010; Labrie et al., 2010; Westra et al., 2012b). Thus, the invasion is checked at different stages of viral life cycle by various types of innate defense system and the adaptive immune system.

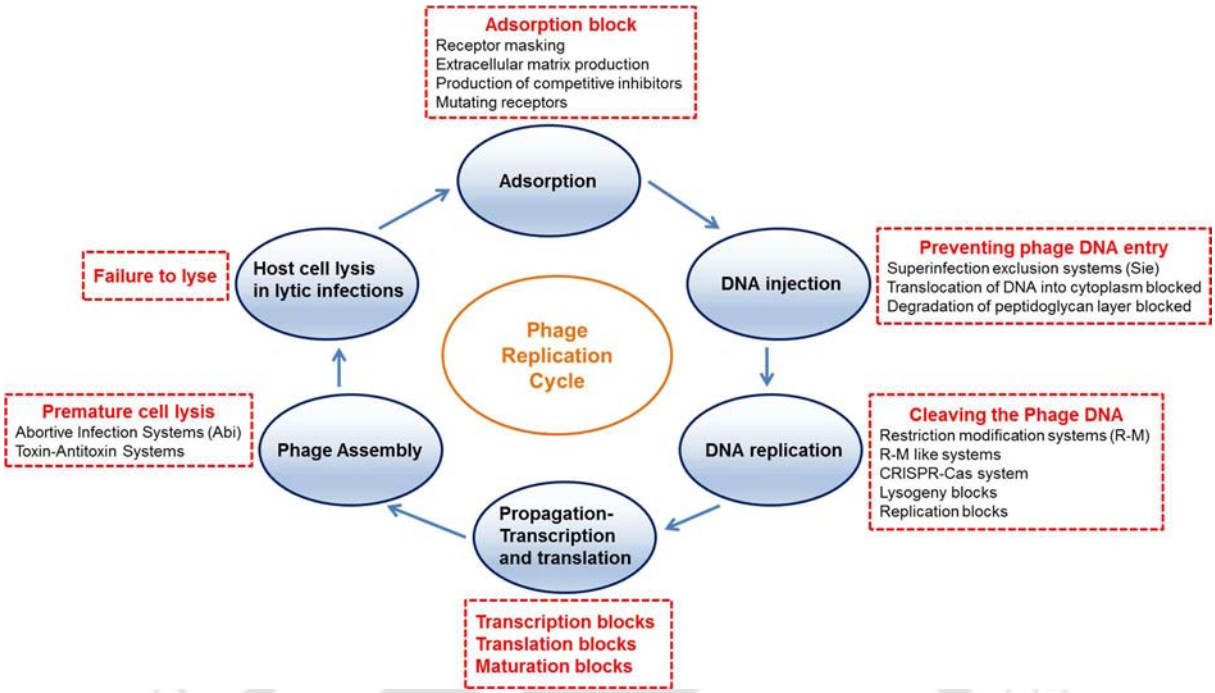


Figure 1.1 ***The various antiviral defense systems of bacteria to counter the different stages of viral life cycle.*** The phage replication cycle in the host cell is shown in the ovals. The antiviral defense operating at particular stages is shown in rectangle corresponding to that stage.

1.1.1. Innate immune systems

The ability of an organism to defend itself against invasion depends on its ability to mount immune responses. All organisms have inborn defense mechanisms that constitutes the innate immunity. Prokaryotes have a number of innate defense systems, as discussed below.

1.1.1.1. Surface exclusions

Surface exclusions or adsorption resistance mechanisms are simple and effective passive defense strategies for acquiring phage resistance, which mainly rely on preventing the attachment of a phage or virus to appropriate receptors on the prokaryotic cell surface. A wide range of surface exposed molecules act as virus receptors, including proteins,

lipopolysaccharides, teichoic acids and capsules. Prokaryotes can alter these receptors either by loss, downregulation, mutation, structural modification (changing the shape of the receptor) or receptor masking through the expression of extracellular polymers like polysaccharides on the cell surface (Hyman and Abedon, 2010), which hinder the recognition by phages, but in some cases, can also make them more susceptible to infection by other viruses (Avrani et al., 2011). Phages counter this resistance by mutations leading to recognition of the altered receptor or another receptor. For example, the downregulation of the receptor (LamB) in *Escherichia coli* to avoid the λ phage is countered by phage by evolving its receptor binding protein (J) through several point mutations to bind a new receptor (OmpF) (Meyer et al., 2012). In *Aeromonas salmonicida* (Ishiguro et al., 1981), *Staphylococcus aureus* (Nordstrom and Forsgren, 1974; Nordstrom et al., 1974), *E. coli* (Nordstrom et al., 1974) and *Lactococcus* spp. (Forde and Fitzgerald, 1999) the receptor masking by extracellular polymers is countered by phages by producing enzymes, which degrade certain polymers (Scholl and Merril, 2005; Sutherland et al., 2004). The composition and efficiency of these polymers, as well as their mode of production, vary widely among bacteria.

1.1.1.2. Blocking DNA injection

The second line of defense operates at the post adsorption level and prevents the genetic material of invaders from entering the host cytoplasm, usually by blocking DNA injection by proteins located in association with or in close proximity of the cell membrane or cell wall and these proteins can be encoded by a plasmid (Garvey et al., 1996) or a prophage (Forde and Fitzgerald, 1999; Lu and Henning, 1994; McGrath et al., 2002; Watanabe et al., 1984). Prophage encoded systems are called superinfection exclusion mechanisms (Sie) and

are mostly common in gram-negative bacteria (Mahony et al., 2008) like in *E. coli* and their T-even phages (Dulbecco, 1952; French et al., 1951) but are also found in some gram-positive bacteria like lactic acid bacteria and their phages (Forde and Fitzgerald, 1999). The bacteria can multiply for many generations carrying the prophage and only at some stage, upon specific stimuli or stress conditions some prophages excise, multiply, and proliferate and kill the host cell. Otherwise, the system provides a selective advantage to the part of the bacterial population carrying the prophage (Brown et al., 2006), by using prophage as a weapon to infect and kill competing (prophage-free) bacteria, whereas the prophage carrying fraction of the population remains intact because of Sie. Under these conditions, the host bacterium and Sie encoding prophage have a mutualistic relationship (Roossinck, 2011).

1.1.1.3. Restriction-modification system

Restriction-modification systems (R-M) are intracellular defense systems, which neutralize the invader by acting directly on its DNA and operates when the adsorption and DNA injection is not prevented. While it restricts or cleaves specific patterns in the incoming foreign DNA, it ignores the same pattern in the host DNA from cleavage by unique biochemical modification. R-M systems are encoded by approximately 90% of sequenced bacterial and archaeal genomes and provide resistance from a wide variety of extrachromosomal elements like phages and plasmids (Hyman and Abedon, 2010; Tock and Dryden, 2005). A typical R-M system consists of a DNA methyltransferase, which modifies specific bases in the host's genomic (endogenous) DNA and a restriction endonuclease, which cleaves the same sequences when unmodified (exogenic DNA). The recognition sequence of these restriction endonucleases is usually 4-8 bp in length. The modifications of endogenous DNA take place after replication of the prokaryotic genome and typically consist

of methylation of adenine or cytosine bases. They can occur either within or at locations up to 1000 bp from the recognition site and requires ATP-dependent translocation of the R-M complex on the DNA. To evade the R-M systems, phages and conjugational plasmids have evolved a variety of mechanisms like stimulation of the host methyltransferase or acquisition of multispecific methyltransferase to methylate the genome in the same pattern as the host in order to avoid recognition (Kruger and Bickle, 1983). Alternatively they also code for inhibitor proteins to obstruct restriction endonucleases like the Ocr protein of phage T7 which blocks the active site of some restriction endonucleases by mimicking 24 bp of bent B-form DNA (Bandyopadhyay et al., 1985) or by carrying enzymes that will hydrolyse or degrade R-E cofactors like S-adenosylmethionine (AdoMet). Phages may also incorporate unusual bases in their genomes to avoid restriction endonucleases recognition like *Bacillus subtilis* phages replace thymine with 5-hydroxymethyluracil (Kruger and Bickle, 1983) and also code for a protein that further inhibits the host protein uracil-DNA glycosalyse from cleaving uracil bases from the phage DNA (Anders et al., 2014). Interestingly, this inhibitory protein is also a DNA mimic (Putnam and Tainer, 2005). The T-even phages T2, T4, and T6 also contain unusual bases in their genomes and may further post-synthetically glycosylate their DNA to avoid restriction by endoRNases (Kruger and Bickle, 1983). Another escape mechanism involves the loss or reorientation or alteration of the restriction recognition sites. For example, phages tend to avoid palindrome sequences in their genome, since type II restriction endonucleases often recognize symmetrical (palindromic) sequences (Rocha et al., 2001). Further, phages may also encode DNA binding proteins which protect the viral DNA after injection.

The R-M systems can be classified into four main groups (I-IV) depending on the subunit combination, characteristics of the recognition, cleavage site and cofactor

requirements (Roberts et al., 2003; Tock and Dryden, 2005). Each type of R-M systems includes restriction enzymes that recognize different recognition sequences.

I) Type I R-M systems

Type I R-M systems form a multiprotein complex of 400-500 kDa, comprising of a nuclease (hsdR), a DNA methyltransferase (hsdM) and a recognition sequence binding specificity subunit (hsdS) with an $hsdR_2$:$hsdM_2$:$hsdS_1$ stoichiometry. It requires AdoMet for DNA methylation and Mg^{2+} and ATP for endonuclease activity. The restriction sites are asymmetric and bipartite and separated into two fragments of 3-4 bp and 4-5 bp interspaced by 6-8 bp of nonspecific sequence (Murray, 2000). The methyltransferase activity of the hsdM subunits is triggered by hemimethylated sequences (Vovis et al., 1974) while the unmethylated sequences induce ATP-dependent DNA translocation by the hsdR subunits and the hsdS subunit remains bound to the restriction site resulting in loop formation in the DNA. R-M complex collision with a second R-M complex or collision block results in subsequent cleavage at sites remote from the restriction site by the hsdR subunits (Murray, 2000; Studier and Bandyopadhyay, 1988).

II) Type II R-M systems

Type II R-M systems have revolutionized molecular cloning and are extensively used in recombinant DNA techniques. They are ATP-independent and cleave DNA at a well-defined position, most often within or very near to the restriction site, which is in contrast to type I and type III R-M systems (Pingoud et al., 2005). Restriction sites are usually 4-8 bp palindromic sequences. It requires Mg^{2+} for endonuclease activity and requires restriction

endonucleases to form homodimers or homotetramers to generate symmetrical cleavage products with either blunt or sticky ends. The DNA sequences are protected by modification that are introduced by a separately encoded DNA methyltransferase. In contrast to the other types, type II methyltransferases and restriction endonucleases work independent of each other (Pingoud et al., 2005), therewith sharing certain properties with selfish addiction modules such as toxin-antitoxin systems (Kobayashi, 2001; Naito et al., 1995). There are 4108 type II enzymes out of 4263 known restriction enzymes, which are listed in REBASE (2015) (Roberts et al., 2015). Type II restriction endonuclease (REase) sequences have frequently been found to be ORFan sequences with no significant similarity to any other protein (Kroger et al., 1984; Mullings et al., 1988), which was initially thought to be a convergent evolution (Wilson and Murray, 1991) but later the structural similarity among type II REases shown by crystallographic studies revealed the possibility of divergence from a common ancestor with a rapid evolutionary rate (Venclovas et al., 1994).

III) Type III R-M systems

Type III R-M systems form a multiprotein complex comprising of DNA methyltransferase (Mod) and a restriction endonuclease (Res) in Mod_2:Res_2 stoichiometry. It requires AdoMet for DNA methylation and Mg^{2+} and ATP for endonuclease activity and two inversely oriented unmethylated nonpalindromic recognition sequences of 5-6 bp. Type III endonucleases cleave 25-27 bp downstream of the recognition site. The methyltransferase provides protection by catalyzing methylation of one strand of the recognition sequence (Bachi et al., 1979; Boyer, 1971; Iida et al., 1983), though during DNA replication some sites become unmodified but remain protected, as these unmodified sites are never inversely oriented (Meisel et al., 1992). The expression of mod genes is phase variable, which seems to

be a way of downregulating host defense to allow uptake of potentially beneficial DNA (Ando et al., 2000; Donahue et al., 2000). Also, these phase-variable type III R-M systems have evolved into regulatory systems that control expression of target genes through their methylation status (Fox et al., 2007; Srikhanta et al., 2010; Tan et al., 2016).

IV) Type IV R-M systems

Type IV R-M systems usually do not encode a modification enzyme and thus have no methyltransferase activity but instead recognize and cleave modified or methylated DNA substrates (Bair and Black, 2007; Roberts et al., 2003). Example, Mrr is a type IV restriction endonuclease which restricts both adenine and cytosine methylated foreign DNA (Waite-Rees et al., 1991). In contrast to all known type IV R-M systems, the BspLU11III present in *Bacillus* sp. LU11 possess two methyltransferases with adenine and cytosine specificity and one endonuclease that also shows adenine-specific methyl transferase activity and is able to protect the DNA against cleavage by itself (Lepikhov et al., 2001). The type IV R-M systems seem to have evolved as a counter defense against escape mechanisms of phage.

R-M-like systems

R-M-like system has been found in *Streptomyces coelicolor* known as phage growth limitation (Pgl) system and is made up of four genes, one of which resembles a methyltransferase (Hoskisson and Smith, 2007). When phages infect host cells having Pgl system, the Pgl modifies the offspring phage DNA, such that they are targeted for destruction (Sumby and Smith, 2002).

1.1.1.4. Abortive infection system

This is the last line of defense and operates in the cases of a successful infection and inhibits phage development at various stages of the phage life cycle inside the host cell such as replication of the phage genome, transcription, protein production and virus assembly (Chopin et al., 2005). The result is the complete inhibition of virus proliferation and death of the host cell. There are number of systems that lead to abortive infection (Abi). The term Abi is used for the controlled self-destruction or suicide of the infected cell in order to completely prevent the release and spread of the new virus particles and is therefore advantageous to the surrounding bacterial population as a whole, since the infection is confined to the sacrificed cell (Chopin et al., 2005; Hyman and Abedon, 2010). The Abi systems vary in their specific molecular mechanisms, with each mechanism targeting only certain groups of phages and have little or no evolutionary relationship. The majority of them are encoded on plasmids consisting of a single gene. In general, the toxic effect of an Abi system is phage induced and results in cleavage of essential cellular components, therefore a tight regulation of these Abi systems is essential. The Abi systems represent a form of bacterial altruism and is common among Gamma-proteobacteria, Actinobacteria, Cyanobacteria and Firmicutes (Chopin et al., 2005) and were first isolated from *lactococci*. Lactic acid bacteria contain many Abi systems, with *Lactococcus lactis* alone containing 20 Abi systems (AbiA to AbiU). Some of them are tightly regulated like AbiD1, which inhibits the resolution of branched DNA structures while others are constitutively expressed at low levels like AbiA, AbiB, AbiK (Chopin et al., 2005). In *E. coli* several Abi systems have been described, which either consist of a single protein like LitA, PrrC, PifA or a two-component system as in the case of Rex (Labrie et al., 2010; Molineux, 1991).

The prokaryotic defense systems like Abi systems, R-M systems and toxin-antitoxin (TA) systems have some similarities. TA systems which are composed of a stable toxin and an unstable antitoxin mediate abortive infection upon phage infection. Normally, the antitoxin remains bound and inhibits the toxin but a decrease in the levels of the unstable antitoxin activates the toxin leading to growth arrest or cell death as in the case of *mazEF* and *hok-sok* TA modules (Hazan and Engelberg-Kulka, 2004; Pecota and Wood, 1996). Interestingly, the AbiQ system is found to function as a protein-RNA TA pair (Fineran et al., 2009). Thus, some TA systems can be considered to be a subtype of Abi systems. Interestingly, the loss of the methyltransferase in case of R-M systems results in a toxic effect of the restriction endonuclease, thus R-M systems can be regarded as TA systems (Kobayashi, 2001). The TA systems have various functions including the maintenance of mobile genetic elements, regulation of pathogenicity islands and induction of stasis or cell death as a stress response (Blower et al., 2011b; Van Melderen, 2010). TA cassettes consist of a bicistronic operon, *i.e.*, a single promoter controlling the expression of a gene pair, encoding for the unstable antitoxin and the stable toxin. Both elements can be proteins or RNA molecules according to the system classification (Bukowski et al., 2011). Formation of a toxin-antitoxin complex inhibits toxin activity, which can manifest in a variety of mechanisms, but is either lethal or restrict cellular growth (Blower et al., 2011b; Van Melderen, 2010; Van Melderen and Saavedra De Bast, 2009). TA systems are widespread in bacteria as well as in archaea, although the three systems – type I-III, that mediate Abi have been characterized from bacteria.

I) Type I TA systems

In type I TA systems the toxin is a protein and the antitoxin is an antisense RNA. The silencing occurs at the RNA level and relies on the complementary base-pairing between antitoxin RNA and the toxin's mRNA. Thus, translation of the mRNA gets inhibited either by degradation via RNase III or by occluding the Shine-Dalgarno sequence or ribosome binding site. The toxin and antitoxin are usually encoded on opposite strands of DNA. The complementary base-pairing between the genes usually occurs in the 5' or 3' overlapping region and has 19-23 contiguous base pairs (Fozo et al., 2008a). Toxins of type I systems are small, hydrophobic proteins and cause toxicity by damaging cell membranes, like the *hok-sok* encoded in plasmid R1 of gram-negative bacteria when stimulated by phage T4 causes cell membrane damage by toxin Hok (Fozo et al., 2008a).

II) Type II TA systems

In type II TA systems both the toxin and the antitoxin are proteins that form a tight complex. This prevents the action of the toxin. However, since the antitoxin is a labile protein, it can be degraded by cellular proteases to release the stable toxin whenever required. The largest family of type II toxin-antitoxin systems is vapBC (Robson et al., 2009). Type II TA systems are organised in operons with the antitoxin usually located upstream of the toxin and inhibits the toxin by downregulating its expression. The proteins are typically around 100 amino acids in length (Fozo et al., 2008b) and exhibit toxicity in a number of ways: CcdB protein affects DNA gyrase by poisoning DNA topoisomerase II (Bernard and Couturier, 1992), whereas MazF protein of *mazEF E.coli* TA cassette is a toxic endoribonuclease that

cleaves cellular mRNAs at specific sequence motifs and aborts phage P1 infection by inhibiting translation (Hazan and Engelberg-Kulka, 2004).

III) Type TA III systems

Type III TA systems have the antitoxin as RNA (ToxI) transcribed by an array of repeat sequences. The antitoxin RNA interacts and inhibits the protein toxin (ToxN), like the plasmid-encoded ToxIN system of gram-negative bacteria (Fineran et al., 2009). The transcriptional terminator stem-loop structure is formed between the two, which mediates the regulation. The phage infection can trigger the disruption of ToxIN transcription or affect the stability of the ToxIN complex, causing mobilization of the toxin, which allows it to exert its toxicity (Blower et al., 2011a).

Phages have also developed counter strategies to overcome the damage caused by various Abi or TA systems and avoid abortion of infection either by recombination with a cognate prophage or by point mutations within the phage genome or by encoding an antitoxin molecule (Chopin et al., 2005).

1.1.2. Adaptive immune system

The adaptive immune system has the hallmark of specificity (restricting the response to a particular antigen and its discrimination from others) and memory (once an immune response is mounted against the antigen, the encounter is remembered and a more rapid and vigorous response is prepared for subsequent encounters). The existence of adaptive immunity in prokaryotes came into picture with the discovery of CRISPR-Cas system.

1.1.2.1. CRISPR-Cas system

CRISPR-Cas system which is the latest addition to the list of known prokaryotic defense systems is unique because of its adaptable and heritable nature. Interestingly, it also holds functional analogy to the eukaryotic RNAi systems (Al-Attar et al., 2011; Bhaya et al., 2011; Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010; Terns and Terns, 2011; van der Oost et al., 2009). This immune system of prokaryotes operates after the invader's nucleic acid has been injected into the host cytoplasm. It utilizes exposure to foreign nucleic acids to subsequently target and destroy the incoming nucleic acid of similar viruses or plasmids. For this, small fragments of foreign nucleic acids are incorporated into the host genome between conserved short DNA repeats known as CRISPR. The CRISPR array is transcribed and processed into small RNAs, which in conjunction with Cas proteins recognize and destroy non-self nucleic acid. Thus, this system adapts or acquires the immunity, which is then passed on to the progeny and therefore it is called an adaptable and heritable immune system. Approximately half of all sequenced bacteria and nearly all sequenced archaea harbour one or more CRISPR loci (Grissa et al., 2007a; Rousseau et al., 2009), which is composed of a series of direct repeats separated by unique spacer sequences (Al-Attar et al., 2011) along with adjacent *cas* genes (Haft et al., 2005; Jansen et al., 2002a). The CRISPR-Cas system show high diversity owing to the dynamic evolution of CRISPR-cas loci, involving numerous rearrangements of the locus architecture and horizontal transfer of complete loci or individual modules, making straight forward phylogenetic classification complicated. The updated CRISPR-Cas system classification combines the analysis of signature protein families, features of the architecture of *cas* loci and the information of the effector modules involved in interference and categorize them into two distinct classes, five types (I-V) and sixteen subtypes (Makarova et

al., 2015). These highly diverse CRISPR-Cas types display major structural and functional differences in their mode of generating resistance against invading nucleic acids. The CRISPR-Cas defense process can be operationally distinguished into three phases – adaptation (spacer acquisition), expression and maturation of CRISPR RNA (crRNA) and CRISPR interference (target degradation), which are temporally separated with each step requiring one or more Cas proteins. The phages have also evolved resistance to CRISPR-Cas interference. Phages with mutated, recombined or even lost protospacer (viral sequence that corresponds to spacer) target sequence become resistant to CRISPR-Cas system (Andersson and Banfield, 2008; Heidelberg et al., 2009). Further, they may also possess mechanisms that can directly target the CRISPR-Cas machinery (Samson et al., 2013) like possession of phage-borne anti-CRISPR genes that encode protein inhibitors that inactivate the CRISPR immunity (Maxwell, 2016; Pawluk et al., 2014). Another mechanism to bypass CRISPR-Cas system is observed in *Vibrio cholerae* phages that encodes their own CRISPR-Cas system to target bacterial antiphage CRISPR-Cas system (Seed et al., 2013).

1.1.3. Common characteristics of the prokaryotic defense systems

All the prokaryotic defense systems share three common characteristics:

1. Propagation mostly through horizontal gene transfer (HGT).
2. Extremely high rates of molecular evolution which might be a consequence of the co-evolutionary arms race with phages (Hoskisson and Smith, 2007).
3. Traits of selfish genetic elements.

HGT takes place by evading the pre-existing resistance mechanisms of a given organism, which is possible because of the rapid evolutionary rates that enable the invading

elements to continuously develop anti-resistance strategies. These defense systems may also undergo HGT between distantly related prokaryotes (Godde and Bickerton, 2006; Gogarten and Townsend, 2005; Haaber et al., 2009; Jeltsch and Pingoud, 1996; Nelson et al., 1999), which is evidenced by the fact that a large number of these systems are encoded by plasmids, phages, prophages or hypervariable loci in the prokaryotic chromosome or genomic islands of foreign origin or are linked to transposase genes and by homology found between distantly related strains. They often exhibit a codon usage bias and GC content different from the rest of the genome (Chopin et al., 2005; Godde and Bickerton, 2006; Gogarten et al., 2002; Kobayashi, 2001). The mobility allows rapid acquisition and dissemination of new defense systems to counteract the invading phage, thus contributing to the extensive co-evolution with the prokaryotic invaders (Hoskisson and Smith, 2007; Stern and Sorek, 2011). Thus, the high evolutionary rate of gene sequences results in high genetic variability (Zheng et al., 2004), which is manifested in the enormous numbers of classes and types of R-M systems and even hyper variability of the target recognition domains in R-M systems (Orlowski and Bujnicki, 2008; Tock and Dryden, 2005), in the variety of abortive infection systems (Chopin et al., 2005) and in various CRISPR-Cas subtypes (Haft et al., 2005; Kunin et al., 2007; Makarova et al., 2011b). The selfish behaviour shown by the prokaryotic defense mechanisms leads to an increase in their relative frequency within a given population. Selfishness is characterized by the following features:

1. Loss of these systems has deleterious effects on the host cell.
2. Preventing the competing genetic element from establishing itself in the population by competitive exclusion between the two, which will lead either to the destruction of the invader or to host death.
3. Extensive mobility between genomes and are often associated with plasmids, phages, transposons and integrons.

The post segregational killing of the carrier that loses the element to establish its maintenance in the population and competitive exclusion between two equally deleterious selfish elements has been observed in the R-M systems (Kobayashi, 2001) and TA modules (Makarova et al., 2009). Thus, the cellular defense incurred by these systems in host is a mere by-product, which primarily seems to be a self-maintenance strategy that happens to be advantageous to the host cell (Kobayashi, 2001; Makarova et al., 2009; Stern and Sorek, 2011).

The prokaryotes evade viruses by various defense mechanisms, while the viruses develop counter strategies to escape the defense system or adapt to new hosts (Labrie et al., 2010; Samson et al., 2013). There is a continuous evolutionary arm race. This remarkable ability to adapt is manifested in several dynamic forms (Figure 1.2). The counter strategies of phages for various prokaryotic defense systems include:

1. Surface exclusion barrier – adaptability to new receptors, digging for receptors, stochastic recognition of variable host receptors.
2. Restriction-modification (R-M) systems barrier – fewer restriction sites or sites in a non-recognizable orientation, modification of restriction sites, masking restriction sites by mimicry of the substrate DNA, coding proteins to target and shut down the restriction endonucleases, possessing unusual bases, degradation of R-M cofactor, stimulation of the modification enzyme or even acquisition of modification enzyme.
3. Abortive-infection (Abi) systems barrier – mutations in specific phage genes, encoding an antitoxin molecule.
4. CRISPR-Cas systems barrier – mutation in the protospacer or in the protospacer adjacent motif (PAM), anti-CRISPR proteins, antibacterial CRISPR-Cas systems.

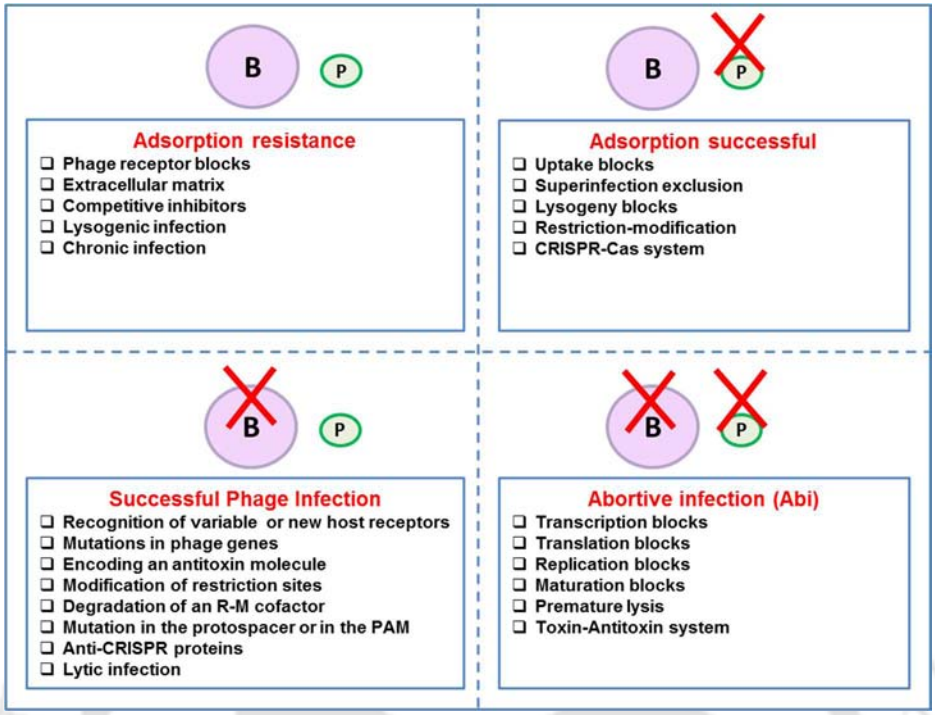


Figure 1.2 ***Types of phage adaptation to the various bacterial defense mechanisms and the relative outcome on their survival.*** The bacterial resistance mechanism operating at different levels are shown along with the adaptations by phage. The bacterium (B) and the phage (P) are represented with circle. The deleterious effect on bacterium/phage is shown by red cross, while the absence of it indicates the survival.

1.1.4. Alternative functions of defense systems

Most of the prokaryotic defense mechanisms show exaptation and evolve to gain a different function in cellular regulation, independent of the defense system. The extreme plasticity and dynamic nature of the prokaryotic genome to adapt can be attributed to exaptation which refers to the evolutionary process where a biological structure or function is used for a purpose other than that for which it was initially evolved (Brosius and Gould, 1992). In the case of R-M systems, owing to the loss of restriction endoRNase, the orphan methyltransferse adopts regulatory role in DNA metabolism and in differential epigenetic modifications of the genome, enhancing pathogenicity (Srikhanta et al., 2010; Stern and

Sorek, 2011). For example, in *E. coli* the methylation by Dam methylase is involved in regulatory processes such as mismatch repair by the MutHLS complex, binding of the replication initiation complex to methylated OriC and regulation of bacterial pathogenicity (Marinus and Casadesus, 2009). In *Caulobacter crescentus*, the CcrM methylase is shown to affect the cell cycle (Marinus and Casadesus, 2009). The phase variable type III R-M systems have numerous regulatory roles including the regulated removal of the barrier against extraneous DNA, which results in uptake of potentially beneficial DNA (Ando et al., 2000; Donahue et al., 2000), autolytic self DNA degradation or cellular suicide (Dybvig et al., 1998; Hamilton and Dillard, 2006; Saunders et al., 1998) and epigenetic gene regulation via differential methylation of the genome which switches different genes on and off (Seib et al., 2002) and association with pathogenicity of bacterial species by allowing colonization, immune evasion and adaptation to novel environments (Srikhanta et al., 2010).

Abortive infection systems and TA systems which get activated in response to various types of cellular stress are also involved in other cellular functions and maintenance of genomic integrity by preventing loss of certain mobile genetic elements (Chopin et al., 2005; Van Melderen, 2010). For example, *mazEF* TA loci, aborts translation by cleaving mRNA molecules in response to different stress signals (Christensen et al., 2003; Pedersen et al., 2002) like phage infection (Hazan and Engelberg-Kulka, 2004) and can have reversible or irreversible effects. Reversible effects include bacteriostatic effects, which allow reduced growth rate of each cell during nutritional stress (Christensen et al., 2003; Pedersen et al., 2002), while the irreversible effects of *mazEF* include programmed cell death that occurs in a subpopulation of cells, permitting the survival of the population as a whole (Engelberg-Kulka et al., 2006). An interesting case is observed in *Myxococcus xanthus* where the toxin MazF, which mediates programmed cell death during multicellular development, has adopted a key

transcriptional regulator as an alternative antitoxin instead of its antitoxin MazE (Nariya and Inouye, 2008).

CRISPR-Cas system also has additional roles in host regulatory and developmental processes. Example, the inhibition of swarming and biofilm formation by the lysogenic infection of *Pseudomonas aeruginosa* with bacteriophage DMS3 requires the CRISPR region in the host. Mutation or deletion of five of the six cas genes and one of the two CRISPRs in this region restored biofilm formation and swarming to DMS3 lysogenized strains. Thus, the type I-F CRISPR-Cas system seems to play a role in modifying the effects of lysogeny on *P. aeruginosa* [133]. Later, the involvement of Cas3 with functional HD and DEXD/H domains was also shown to suppress biofilm formation in DMS3 lysogens [132]. The regulatory role of type III-B systems were also found, which seem to exert the effect by cleaving the complementary mRNAs (Hale et al., 2012). The involvement of Cas1 in DNA repair has been indicated by interaction of Cas1 from *E. coli* with the essential DNA repair enzymes such as RecB, RecC and RuvB (Babu et al., 2011), which is in coherence with the preliminary hypothesis of CRISPR-Cas being a DNA repair system (Makarova et al., 2002). Cas1 deletion strains show defects in chromosome segregation and DNA repair, which suggests Cas1 interactions with DNA repair components. Also, the CRISPR involvement in replicon partitioning has been predicted (Mojica et al., 1995). Further, in *Pyrococcus furiosus* the upregulation of *cas* genes has been observed in response to gamma irradiation suggesting a role in DNA repair (Williams et al., 2007). Later, it was revealed that CRISPR-Cas system have a dual function – role in bacterial antiviral immunity and DNA repair (Babu et al., 2011). Another function of CRISPR-Cas system is shown by the linkage of regulation of *cas* operon to dev operon, a cluster involved in fruiting body development in *Myxococcus xanthus* (Viswanathan et al., 2007). Thus, these defense systems are also involved in other regulatory processes as well.

1.2. Salient features of adaptable and heritable CRISPR-Cas immune system

The CRISPR-Cas defense comprises of a novel system having the ability to incorporate short sequences of non-self genetic material known as spacers, at specific locations within CRISPRs in the host genome. Subsequently, these are transcribed and processed into small noncoding RNAs called the crRNA, which in conjunction with specific Cas protein complexes form a surveillance complex of this adaptive immune system. This nucleoprotein complex recognizes and binds to the incoming foreign genetic material that is complementary to the spacer region of the crRNA and triggers the destruction or silencing of the invading DNA (Jiang and Doudna, 2015). Since this process exploits the previous exposure to a virus or plasmid to target recently detected foreign DNA or RNA it is termed as adaptive/acquired immune system (Figure 1.3).

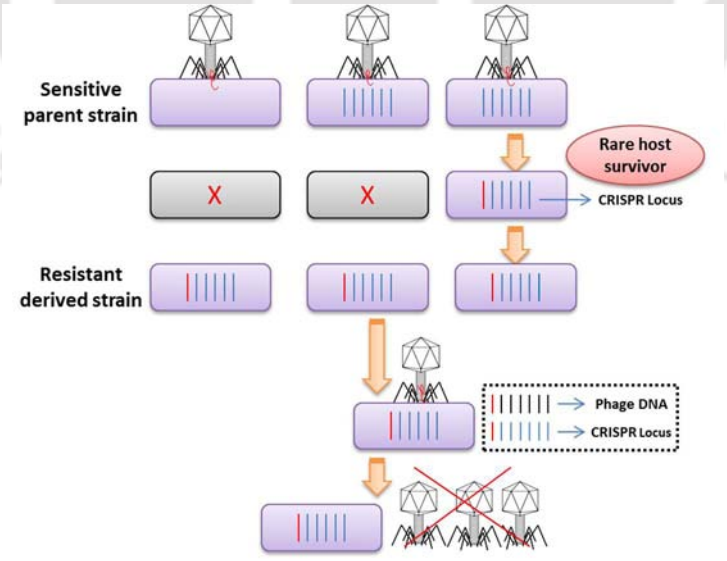


Figure 1.3 ***The overview of CRISPR-Cas system.*** The defense provided by CRISPR-Cas system is shown schematically. The host cells are shown in purple rectangles. When the host cells encounter the phage infection for the first time, most of them die, but the rare host survivors are able to incorporate a portion of invaders genome (shown in red) in the CRIPSR locus, which is passed on to the progeny. When the same infection is encountered again, the host efficiently degrades the invader nucleic acid and is therefore resistant to the infection.

1.2.1. CRISPR-Cas discovery

One of the possible reasons for the late discovery of the CRISPR-Cas immune system can be the tight regulation of CRISPR-Cas systems in model organisms, such as *E. coli* and *Salmonella enterica*, wherein their expression is silenced under normal laboratory growth conditions. The existence of unusual structure of repetitive DNA adjacent to the isozyme-converting alkaline phosphatase (iap) gene in *E. coli* K12, which comprises of a cluster of 29 nt invariant direct repeats interspaced with 32 nt variable spacing sequences, was discovered in 1987 by Ishino *et al.* (Ishino et al., 1987) and in 1989 by Nakata *et al.* (Nakata et al., 1989). This family was first identified as a distinct class of interspaced short sequence repeats (SSR). Other bacteria and archaea such as *Mycobacterium tuberculosis* (Groenen et al., 1993; Hermans et al., 1991; van Embden et al., 2000), *Haloferax mediterranei* (Mojica et al., 1995), *Streptococcus pyogenes* (Hoe et al., 1999), Anabaena (Masepohl et al., 1996) and *Thermotoga maritima* (Nelson et al., 1999) also showed the presence of similar repeat clusters, which were recognized as a defined prokaryotic family of short regularly spaced repeats (SRSR) (Mojica et al., 2000) or SPacers Interspaced Direct Repeats (SPIDR) (Jansen et al., 2002b), but their function remained unknown. In 1993, the first typing method based on the repetitive elements in *M. tuberculosis* was developed (Groenen et al., 1993). The term CRISPR was coined for these repeats in 2002 along with discovery of associated *cas* genes, which encode proteins such as nucleases, helicases and polymerases (Jansen et al., 2002a). The turning point came in 2005, when three groups independently reported that the hypervariable spacers show sequence homology to viruses or bacteriophages or plasmids and hypothesized that CRISPRs and associated proteins could play a role in immunity against transmissible genetic elements (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Bioinformatics analyses predicted their involvement in chromosome partitioning (Mojica et

al., 1995) and DNA repair (Makarova et al., 2002) and conjecture was put forward by Makarova *et al.* that CRISPR-Cas system might be a defense system akin to eukaryotic RNAi (Makarova et al., 2006). The first experimental evidence that this system indeed confers resistance to phage infection came in 2007 (Barrangou et al., 2007), which was followed by the demonstration of CRISPRs ability to prevent plasmid transfer in 2008 (Marraffini and Sontheimer, 2008). Soon, the ribonucleoprotein complex called Cascade (CRISPR associated antiviral defense complex) was identified to be responsible for the maturation of CRISPR RNA (Brouns et al., 2008). Further experiments showed that the CRISPR locus is transcribed as a single RNA transcript, which is processed by Cas proteins to release smaller crRNA units, each including one targeting spacer, which are used as interfering units with the incoming foreign genetic material – either DNA (Brouns et al., 2008) or RNA (Hale et al., 2009) – by complementary base-pairing. This gave impetus to the research in the area of crRNA maturation and led to the characterization of crRNA processing pathway (Carte et al., 2008). Soon, the term protospacer was coined for sequence on viral genome that finally gets incorporated as spacer in the host genome (Deveau et al., 2008). This was followed by the recognition of the regions adjacent to protospacer as the Protospacer Adjacent Motif (PAM), which is utilized by CRISPR machinery for spacer uptake as well as in recognition self from non-self (Mojica et al., 2009).

Thus, it became known that the adaptable and heritable nucleic-acid based immune system comprising of the CRISPR locus along with the *cas* genes, protects the bacteria and archaea from infection by foreign genetic elements (Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010; Sorek et al., 2008; van der Oost et al., 2009; Wiedenheft et al., 2012). In the year 2012, the potential of CRISPR-Cas system in genome engineering (Gasiunas et al., 2012; Jinek et al., 2012) was demonstrated and the following year of 2013 witnessed the

explosion in the application of CRISPR-Cas system (Bassett et al., 2013; Cho et al., 2013; Cong et al., 2013; Friedland et al., 2013; Jinek et al., 2013). Since then, CRISPR-Cas systems have been efficiently repurposed for genome engineering of bacteria, archaea, insects, plants, mammals and even humans (Altenbuchner, 2016; Bassett and Liu, 2014; Calarco and Friedland, 2015; Doudna and Charpentier, 2014; Ebina et al., 2013; Kleinstiver et al., 2016; Kumar and Jain, 2015; Pohl et al., 2016; Shan et al., 2013; Trevino and Zhang, 2014).

1.2.2. Functional components of CRISPR-Cas system

The working machinery of the CRISPR-Cas system has two distinguishable components – the protein component (Cas proteins) and the nucleic acid component (CRISPR array) (Tsui and Li, 2015; van der Oost et al., 2014) (Figure 1.4). The CRISPR array is transcribed and processed to form a mature crRNA, which gets loaded onto the Cas proteins to forms a nucleoprotein surveillance complex of the system. The surveillance complex then scans the cell for the target having sequence complementarity with its loaded crRNA, a match ensues the target degradation (Al-Attar et al., 2011; Deveau et al., 2010; Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010; Sorek et al., 2008; van der Oost et al., 2009; Wiedenheft et al., 2012).

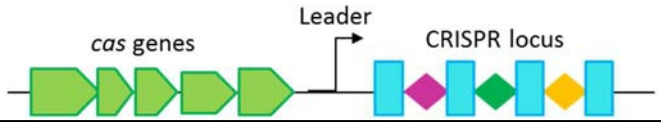


Figure 1.4 ***Representation of a CRISPR locus and the adjacent cas gene operon in a prokaryotic genome.*** The leader region is followed by invariant repeats shown in blue vertical rectangles and variable spacers are shown in diamonds (differently coloured to show the variability). The *cas* genes are shown in green pentagons.

1.2.2.1. Protein component

A number of protein encoding genes associated with CRISPR loci are designated as *cas*, with no homologues in eukaryotic or CRISPR-negative genomes (Haft et al., 2005; Jansen et al., 2002a; Makarova et al., 2011a). The order, orientation, and groupings of *cas* genes appear to be extremely variable. Species having multiple CRISPR loci with the same repeat sequence contain only one set of *cas* genes, but if multiple loci with varied repeat sequence are present, then a respective number of *cas* gene sets are found. In 2002, the neighbourhood of *cas* genes comprising of more than 20 different tandemly arranged genes with no preferential direction of their reading frames were initially identified and characterized by genomic context analysis. They were found on either side of a CRISPR locus. Based on computational analyses, Cas proteins were predicted to contain identifiable domains characteristic of helicases, nucleases, polymerases, and RNA-binding proteins, which led to the initial speculation that they may be part of DNA repair system (Makarova et al., 2002). Initially, four gene families *cas1-4* (Jansen et al., 2002a) were identified, which were later extended to include *cas5* and *cas6* (Bolotin et al., 2005; Haft et al., 2005) and 45 distinct Cas protein families were identified by Hidden Markov models (Haft et al., 2005). The categorization was refined taking into account genomic context information, resulting in 25 Cas protein families (Makarova et al., 2006) which were proposed to be involved in the generation, expansion, maintenance, transfer between genomes and function of the CRISPR elements. Based on the phylogeny of the highly conserved Cas1 protein and the organization of *cas* operon the CRISPR-Cas system was classified into eight subtypes, which were named after eight representative organisms that contained a single CRISPR-Cas locus. For example, Cas proteins from *E. coli* were designated as Cse (CRISPR system of *E. coli*) and *cse*1 represented gene1 of the *cas* operon. Similarly, Cas proteins of the subtypes were named as

Csa of *Aeropyrum pernix*, Csd of *Desulfovibrio vulgaris*, Csh of *Haloarcula marismortui*, Csm of *Mycobacterium tuberculosis*, Csn of *Neisseria meningitidis*, Cst of *Thermotoga neapolitana* and Csy of *Yersinia pestis.* But as the research progressed and more Cas proteins were identified, the existing CRISPR-Cas classification systems became inadequate to depict the emerging phylogenetic relationships between the distantly related Cas proteins, the extensive variability of *cas* operons and the organisms that contain multiple CRISPR loci. The elucidation of many Cas protein structures from different families and analysis of an increasing number of gene sequences led to the identification of homologous relationships, previously undetected, which enabled the unification of certain Cas families and the identification of novel ones (Makarova et al., 2011a). This resulted in the development of a polythetic classification of CRISPR-Cas systems, which included gene composition, operon organisation and the phylogenetic and functional relationships between *cas* genes (Makarova et al., 2011b). In this classification scheme, the CRISPR-Cas system was divided into two partially independent subsystems with the first containing the information processing module and requires the universally conserved core proteins, Cas1 and Cas2, which are involved in new spacer acquisition, while the second contained the executive subsystem, which is required for processing of primary CRISPR transcripts and recognition and degradation of invading foreign nucleic acid. The second module shows high diversity as in certain CRISPR sub-types, a single multifunctional protein is involved in the processing of the crRNA, while others have multisubunit complex formed by Cas proteins called Cascade. In addition, the pre-crRNA processing involves repeat-associated mysterious proteins (RAMPs), which constitute a large superfamily of Cas proteins and contain at least one ferredoxin-fold domain (also called as RRM, RNA recognition motif) (Brouns et al., 2008; Carte et al., 2008; Ebihara et al., 2006; Haurwitz et al., 2010). Based on this classification that integrates phylogeny, sequence, locus organization and content of *cas* genes, CRISPR-Cas systems were divided

into three types (I-III) and ten subtypes (Makarova et al., 2011b) (Figure 1.5). The type I system had six subtypes (type I-A to I-F) and both type II and III had two subtypes. The type-dependent proteins were typically involved in expression and/or interference and signature genes are involved in interference (Figure 1.5). The type I system is found in both bacteria and archaea, but the type II system is exclusively present in bacteria, whereas the type III systems appear more commonly in archaea, although it is also found in bacteria (Makarova et al., 2011b; Terns and Terns, 2011).

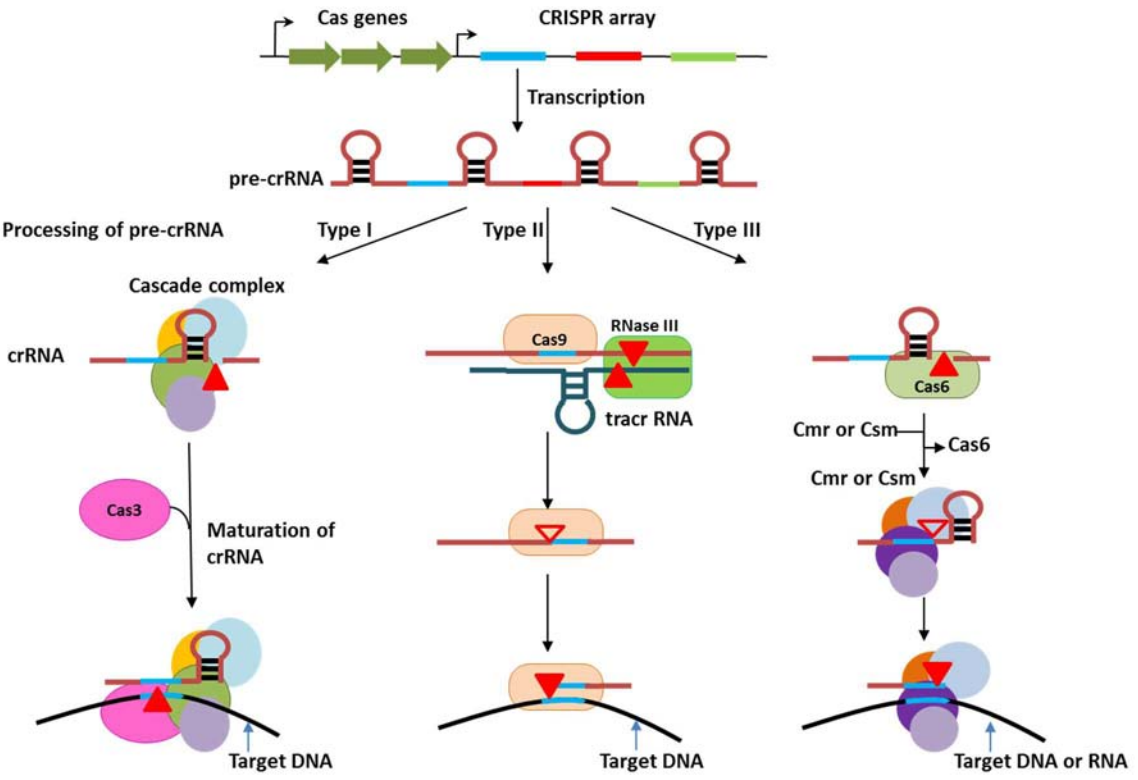


Figure 1.5 ***The type I, II, and III CRISPR-Cas systems and their mechanism of action.*** Transcription of the primary transcript (pre-crRNA) by RNA polymerase followed by processing by respective endoRNases in the three systems to produce of mature crRNAs is shown. In type I, the multisubunit Cascade binds pre-crRNA, which is cleaved by Cas6 or its homologs, to create crRNAs with a typical 8 nt extension at the 5'-end called the 5' handle, followed by the spacer and part of the repeat region, which can form a hairpin structure at the 3'-end called the 3' handle. In type II, a *trans*-encoded small RNA (tracrRNA) base-pairs with the repeat region and is cleaved by host RNase III in presence of Cas9. In type III, processing of crRNA requires Cas6, but the crRNAs appear to be transferred to a specific Cas complex (Cmr in subtype III-B and Csm in subtype III-A). In subtype III-B, the 3'-end of crRNA is further trimmed. The final stage is the interference which results in cleavage of targeted foreign nucleic acid and proceeds differently in all systems. In type I, crRNA bound Cascade recognizes the complementary target DNA (via PAM) and Cas3 cleaves the target DNA. In type II, Cas9 along with crRNA targets DNA for cleavage in a process that requires the PAM. Subtype III-A can target DNA (and RNA in some cases), whereas subtype III-B can target RNA. PAM does not appear to be required for the activity of type III systems. Filled triangles represent experimentally demonstrated nuclease activity while open triangles represent unidentified activity.

Later, with the information available on the final effector modules the CRISPR-Cas systems were broadly divided into two classes, *i.e.*, 'class 1' encompasses the systems that utilize a multisubunit complex (like Cascade, Csm or Cmr complexes) during interference and 'class 2' systems utilize a single effector protein (Cas9 or Cpf1). Therefore, it further led to the reclassification of the CRISPR-Cas system, which combined the information of the effector modules involved in interference stage with the updated analysis of signature protein

families and features of the architecture of *cas* loci. This new classification (Makarova et al., 2015) retains the overall structure of the previous version (Makarova et al., 2011b) but is expanded to encompass two distinct classes, five types (I-V) and sixteen subtypes (Figure 1.6).

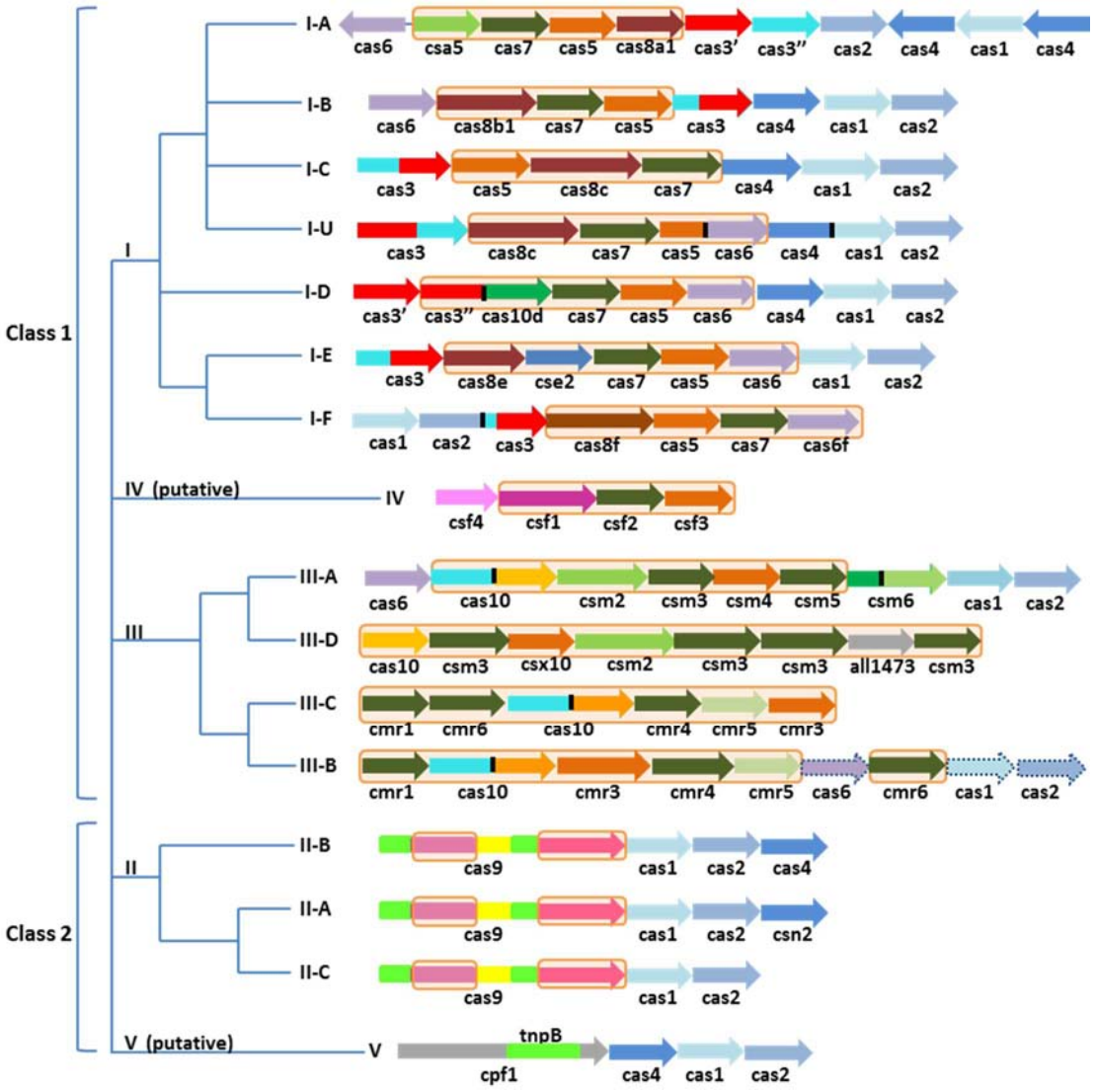


Figure 1.6 ***Architecture of the genomic loci for the subtypes of CRISPR-Cas systems.*** Typical operon organization is shown for each CRISPR-Cas subtypes, though the gene order may vary in each organism. Gene families are colour coded and the family name can be shown under each gene. The small subunit is encoded by either csm2, cmr5, cse2 or csa5. Genes or gene regions encoding components of the interference effector complex are boxed in orange. The cas1, cas2 and cas6 are dispensable in subtypes III-A and III-B (dashed lines). Gene regions coloured cyan represent the HD nuclease domain. The HD domain in Cas10 is distinct from that of Cas3 and Cas3''. The cas9 regions, RuvC-like nuclease (lime green), HNH nuclease (yellow), recognition lobe (purple) and protospacer adjacent motif (PAM)-interacting domains (pink) are shown. The functionally uncharacterized gene (all1473) in subtype III-D is shown in grey. The types and subtypes linkage is also shown.

Further, the availability of structural and functional information of core Cas proteins (Cas1-Cas10) has enabled the classification of Cas proteins into four distinct functional modules – adaptation or spacer acquisition, expression involving crRNA processing and target binding, interference or target degradation and ancillary or regulatory functions (Figure 1.7). The adaptation, expression/maturation and interference modules are discussed in details in section 1.2.4. The adaptation module is largely uniform across CRISPR-Cas systems and consists of the Cas1 and Cas2 proteins, with additional involvement of Cas4 (a restriction endonuclease superfamily enzyme) (Hooton and Connerton, 2014) and in type II systems, Cas9 (an interference protein) (Heler et al., 2015). The expression and interference modules use multisubunit nucleoprotein complexes formed of crRNA and effector Cas proteins (Brouns et al., 2008; Jackson et al., 2014; Mulepati et al., 2014; Rouillon et al., 2013; Staals et al., 2013; Staals et al., 2014; Tamulaitis et al., 2014; Taylor et al., 2015; Zhao et al., 2014) or a single large multifunctional protein like Cas9 (Deltcheva et al., 2011; Jinek et al., 2012; Jinek et al., 2014) or Cpf1 (Fonfara et al., 2016; Zetsche et al., 2015) (Figure 1.7). The ancillary module is a combination of various proteins and domains that are often found outside of CRISPR-cas loci and are less common than the core Cas proteins in CRISPR-Cas systems with the exception of Cas4. In addition to its role in adaptation, Cas4 seems to be involved in CRISPR-Cas coupled programmed cell death (Koonin and Makarova, 2013; Makarova et al., 2012). Other components of the ancillary module include – a diverse set of proteins containing the CRISPR-associated Rossmann fold (CARF) domain (Makarova et al., 2014; Vestergaard et al., 2014), which might be involved in regulation of CRISPR-Cas activity in type I and type III system and Csn2 in type II system is a inactivated P-loop ATPase, which forms a homotetrameric ring to accommodate linear double stranded DNA in the central hole (Arslan et al., 2013; Koo et al., 2012; Lee et al., 2012; Nam et al., 2011).

Csn2 is not required for interference but apparently has a role in spacer integration, possibly preventing damage from the double strand break in the chromosomal DNA (Arslan et al., 2013; Barrangou et al., 2007). DinG (damage inducible G) family helicase association is seen in some variants of type IV CRISPR-Cas systems (Makarova et al., 2015). DinG helicase in bacteria seems to be a homolog of XPD (xeroderma pigmentosum complementation group D) helicase (White, 2009). The XPD helicase family comprises of a number of superfamily 2 DNA helicases, the founding member of which are conserved in archaea and eukaryotes. The family members are 5' to 3' DNA helicases and contain an essential iron-sulfur cluster binding domain (White, 2009).

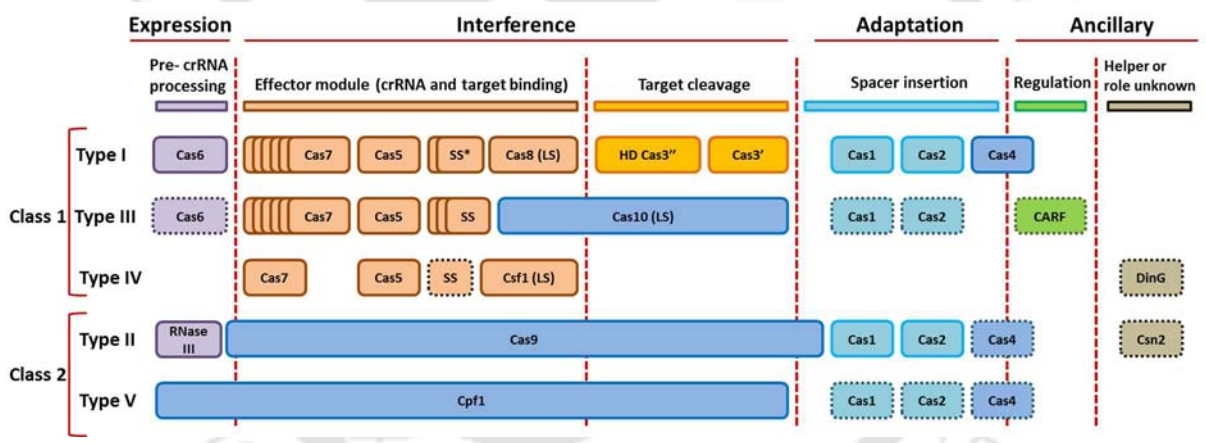


Figure 1.7 ***Functional classification of Cas proteins.*** In several type I subtypes the putative small subunit (SS) protein is fused to Cas8 (the type I system large subunit, LS), which is indicated by an asterisk. The type III system LS and type IV system LS are Cas10 and Csf respectively. In type I systems, Cas3' represents the fusion of HD nuclease domain to the superfamily 2 helicase and Cas3'' represents its encoding as a separate gene. Dashed outlines indicate the dispensable components. The functions shown for type IV and type V system components are proposed based on homology to the cognate components of other systems.

1.2.2.2. Nucleic acid component

The CRISPR loci

The CRISPR loci are present in ~40% of sequenced bacterial genomes and ~90% of archaeal genomes (Makarova et al., 2011b; Sorek et al., 2008), though the repetitive DNA accounts for less than 5% in most prokaryotic genomes (Ussery et al., 2004). The number of CRISPR loci can range from 0 to 20 per genome, with the highest number in *Methanocaldococcus jannaschii*, which harbours 18 distinct CRISPR loci (Bult et al., 1996; Godde and Bickerton, 2006; Jansen et al., 2002b; Lillestol et al., 2006; Sorek et al., 2008). In general, the archaeal clusters, especially from thermophilic organisms are multiple and larger than the bacterial ones. Clusters are also present in bacterial mega plasmids such as pTT27 of *Thermus thermophilus* and archaeal conjugative plasmids such as pNOB8 and pKEF9 of *Sulfolobus* species (Godde and Bickerton, 2006; Lillestol et al., 2006). The CRISPR locus comprises of invariant repeats interspaced by variable spacer sequences which are acquired from the invader genome. The repeat sequences vary between 21 and 48 nucleotides, while spacer sequences vary between 26 and 72 nucleotides in length (Al-Attar et al., 2011; Godde and Bickerton, 2006). Within a single CRISPR locus, all of the repeat sequences are nearly identical, except the final repeat which is frequently degenerate (Grissa et al., 2007a). The sequence of the repeat units in different CRISPR loci is not conserved and phylogenetically distant species generally show greater variation of the repeat sequences than closely related species, although there are partially conserved sequences such as a GTTT(G/C) motif at the 5'-end and a GAAAC motif at the 3'-end (Diez-Villasenor et al., 2010; Godde and Bickerton, 2006; Jansen et al., 2002b; Kunin et al., 2007). Because of the partially palindromic nature of the repeats, the transcripts from these regions can form stable, highly conserved RNA

secondary structures (Kunin et al., 2007; Makarova et al., 2006). Based on sequence similarity and secondary structure formation, the repeat sequences were classified into 12 different clusters, some of which form a stable hairpin structure due to their quasi-palindromic nature, while others remain unstructured or loosely structured (Kunin et al., 2007). The number of spacers in a CRISPR locus vary widely, from one to several hundred with as many as 587 spacers at a specific CRISPR locus in the myxobacterium *Haliangium ochraceum* DSM 14365 (Grissa et al., 2007b). The spacer sequences exhibit significant similarity to sequences from phage DNA and conjugative plasmid sequences and therefore were concluded to have originated from foreign genetic elements (Mojica et al., 2005; Pourcel et al., 2005), which is supported by the finding that 40% of the spacers in the CRISPR loci of lactic acid bacteria were homologous to streptococci phage genomes and the respective conjugative plasmids (Bolotin et al., 2005). Similarly, the crenarchaeal CRISPR spacers had shown matches to fuselloviruses, rudiviruses and β-lipothrixviruses, thereby further confirming the exogenic origin of spacers. The spacer sequences were derived from both the sense and anti-sense strands as well as from both coding and intergenic regions of phage genomes (Shah et al., 2009). The sequence in viral or plasmid DNA that is complementary to a given spacer sequence is called as a protospacer (Deveau et al., 2008).

Usually a CRISPR-Cas subtype associates with a single repeat cluster, but not all CRISPR loci have adjoining *cas* genes. Only the CRISPR loci that have adjacent *cas* genes are functionally active, whereas the others can either be non-functional or rendered functional by Cas proteins *in trans* (Diez-Villasenor et al., 2010; Horvath et al., 2009). The availability of dedicated databases – CRISPRdb and CRISPi have helped the identification of CRISPRs and Cas proteins on sequenced genomes (Grissa et al., 2007a; Rousseau et al., 2009).

PAM

The protospacer adjacent motif (PAM) is a DNA sequence immediately following the protospacer. These are short conserved regions typically 2-3 nt occurring in close proximity protospacer sequence, as revealed by the alignment of protospacer flanking sequences in phage genomes (Bolotin et al., 2005; Deveau et al., 2008; Mojica et al., 2009; Semenova et al., 2009). The presence of these motifs reveal that protospacers are not randomly selected, as these conserved sequences provide a recognition signal for the selection of target sequences that will become new spacers. PAM are also recognized by CRISPR machinery during interference stage for targeting but in some cases phages evade CRISPR immunity by mutating residues of the PAM (Deveau et al., 2008; Semenova et al., 2009).

Leader sequence

Leader sequence is located directly upstream of the repeat cluster, with respect to the strand orientation of the repeat sequence. The sequence is rich in homopolynucleotide regions having high adenine or thymine (A-T) content, lacking open reading frames and typically hundreds of nucleotides long (100-550 bp) (Jansen et al., 2002a). They are generally not conserved between distantly related species and show high variability (Jansen et al., 2002a), but exhibit similarity between related species. Analysis of the primary transcripts of CRISPR loci in several species, revealed that transcription initiates within the leader region (Hale et al., 2009; Lillestol et al., 2006) and subsequently the promoter elements were identified *in vitro* and *in vivo* in the leader regions of *E. coli* K12 (Pougach et al., 2010; Pul et al., 2010) and *Sulfolobus acidocaldarius* (Lillestol et al., 2006; Lillestol et al., 2009). Another study in *Staphylococcus epidermidis* RP62a also showed the requirement of leader sequence for

transcription of the CRISPR locus (Marraffini and Sontheimer, 2008). Thereby, confirming that these regions act as transcription promoters for the sense strand of the CRISPR arrays (Hale et al., 2012) and are also the binding sites for regulatory proteins. Moreover, it was initially deduced by comparative analysis (Lillestol et al., 2006) and subsequently confirmed by genetic studies in *Streptococcus thermophilus* (Barrangou et al., 2007) that novel spacers are incorporated along with a new repeat into the leader proximal end of the CRISPR loci, thus exhibiting polarized acquisition of new spacer sequences at the leader end of the CRISPR array (Figure 1.8). Therefore, leader regions seem to be playing the dual role of controlling CRISPR transcription and the growth of the array, by interacting with the appropriate proteins for the addition of new spacers. In other words, the leader sequences may contain binding sites for Cas proteins involved in CRISPR adaptation (Al-Attar et al., 2011).

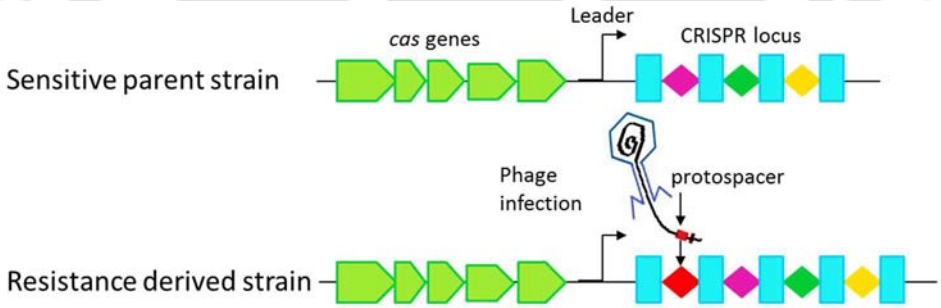


Figure 1.8 ***The polarized acquisition of spacers near the leader end of CRISPR locus.*** The leader region is followed by invariant repeats shown in blue vertical rectangles and variable spacers are shown in diamonds (differently coloured to show the variability). The *cas* genes are shown in pentagons. The acquisition of the new spacer at the leader proximal end of the CRISPR array is shown by diamond in red. This polarized acquisition maintains the chronology of the infection, as indicated by the newly acquired spacer near the leader.

CRISPR RNA (crRNA)

The CRISPR array is transcribed to form a single transcript, which is processed by endoRNases of the CRISPR-Cas system to form short crRNA, that serves as guide to target the invading DNA with the help of effector protein or protein complex called as Cascade. During processing, the endoRNases cleave in-between the repeats of CRISPR transcript (consisting of repeat-spacer units), liberating each spacer unit flanked by the repeat sequence on both sides, which in some cases may get further trimmed on edges forming a mature crRNA. This mature crRNA associates with effector Cas proteins to form a surveillance complex ready for target degradation using the sequence complementarity provided by the spacer region of crRNA. Thus, crRNA serves as a guide.

1.2.3. Parallels and distinction between CRISPR-Cas system and RNAi

The prokaryotic CRISPR-Cas systems holds functional analogy to the eukaryotic RNA interference (RNAi) system (Makarova et al., 2006) (Figure 1.6). RNAi is a process of sequence specific gene silencing in eukaryotes, which is mediated by a variety of non-coding small RNA species with the aid of specific protein complexes. The various pathways of this mechanism involve RNA species of different origin which serve different functions (Carthew and Sontheimer, 2009; Hannon, 2002; Malone and Hannon, 2009; Meister and Tuschl, 2004). These include the genome encoded microRNAs (involved in post-transcriptional regulation of gene expression in the miRNA pathway), the short interfering RNAs (siRNAs) processed from the exogenous dsRNA (involved in silencing of the invading element) and the endogenous piwi-interfering RNAs (involved in transposons silencing in animal germ cell lines). In general, RNAi pathway operates in two stages – the generation of small RNAs

which will serve as guide followed by targeting using the ribonucleprotein complex called RNA-Induced Silencing Complex (RISC) (Carthew and Sontheimer, 2009; Siomi and Siomi, 2009). In case of siRNA generation, the long dsRNA molecules of exogenous origin are processed into 21-25 nt duplex RNA fragments having a 3' overhang (siRNAs) by an RNase III endonuclease called Dicer, which is a nonspecific nuclease but operates as a molecular ruler to determine the size of the processed siRNAs. In contrast, the precursor molecules for miRNAs are transcribed from the organism's chromosome usually encoded within introns and fold into imperfect stem-loop structures which are processed by Drosha in nucleus and then translocated to cytoplasm for further processing by Dicer, to yield miRNAs of 21-25 nt length. The generated small RNAs are unwound and one strand termed the guide strand is selected for incorporation in a large protein complex to form RISC. The composition of the RISC varies in different organisms, but the core component is always a member of the Argonaute protein family, also known as Slicer. This protein binds the single stranded guide RNA and uses it to locate complementary RNA strands, which are subsequently cleaved by a conserved domain of the Argonaute (PIWI domain). The cleavage site is determined again by a molecular ruler mechanism, based on the size of the guide RNA. The RISC complex can then be recycled and used for repeated silencing. The outcome of the RISC encounter with the target RNA depends largely on the degree of complementarity exhibited between the guide and the target RNA, the type of Argonaute protein, the specific subunits of the RISC complex and other proteins interacting with the target and/or the RISC (Figure 1.9). Though the CRISPR-Cas system seems more similar to PIWI-interacting RNAs (piRNAs), which protect genome integrity from parasites such as transposons (Siomi et al., 2011), the molecular commonalities between CRISPR and RNAi include:

1. Both are mediated by small noncoding RNAs, which in conjunction with a ribonucleoprotein (RNP) complex mediates the sequence-specific cleavage of target nucleic acids.
2. There are similarities in guide RNA. They are derived from long RNA precursors and contain invader-derived sequences.
3. There are functional similarities between the proteins involved in the biogenesis of small interfering RNAs.
4. There are mechanistic and structural commonalities such as the formation of multisubunit RNP surveillance complex between the eukaryotic RISC and prokaryotic Cascade complex.
5. There are abundance of conserved RNA and DNA-manipulating domains in the proteins associated with each system (Marraffini and Sontheimer, 2010).

In spite of these similarities there are notable differences between the CRISPR-Cas and RNAi system, which include:

1. The primary target for CRISPR-interference is dsDNA, although RNA can be targeted by some CRISPR-Cas systems like the Cmr complex in *P. furiosus* (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Hale et al., 2009).
2. The protein machineries and the small RNA biogenesis of both the systems are different. Some key proteins do not have direct functional or structural equivalents like some of the CRISPR universal and signature proteins and the eukaryotic Argonaute.
3. Each system seems to have distinct physiological roles, with RNAi involved in gene regulation, chromosome stability, transposon and invader silencing while the CRISPR system seems to be predominantly an immune system (Horvath and Barrangou, 2010;

Marraffini and Sontheimer, 2010), though it also has regulatory role in host cells owing to its modular architecture, discussed in section 1.1.4.

4. The CRISPR-Cas system differs from RNAi system in the time-resolved activity, *i.e.*, the CRISPR-Cas system can readily acquire new spacers or conversely lose the old spacers which allows it to respond dynamically to viral predators, which are also evolving at high rates. Further, this ability of dynamic acquisition of foreign DNA and its subsequent use to fight the invading genetic material forms the acquired and heritable immunity, that passes the updated spacer set to the progeny, which reflects a Lamarckian mode of evolution, that does not occur in eukaryotes (Barrangou and Marraffini, 2014; Haerter and Sneppen, 2012; Koonin and Makarova, 2013). Thus, the CRISPR-Cas system provide a unique opportunity to observe and model coevolution between host and virus in natural environments or in controlled settings, because acquisition and immunity occur on short time scales and evidence of past genetic aggressions can be deduced.

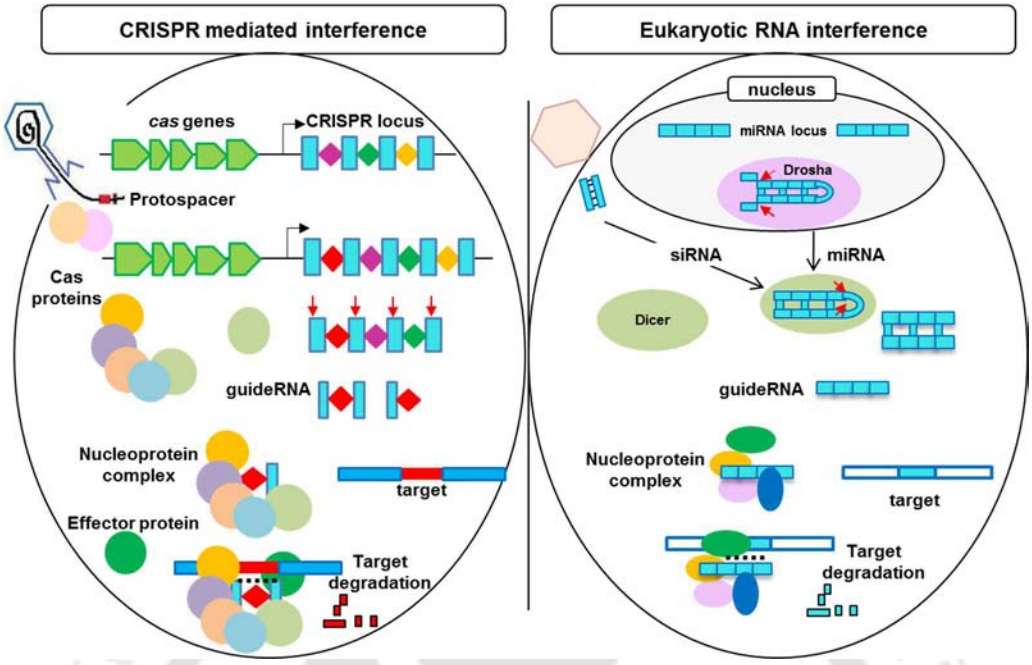


Figure 1.9 ***Parallels and distinctions between CRISPR-Cas and RNAi systems.*** CRISPR-Cas system acquires the protospacer (the exogenic DNA) as spacers in between its repeats and processes the transcript to generate guide RNA called crRNA. While in RNAi the sources of long dsRNA precursors can be genomic or exogenic, which is processed by Dicer to generate the mature siRNAs or miRNAs. The guide strand of each crRNA or siRNA/miRNA gets loaded onto the Cas proteins complex or Argonaute protein to form nucleoprotein complex called Cascade and RISC respectively and used to recognize the complementary target.

1.2.4. Stages of CRISPR-Cas immunity

The functioning of CRISPR-Cas defense system can be categorized under three mechanistically distinct stages – (i) spacer acquisition, (ii) expression and maturation, and (ii) interference (Brouns et al., 2008; Makarova et al., 2006; van der Oost et al., 2009; Wiedenheft et al., 2012) (Figure 1.10). Though all these processes can work independently, but to operate as a defense system all three stages must be functional.

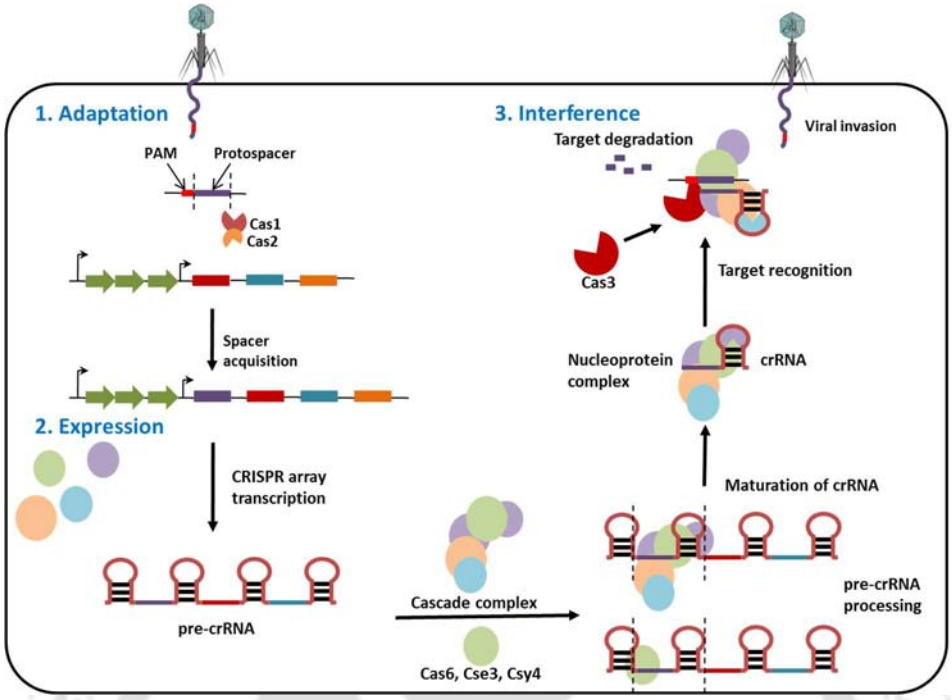


Figure 1.10. ***Mechanistic overview of CRISPR-mediated immunity.*** The different stages of CRISPR-Cas immunity as exemplified in type I system is shown. Stage 1 – An infected prokaryote incorporates a piece of invading nucleic acid (the protospacer) into a genomic CRISPR locus as a new spacer at the leader proximal end of the CRISPR. Stage 2 – The CRISPR locus is transcribed as a single pre-crRNA transcript that is processed by Cas endoRNase either alone or in association with other Cas proteins into mature crRNAs. Stage 3 – The crRNA aided large multiprotein complex targets the invading nucleic acid. The sequence complementarity between the crRNA spacer and the invading nucleic acid triggers its degradation.

1.2.4.1. Stage I: Spacer selection and integration into CRISPR arrays

The first stage is adaptation (Garneau et al., 2010; Marraffini and Sontheimer, 2010), immunization (Horvath and Barrangou, 2010) or spacer acquisition (Karginov and Hannon, 2010; van der Oost et al., 2009), which involves encounter with the invading element followed by the protospacer selection and its subsequent incorporation into the CRISPR array, as a novel spacer between two adjacent repeat units. The CRISPR locus shows polarized acquisition of spacers, *i.e.*, integration occurs at the end of the array closest to promoter/leader and an additional repeat is synthesized for every new spacer acquired (Arslan et al., 2014; Datsenko et al., 2012; Diez-Villasenor et al., 2013; Levy et al., 2015; Nunez et

al., 2015b; Swarts et al., 2012; Yosef et al., 2012). This positional information represents a timeline of spacer acquisition events. The sequences within 10 bp of the site of integration, in both the leader and repeat of the CRISPR are required for the process. The CRISPR leader-repeat junction information seems to be critical for adaptation in Type II-A system (Wei et al., 2015a) and likely in other CRISPR-Cas systems as well. Interestingly, the leader and a single repeat of the CRISPR locus were shown to be sufficient for adaptation in type II-A system from *S. thermophilus* (Wei et al., 2015a). Thus, in addition to the presence of promoter, the leader is also important in acquisition. The acquisition of new spacers largely depends on RecBCD-mediated processing of double stranded DNA breaks, occurring primarily at replication forks (Levy et al., 2015). The higher number of forks on the foreign DNA results in its preference, while spacer acquisition from the self DNA is limited due to the higher density of Chi sites (sequence octamers) (Levy et al., 2015). In CRISPR-Cas system, two modes of CRISPR adaptation are known – Naive/non-primed and primed adaptation. The minimal requirement of non-primed adaptation are the two nucleases, Cas1 and Cas2 (Arslan et al., 2014; Yosef et al., 2012). To generate immunological memory, Cas1 and Cas2 together form an integrase complex, responsible for capturing a 30-40 base pair segment of foreign DNA or protospacer and catalyzes the integration into the host genome as unique spacer sequence (Datsenko et al., 2012; Levy et al., 2015; Nunez et al., 2015a; Nunez et al., 2014; Swarts et al., 2012; Wang et al., 2015; Yosef et al., 2012). Though biased toward invading DNA, it also leads to acquisition of multiple spacers from host DNA (Yosef et al., 2012). Since many spacers acquired during non-primed adaptation originate from protospacers without a consensus PAM, they are incapable of immune response (Yosef et al., 2012). Recently, the essential role of the integration host factor (IHF) protein in specifying the leader-proximal spacer acquisition during CRISPR-Cas adaptive immunity and the requirement of target DNA bending for Cas1-Cas2 mediated spacer integration has been

identified (Nunez et al., 2016). The IHF protein binds to the leader sequence, inducing a sharp bend in DNA, which allows the Cas1-Cas2 integrase to catalyze the first integration reaction at the leader-repeat border. Further in type II systems, the additional involvement of Cas4 (Hooton and Connerton, 2014) and Cas9 are shown (Heler et al., 2015). The non-primed adaptation in type II systems, requires Cas9 to ensure that spacers are selected from protospacers with correct PAMs (Heler et al., 2015). Primed adaptation, on the other hand requires all Cas proteins and a crRNA recognizing a partially complementary spacer or a spacer with a nonconsensus PAM (Datsenko et al., 2012). It leads to highly efficient and selective acquisition of spacers with consensus PAM from protospacers located in cis with respect to the priming protospacer and most spacers are capable of protecting the host (Datsenko et al., 2012; Swarts et al., 2012). In addition to type I-E CRISPR-Cas system, primed acquisition has been demonstrated in type I-B system from an archaea *Haloarcula hispanica* (Li et al., 2014), and type I-F system from bacteria *Pectobacterium atrosepticum* (Richter et al., 2014). The existence of alternative, non-primed adaptation, was not demonstrated in these cases. Moreover, it was suggested that the *H. hispanica* adaptation is strictly dependent on priming (Li et al., 2014). Interestingly, in type I-F CRISPR-Cas system from *P. aeruginosa*, both the modes of adaptation required intact Csy complex (an ortholog of the *E. coli* Cascade) and crRNA in addition to Cas1 and Cas2, which in the case non-primed adaptation does not have to match the target DNA (Vorontsova et al., 2015). Thus, in type I-F CRISPR-Cas system, the adaptation requires all components of the interference machinery. Further in primed adaptation, the acquisition efficiency is a function of distance from the priming site and a strand bias consistent with existence of single stranded adaption intermediates is observed (Vorontsova et al., 2015). Similarly, in type II-A system from *S. thermophiles*, all the Cas proteins (Cas1, Cas2, Csn2 and Cas9) are required for adaptation (Wei et al., 2015b).

1.2.4.2. Stage II: CRISPR expression and biogenesis of crRNA

In the second stage termed as CRISPR expression, the CRISPR locus is transcribed to form a single primary transcript called the pre-CRISPR RNA (pre-crRNA) by RNA polymerase. Next, the pre-crRNA is cleaved endonucleolytically by specific endoribonuclease to yield mature crRNA, which then binds to Cas effector proteins and serves as guide in the third stage of CRISPR-mediated defense. Thus, based on its function, the crRNA is also referred as prokaryotic silencing (psiRNA) (Hale et al., 2009; Makarova et al., 2006) or guide RNA (Brouns et al., 2008; Carte et al., 2008). The expression stage can be divided into two steps – The CRISPR locus transcription and the crRNA processing, both of which are required for successful interference.

The transcription of CRISPR locus

Transcription of a CRISPR locus into a primary transcript or pre-crRNA was first observed in high throughput analyses of non-coding RNAs in the archaea *Archaeoglobus fulgidus* and *Sulfolobus solfataricus* P2 (Tang et al., 2002; Tang et al., 2005). The transcripts ranged from a minimum length corresponding to the distance between two successive repeats in the CRISPR cluster to higher order multiples of this single repeat-spacer unit. The detected sequences corresponded to various positions of the CRISPR array suggesting that the whole locus is transcribed as long transcript which is subsequently processed into smaller repeat-spacer units. Later, the transcription of CRISPR loci was shown in a number of species, such as bacterial species *E. coli* (Brouns et al., 2008; Pougach et al., 2010; Pul et al., 2010), *T. thermophilus* (Agari et al., 2010), *Xanthomonas oryzae* (Semenova et al., 2009) and archaeal species *P. furiosus* (Hale et al., 2009), *Staphylococcus epidermidis* (Marraffini and

Sontheimer, 2008), *S. solfataricus* and *S. acidocaldarius* (Lillestol et al., 2009; Tang et al., 2005). All these studies suggested the unidirectional transcription from the leader proximal end of the locus. The analysis of the transcription start sites and leader regions of the *Sulfolobales* revealed putative BRE and TATA box motifs within 25 nt of the transcription start site in the leader sequence, suggested the existence of promoter in the leader region (Lillestol et al., 2009). Also, the reverse transcripts of the repeat clusters were detected in *S. solfataricus* and *S. acidocaldarius* (Lillestol et al., 2009), suggesting the existence of putative BRE and TATA box elements downstream of the CRISPR arrays, but their processing seems to be less efficient and therefore it remains unknown whether they produce functional repeat-spacer units. Generally, Cas proteins and pre-crRNA are expressed constitutively but under certain conditions these levels can be regulated, suggesting the chances of background defensive monitoring of presence of invasive nucleic acid and the flexibility to mount a more concerted counter-attack when required. The process has striking differences among various CRISPR-Cas system, which highlights the remarkable versatility and ability of CRISPR systems to adapt and evolve according to environmental pressures. The following types of regulation have been observed in the transcription of CRISPR arrays and *cas* genes –

1. crRNAs are often identified as quantitatively dominant amount of small RNAs in bacteria and archaea, suggesting the constitutive expression of CRISPR loci, which can be further induced by viral challenge. This is consistent with a surveillance mode of action and is shown in all studied archaea (Hale et al., 2009; Lillestol et al., 2009; Semenova et al., 2009; Tang et al., 2002).
2. Expression can be upregulated in response to phage infection, under control of the cAMP receptor protein (Agari et al., 2010). This pathway also gets activated during

carbon limitation stress. Another study suggests the upregulation of *cas* gene expression in response to envelope stress (Perez-Rodriguez et al., 2011).

3. The negative regulation of *cas* operon by DevS along with the *dev* operon which control developmental stages has been observed in *Myxococcus xanthus* (Viswanathan et al., 2007). In *E. coli*, transcription is suppressed by the Heat-stable Nucleoid Structuring protein (H-NS, a typical transcriptional repressor in gram negative bacteria), which binds the promoter region in the leader sequence of the CRISPR locus (Pul et al., 2010). This repression is relieved by the transcriptional regulator LeuO, by binding to the same genomic region and reversing the cooperative binding of H-NS dimers along the DNA and also by directly or indirectly causing the enhancement of CRISPR-associated transcription (Westra et al., 2010).
4. Allosteric regulation can be exerted by a putative transcriptional regulator family, *csa3* (*casRa*), found in archaea (Lintner et al., 2011a).

Pre-crRNA processing

Precursor transcripts encompassing the full length CRISPR locus (pre-crRNA) are transcribed and cleaved within each repeat sequence to generate mature crRNA that consists of a spacer sequence flanked by portions of the repeat sequence. CRISPR-Cas immune systems from the five types utilize distinct sets of enzymes to process pre-crRNAs (Makarova et al., 2015) and thus the processing differs based on the type of CRISPR-Cas system. These endoRNases perform two specific functions, first being the recognition and processing of the precursor transcript to generate the mature form of crRNAs and second being the retention of the mature crRNA for subsequent utilization by the respective effector proteins or complexes that mediate interference. In the type I and type III systems, a CRISPR-specific

endoribonuclease Cas6 (associated with subtypes I-A, I-B, I-D, III-A and III-B) or its homolog (Cas6e formerly known as Cse3 or CasE and Cas6f formerly known as Csy4 associated with subtypes I-E and I-F respectively) bind and cleave the repeat elements in a sequence specific manner (Brouns et al., 2008; Carte et al., 2010; Carte et al., 2008; Gesner et al., 2011; Haurwitz et al., 2010; Haurwitz et al., 2012; Lintner et al., 2011b; Sashital et al., 2011; Shao and Li, 2013; Shao et al., 2016; Sternberg et al., 2012; Wang et al., 2011). These proteins are part of the RAMP superfamily (Repeat Associated Mysterious Proteins) which encompasses a large variety of protein families having tandem or single ferredoxin-like folds, also called as RNA recognition motifs (RRM) for RNA binding (Carte et al., 2010; Haft et al., 2005; Haurwitz et al., 2010; Makarova et al., 2002; Makarova et al., 2006; Wang et al., 2011). They also form large heteromeric complexes and take part in invader silencing (Brouns et al., 2008). Though these enzymes share no detectable primary sequence similarity, they all adopt ferrodoxin-like folds (Wiedenheft et al., 2012). But despite their shared fold and structural topology, each subtype endoRNase exhibits remarkably different mechanism for target RNA recognition and cleavage, although the final product is similar. This functional versatility is related to the specific repeat family of each subtype, which can form structured, unstructured or weakly structured repeats, that influences the mode of recognition of repeats and binding by the respective Cas proteins (Kunin et al., 2007). The propensity of each repeat sequence to form stable secondary structures typically a stem-loop structure depends on the palindromic nature of the repeat sequence.

The endoRNases in some cases remain bound to the processed product and become a part of the Cascade but in others only the mature crRNA is loaded onto Cascade for targeting. For example, the Cas6 from type I-A remains weakly bound to the archaeal Cascade (Lintner et al., 2011b) while Cas6e and Cas6f, the endoRNases of type I-E and I-F respectively, are single turnover catalysts and remain bound to the crRNA product after cleavage and become

part of the targeting complex along with crRNA (Brouns et al., 2008; Haurwitz et al., 2012; Hayes et al., 2016; Jackson et al., 2014; Jore et al., 2011; Sashital et al., 2011; Sternberg et al., 2012; Wiedenheft et al., 2011a; Zhao et al., 2014). In type III-A system, the CRISPR processing endonuclease Cas6 is not a part of the targeting Csm complex (Hayes and Ke, 2015; Rouillon et al., 2013; Staals et al., 2014), with the exception of *S. thermophilus* Csm complex that shows weak transient interactions with Cas6 (Tamulaitis et al., 2014). Similarly, Cas6 which is also a crRNA maturase in type III-B system is not a component of the targeting Cmr complex (Benda et al., 2014; Hale et al., 2009; Osawa et al., 2015; Spilman et al., 2013; Staals et al., 2013; Taylor et al., 2015; Zhang et al., 2012).

Interestingly, these endoRNases can process the pre-crRNA either individually or as a part of Cascade. Like in the case of type I-E system, the crRNA is processed by Cas6e as a part of Cascade, which was first characterized in *E. coli* (Brouns et al., 2008; Wiedenheft et al., 2011a). The repeat sequences of this system form a stable hexanucleotide stem with a tetranucleotide loop. Soon the structure of Cas6e from *T. thermophilus* became available (Gesner et al., 2011; Sashital et al., 2011), which revealed the presence of double ferredoxin-like fold, with a four strand antiparallel β-sheet forming the central positively charged cleft of the protein, where the phosphate backbone of the 3' strand of the stem loop is bound (Figure 1.11A). Cas6e shows a conformational change upon RNA binding, whereby a previously disordered region forms β-hairpin and recognizes the major groove of repeat stem-loop and a previously disordered loop interacts with the base of the stem loop, thus positioning the scissile phosphate in the active site. The unwinding of the base pair at the base of the stem-loop takes place, which is necessary for cleavage (Gesner et al., 2011; Sashital et al., 2011). Additionally, the protein interacts specifically with four residues located on either side of the stem loop in the single stranded region. Cleavage occurs at a G-A bond at the 3' base of the stem-loop. The mature crRNA of this system comprises of a complete spacer sequence

flanked by 8 nt of repeat derived sequence at the 5'-end (5' handle/psi-tag resulting from cleavage in the repeat 8 nt upstream of the beginning of spacer) and the remaining 21 nt of repeat containing the stem-loop on the 3'-end (3' handle), though a degree of heterogeneity was observed for the 3'-end, that highlighted the importance of the 5' handle for potential involvement in protein recognition and in self *vs.* non-self discrimination (Brouns et al., 2008; Jore et al., 2011). In type I-F systems, Cas6f from *P. aeruginosa* also adopts an N-terminal ferredoxin-like fold but its C-terminal region adopts an extended conformation, although the basic secondary structure connectivity resembles a ferredoxin-like fold (Haurwitz et al., 2010) (Figure 1.11B). An arginine rich helix in the C-terminal domain interacts extensively with the major groove of the RNA stem-loop and uses two amino acid side chains to read out the identity of the bottom two base pairs of the hairpin. The base of the stem of the repeat RNA is positioned in the positively charged cleft between the two domains. Cas6f makes a sequence specific interaction with the first single stranded nucleotide upstream of the stem-loop, but does not interact with any of the nucleotides downstream of the stem-loop (Haurwitz et al., 2010; Haurwitz et al., 2012). Sequence specific hydrogen bonds tether the substrate in the active site so that the cleavage takes place immediately downstream of the hairpin, *i.e.,* 8 nt upstream of the spacer sequence. Cas6f remains bound to the cleavage product via the base-specific and electrostatic interactions formed with the RNA, enabling the subsequent use of the mature crRNA by Csy complex (Rollins et al., 2015; Sternberg et al., 2012). A representative of the Cas6 family protein associated with subtypes I-A, I-B, I-D, III-A and III-B has been characterized in *P. furiosus, S. solfataricus and Methanococcus maripaludis* (Carte et al., 2010; Carte et al., 2008; Shao and Li, 2013; Shao et al., 2016; Wang et al., 2011). Although the architecture of Cas6 also consists of two ferredoxin-like domains, the molecular mechanism for recognition and cleavage of the pre-crRNA seems to have evolved to accommodate the unstructured repeat associated with these

subtypes (Kunin et al., 2007). In *P. furiosus* the conserved positively charged central cleft between the two ferredoxin-like domains is responsible for interaction with single stranded repeat RNA, where conserved residues form contacts with nucleotides near the 5'-end of the CRISPR repeat, anchoring it in position for the cleavage at 3'-end of the repeat on the opposite surface of the protein (Figure 1.11C). The RNA likely wraps around the protein to the opposite face where Cas6 cleaves the RNA in an A-A dinucleotide motif (Carte et al., 2010; Carte et al., 2008; Wang et al., 2011). The catalytic active site and binding site are physically distinct, and linked by substrate, which is interacting weakly or transiently with the signature Gly-rich loop. The metal independent cleavage of the pre-crRNA transcript occurs 8 nt upstream of each spacer generating the conserved 5' handle present in the mature crRNA and 22 nt repeat derived sequence at the 3'-end. The product remains bound to Cas6 until transferred to the respective effector complex (Cmr complex or an archaeal version of Cascade, in the case of *P. furiosus* which contains both type I and III systems). The mature crRNAs loaded into the targeting complex in *P. furiosus* are smaller than the initial Cas6 cleavage products which might be a result of either endonuclease or exonuclease trimming at the 3'-end (Hale et al., 2009). But this seems to vary in different organisms, like *S. solfataricus* where no trimming is observed (Shao and Li, 2013). Further, in case of *S. solfataricus* the pre-crRNA is processed through stabilization of a 3 base pair stem loop by Cas6 (Shao and Li, 2013), suggesting an intrinsic ability of *S. solfataricus* Cas6 in reinforcing a stem-loop near the cleavage site (Figure 1.11D). Another distinct mechanism of unstructured repeat recognition by Cas6 is observed in *M. maripaludis*, where Cas6 binds to two sites (cleavage site and a distal site) of the long CRISPR repeat, *i.e.,* shows dual binding (Figure 1.11E). The Cas6 recognizes a 2 base pair stem and an AAYAA loop and supplies a tyrosine residue as a nucleobase mimic, that interacts with an adenine, which helps in stabilizing the stem and thereby resulting in efficient cleavage of the pre-crRNA by the Cas6

bound to the cleavage site (Shao et al., 2016). In type III, the Cas6 family proteins have not been found to associate tightly with any effector Cas protein or complex, which possibly grants them the flexibility needed to associate with multiple subtypes that potentially differ at the interference stage. In type I-C system, Cas5d is shown to be involvement pre-crRNA processing in the absence of Cas6 (Garside et al., 2012; Koo et al., 2012; Nam et al., 2012; Punetha et al., 2014).

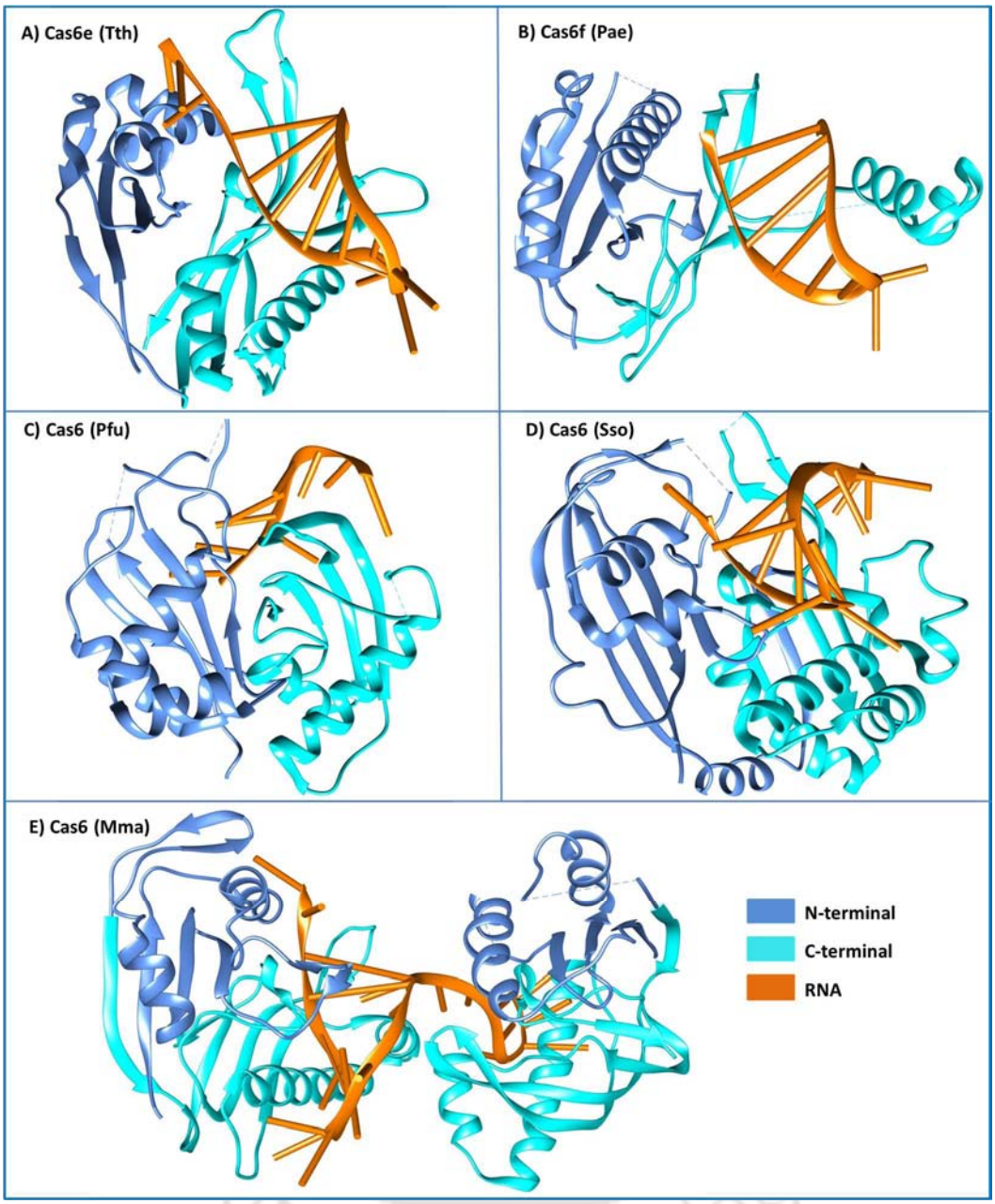


Figure 1.11. ***The distinct modes of repeat RNA recognition by Cas6 endoRNases of Type I and Type III CRISPR-Cas systems.*** The crystal structures of (A) Cas6e from *T. thermophilus* with 20 nt repeat RNA (PDB ID: 2Y8W), (B) Cas6f from *P. aeruginosa* with 16 nt repeat RNA (PDB ID:2XLK), (C) Cas6 from *P. furiosus* with 10 nt repeat RNA (PDB ID:3PKM), (D) Cas6 from *S. solfataricus* with 24 nt repeat RNA (PDB ID:4ILL) and (E) Cas6 from *M. maripaludis* with 31 nt repeat RNA (PDB ID:4Z7K) are shown. The bound RNA in some cases is only the product mimic or the minimum cleavable repeat instead of the complete repeat RNA. The N-terminal is shown in blue and C-terminal in cyan and RNA in orange. This figure was rendered using Chimera (Pettersen et al., 2004).

Thus, distinct sequence and structure specific recognition mechanisms are employed by these endoRNases in substrate selection, which could be a result of coevolution of CRISPR repeat sequences and Cas proteins (Shah and Garrett, 2011). Notwithstanding the mode of RNA recognition that seem to vary between type I and III, the endoRNases seem to follow a metal independent acid-base hydrolysis mechanism producing a cyclic 2'-3' phosphate intermediate and the final product having 5' hydroxyl group (5'-OH) and 3' phosphate (3'-P) ends (Carte et al., 2008; Gesner et al., 2011; Haurwitz et al., 2010). The deprotonated hydroxyl at the 2' position of the ribose functions as a nucleophile. The catalytic sites of all characterized Cas6-like enzymes are composed of an invariant histidine residue and a tyrosine residue in the active site along with a variable lysine or serine (Figure 1.12). However, the relative positions of these residues are poorly conserved, which might explain the observed functional variations in Cas6 activity. Moreover, these RAMP proteins employ a glycine rich loop to orient the substrate correctly.

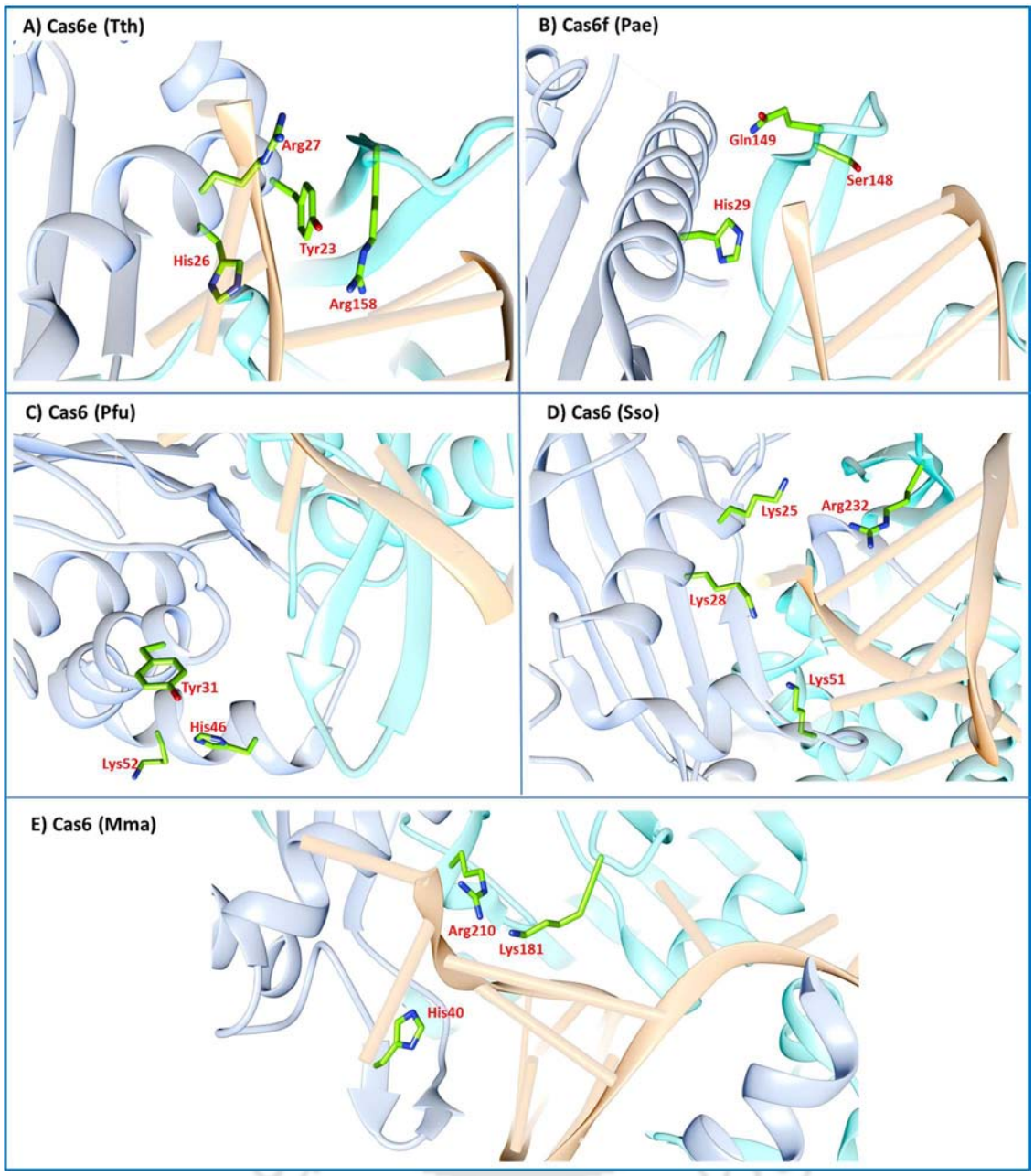


Figure 1.12 ***The active site residues of Cas6 and its homologs.*** The active site residues are shown in the crystal structures of (A) Cas6e from *T. thermophilus* with 20 nt repeat RNA (PDB ID: 2Y8W), (B) Cas6f from *P. aeruginosa* with 16 nt repeat RNA (PDB ID:2XLK), (C) Cas6 from *P. furiosus* with 10 nt repeat RNA (PDB ID:3PKM), (D) Cas6 from *S. solfataricus* with 24 nt repeat RNA (PDB ID:4ILL) and (E) Cas6 from *M. maripaludis* with 31 nt repeat RNA (PDB ID:4Z7K). The bound RNA in some cases is only the product mimic or the minimum cleavable repeat instead of the complete repeat RNA. This figure was rendered using Chimera (Pettersen et al., 2004).

Another interesting feature of these CRISPR-Cas system is that apart from the representative organism for a particular type, other organisms can harbour more than one type

of CRISPR-Cas system. This can be attributed to horizontal gene transfer either via plasmids that harbour CRISPR-cas loci or by other gene transfer mechanisms such as transposon activity, which result in the movement of CRISPR-cas loci across widely diverged lineages (Godde and Bickerton, 2006; Horvath et al., 2009; Portillo and Gonzalez, 2009). Therefore, there can be more than one endoRNase in an organism, each of which remains associated with its specific repeat cluster. For example, *T. thermophilus* has three Cas6 (TTHB192, TTHA0078 and TTHB231), which are associated with different repeat clusters and have distinct mode of substrate recognition (Niewoehner et al., 2014). Despite the differences in the endoRNases structure and the mode of crRNA recognition, the pre-crRNA cleavage catalysed by all Cas6 and its homologs yields a mature crRNA with an 8 nt 5' handle and a less well defined boundary at the 3'-end (which can be a partially cut repeat or may be trimmed product in some cases) (Brouns et al., 2008; Carte et al., 2008; Haurwitz et al., 2010). The generated crRNA therefore consists of three elements, a conserved repeat derived 5' handle (responsible for recognition and binding by the Cascade-like effector complexes and discrimination of target from self), a spacer sequence (responsible for target recognition by base pairing) and a heterogeneous repeat derived 3'-end, with a size range from 0 to 22 nt (Brouns et al., 2008; Carte et al., 2008; Hale et al., 2009; Haurwitz et al., 2010; Lintner et al., 2011b). The processing events that result in trimming of the 3'-end and the functional significance of this heterogeneity are still not known.

The type II CRISPR-Cas systems have quite different procedure for CRISPR RNA maturation in which a host factor is implicated in CRISPR RNA processing (Deltcheva et al., 2011; Jinek et al., 2012). Also, a novel RNA species identified as the transcript of the opposite strand of a region upstream from the start of the *cas* operon and the CRISPR array is found in high copy number. Interestingly, in *S. pyogenes* 25 nt region of this transcript, termed tracrRNA (trans-activating CRISPR RNA) is complementary to the unstructured

repeat sequence with only one mismatch. An RNA duplex formation between the tracrRNA and a repeat sequence of the pre-crRNA is facilitated by Cas9, which is sufficient to guide the cleavage of both strands at specific positions within the duplex region by the host RNase III, producing crRNA units that consist of a complete spacer sequence flanked by the partial repeats. Further processing takes place on the 5'-end of the spacer sequence by an unidentified nuclease, resulting in the mature crRNA (Deltcheva et al., 2011). The mature crRNA comprises of a 5' 20 nt spacer derived sequence and a 19-22 nt repeat derived sequence on the 3'-end which is strikingly different from the mature form of crRNAs found in types I and III in that it lacks the characteristic 5' repeat derived handle. This feature indicates a distinct mechanism for crRNA recognition by Cas9 (Csn1), which mediates the interference.

In the recently identified type IV systems, the crRNA processing still needs to be discovered (Makarova et al., 2015), while in type V system there is a single multifunctional protein, Cpf1 which like Cas9 of type II systems is involved in interference stage, but in contrast to Cas9, it is involved in CRISPR RNA processing also (Fonfara et al., 2016). The dual nuclease, Cpf1 from *Francisella novicida* is specific to crRNA biogenesis and target DNA interference. It cleaves pre-crRNA upstream of a hairpin structure formed within the CRISPR repeats and generates intermediate crRNAs which are further processed to form mature crRNA. Since Cpf1 is both RNase and DNase, the dual nuclease requires sequence and structure specific binding to the hairpin of crRNA repeats, for which it uses distinct active domains and cleaves nucleic acids in the presence of magnesium or calcium.

Thus, the vivid modes of recognition and crRNA generation highlight the exceptional economy and versatility of CRISPR-Cas systems.

1.2.4.3. Stage III: Recognition of invader sequences and target interference

In the third and final stage called as interference (Deveau et al., 2010) or immunity (Garneau et al., 2010), the crRNAs within a multiprotein effector complex, called Cascade recognizes and base pairs specifically with regions of incoming foreign DNA (or RNA) that have perfect or almost perfect complementarity with the spacer region of the guide RNA (Brouns et al., 2008), which triggers the Cas-mediated degradation of the invading nucleic acid (Garneau et al., 2010; Wiedenheft et al., 2012). Not all regions of a crRNA spacer are equally important for target recognition, it usually requires a seed sequence (a region of 7 to 8 nt at the 5'-end of the crRNA spacer with high affinity binding). Even single mutations in the seed region abolishes the targeting capability, whereas as many as five mutations can be tolerated elsewhere in the spacer sequence (Semenova et al., 2011; Wiedenheft et al., 2011b). A successful targeting and silencing also requires the presence of a PAM sequence in the viral or plasmid genome (Mojica et al., 2009). A non-functional PAM or mutations in PAM disrupts the ability of the CRISPR system to target and destroy invading DNA, despite perfect matches between spacer and protospacer sequences (Deveau et al., 2008; Garneau et al., 2010; Semenova et al., 2011). The PAM is so crucial because it plays a role in distinguishing between self and non-self target sequences, as the genomic CRISPR loci themselves harbour the appropriate targeting sequence, the cleavage of which would be lethal. Therefore, the presence and recognition of the PAM in invading DNA is necessitated (Semenova et al., 2011). The unique occurrence of the PAM on the invading foreign DNA (and conversely, its absence in the host spacer sequence) plays a dual role – first in spacer selection during acquisition and second in the interference process for discrimination of self *vs*. non-self. The CRISPR-Cas system versatility is evident by the *in vivo* activity assays, which demonstrated that crRNA can target either coding or non-coding regions and template

or non-template strands of viral and plasmid DNA and result in silencing, suggesting the direct targeting of the viral or plasmid DNA rather than its messenger RNA (mRNA) (Barrangou et al., 2007; Brouns et al., 2008; Deveau et al., 2008; Garneau et al., 2010; Gudbergsdottir et al., 2011; Manica et al., 2011; Marraffini and Sontheimer, 2008).

The 'class 1' categorized type I, type III and recently discovered type IV system employ a multisubunit complex (like Cascade, Csm or Cmr complexes) during interference, while 'class 2' category type II and recently found type V system utilize a single effector protein (like Cas9 and Cpf1) (Makarova et al., 2015). The effector proteins involved in target degradation in various systems include Cas3 (in type I), Cas10 (in type III), Cas9 (in type II) and Cpf1 (in type V), while the type IV effector protein still needs to be explored. The availability of the structural information of various effector complexes has provided insight into the functioning of the CRISPR-Cas system. In type I systems, the multiprotein complexes responsible for targeting have been identified in several subtypes. In type I-E, the cascade consists of a non-stoichiometric distribution of five Cas proteins ($Cse1_1$:$Cse2_2$:$Cse4_6$:$Cas5_1$:$Cas6e_1$) and a single crRNA (Brouns et al., 2008; Jore et al., 2011; Wiedenheft et al., 2011a). High resolution crystal structures are now available for Cascade (Jackson et al., 2014; Zhao et al., 2014) and Cascade bound to the target – single stranded DNA (Mulepati et al., 2014) and double stranded DNA (Hayes et al., 2016; van Erp et al., 2015). The structure reveals that the crRNA runs the entire length of the complex, with the six Cse4 (CasC) backbone subunits displaying the spacer sequence of crRNA and the crRNA processor Cas6e remains bound to the hairpin sequence of the CRISPR repeat. In type I-F, the cascade complex (Csy complex) comprises of four Cas proteins and a crRNA ($Csy1_1$:$Csy2_1$:$Csy3_6$:$Cas6f_1$), as demonstrated in *P. aeruginosa* (Rollins et al., 2015; Wiedenheft et al., 2011b). A low resolution small angle X-ray scattering (SAXS) reconstruction of the Csy complex revealed similarity to the type I-E Cascade in gross

morphology (Wiedenheft et al., 2011b). In both type I-E and I-F systems the target is recognized by base pairing to an 8 nt or 7 nt seed sequence located at the 5'-end of the spacer sequence in the crRNA. Affinity for the rest of the spacer sequence was much smaller, accounting for the tolerance for protospacer mismatches as observed in many cases (Deveau et al., 2008; Gudbergsdottir et al., 2011). In type I-A system, the archaeal Cascade (aCascade) contains Csa2, Cas5a, Cas6, Csa5, and crRNA as shown in *S. solfataricus* (Lintner et al., 2011b). The minimal requirements for stable complex formation and single stranded DNA target binding are Csa2 and Cas5a. The overall structure is similar to Cascade and the Csy complex (Lintner et al., 2011b; Reeks et al., 2013). In type I systems, the cascade together with Cas3 that contains an HD nuclease domain and a DExD/H box helicase domain (Haft et al., 2005; Makarova et al., 2006), targets invading complementary DNA (Brouns et al., 2008). The ATP dependent helicase activity of Cas3 facilitates the unwinding of DNA-DNA or DNA-RNA duplexes and the DNase activity plays a key role in degrading the target bound to Cascade (Howard et al., 2011; Sinkunas et al., 2011; Westra et al., 2012a). The HD nuclease domain can either be fused to the superfamily 2 helicase Cas3' (Gong et al., 2014; Huo et al., 2014; Sinkunas et al., 2011) or encoded by a separate gene, Cas3''(Beloglazova et al., 2011; Huo et al., 2014). The crystal structure of Cas3 HD domain from *T. thermophilus* (Mulepati and Bailey, 2011) became available, followed by the availability of complete structures of Cas3 (Gong et al., 2014; Huo et al., 2014), which provided mechanistic insights into Cascade activated DNA unwinding and degradation. Recently, it was shown that the anti-CRISPR protein AcrF3 produced by bacteriophage binds tightly to Cas3 leading to inhibition of target interference (Wang et al., 2016) suggesting the innovative counter measure effected by phages to escape the CRISPR immunity.

In type II systems a single large multi-functional protein, Cas9 is required for DNA targeting and degradation guided by a dual-RNA heteroduplex consisting of a crRNA and a

tracrRNA (Deltcheva et al., 2011; Jinek et al., 2012). Cas9 contains a McrA/HNH nuclease domain and a RuvC-like (RNase H fold) nuclease domain (Makarova et al., 2006) and produces double stranded DNA breaks in the vicinity of the PAM (Barrangou et al., 2007; Jinek et al., 2012; Nishimasu et al., 2014), by each domain cleaving one strands of the target DNA. The structure of Cas9 revealed an RNA-mediated conformational activation (Jinek et al., 2014). The dual crRNA:tracrRNA, when engineered as a chimeric single guide RNA by connecting the 3'-end of crRNA to the 5'-end of tracrRNA with a linker sequence, efficiently directs Cas9 to cleave target DNA sequences matching the 20 nt guide RNA sequence. The Cas9 along with the single guide RNA can be used to introduced site specific dsDNA breaks in eukaryotic cells, which can be repaired by non-homologous end joining (NHEJ) or homology-directed repair (HDR) resulting site specific modifications (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013). Thus, Cas9 can be programmed to target any DNA site adjacent to PAM which provide effective tool for genome editing and regulation in wide range of organisms (Charpentier and Marraffini, 2014; Hsu et al., 2014). Realizing the potential of Cas9, several variants were engineered, like the nuclease inactive Cas9 (D10A/H840A, dCas9) from *S. pyogenes*, the structure of which in complex with single guide RNA is available (Jiang et al., 2015). Soon, various crystal structures of Cas9 in complex with guide RNA and double stranded target DNA became available (Hirano et al., 2016a; Jiang et al., 2016; Nishimasu et al., 2014; Olieric et al., 2016), which increased our understanding of type II system machinery including its target recognition and priming the cleavage by forming the R-loop. It also revealed the importance of PAM in target recognition in type II systems (Anders et al., 2014), which led to determination of the structural basis of PAM recognition (Hirano et al., 2016b). Structural plasticity of PAM recognition is observed in case of engineered variants of the RNA-guided endonuclease Cas9 (Anders et al., 2016).

In case of type III systems, the target degradation occurs by a distinct HD nuclease domain fused to Cas10, which cleaves single stranded DNA during interference (Jung et al., 2015) in presence of Cas7 (Brendel et al., 2014; Osawa et al., 2015; Ramia et al., 2014; Samai et al., 2015; Tamulaitis et al., 2014; Taylor et al., 2015; Zhu and Ye, 2015). The type III systems typically utilize the large multiprotein assemblies known as the Csm (in III-A and III-D) and the Cmr (in III-B and III-C) complexes, to target the invading DNA or RNA respectively, which are directed by cognate crRNA. The *csm* operon of the type III-A and type III-D system encodes Csm proteins, that form a ribonucleoprotein complex with the mature crRNA (Hatoum-Aslan et al., 2013; Makarova et al., 2015; Rouillon et al., 2013). The immunity provided by type III-A CRISPR-Cas system has been characterized (Hatoum-Aslan et al., 2014; Hayes and Ke, 2015). In type III-A the Cas7-like Csm3 forms the backbone of the complex and binds RNA in a sequence-independent manner (Hrle et al., 2013; Koonin and Makarova, 2013; Rouillon et al., 2013). After primary cleavage of the pre-crRNA, the guide RNA formation involves secondary trimming at the 3'-end by a ruler like mechanism (Hatoum-Aslan et al., 2011), in which each Csm3 subunit binds and extends 6 nt segments of the mature crRNA and exposes unbound 3'-end for cleavage by an unknown nuclease (Hatoum-Aslan et al., 2013). The structure of the CRISPR interference complex Csm holds key similarities with type I Cascade (Rouillon et al., 2013). Initially, the type III-A systems were known to target DNA but recently, the type III-A Csm complex were found to target and degrade RNA *in vitro* in *T. thermophilus* (Staals et al., 2014) and *S. thermophilus* (Tamulaitis et al., 2014). Also the crystal structure of the Csm3-Csm4 subcomplex in the type III-A CRISPR-Cas interference complex is available (Numata et al., 2015). The Cmr complexes (in type III-B and type III-C) are composed of Cmr proteins and a crRNA. The type III-B is well characterized, while type III-C is recently identified. The Cmr complexes from archaea *P. furiosus* (Hale et al., 2014; Hale et al., 2012; Hale et al., 2009) targeted

single stranded RNA while the Cmr complex of archaea *S. solfataricus* (Zhang et al., 2012) targeted mRNA. The crystal structure of the Cmr1 subunit (Jung et al., 2015; Sun et al., 2014) and Cmr4 subunit (Benda et al., 2014; Zhu and Ye, 2015) of the Cmr interference complex is available. The crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog are available, which reveals the mechanism for specifying the periodic target cleavage sites from the crRNA 5' tag (Osawa et al., 2015). Cmr3 recognizes the crRNA 5' tag and defines the start position of the guide-target duplex, using its idiosyncratic loops. Cmr4 is the slicer in the RNA-targeting Cmr CRISPR complex. The β-hairpins of three Cmr4 subunits intercalate within the duplex, causing nucleotide displacements with 6 nt intervals, and thus periodically placing the scissile bonds near the crucial aspartate of Cmr4. The electron microscopy of the *S. solfataricus* Cmr complex reveals a morphology similar to a crab claw attached to a protruding region, which is not similar to Cascade or aCascade or the Csy complex (Zhang et al., 2012). Moreover, the Cmr system of *P. furiosus and S. solfataricus* do not require a PAM for efficient targeting (Hale et al., 2012; Zhang et al., 2012). All type III CRISPR-Cas systems contain genes encoding RAMP proteins, which is suggestive of RNA as the target. A large number of archaeal species contain more than one CRISPR-Cas system, which expands the repertoire of targets that can be recognized by an organism and adds more levels of regulatory control.

The type IV and type V systems are recently identified and very less information is available. The type IV system seems to utilize multiprotein assembly during interference, while the type V system utilizes a single large multifunctional protein, Cpf1 guided by the single mature repeat-spacer crRNA for target interference, which after recognizing a 5'-YTN-3' PAM on the non-target DNA strand introduces double stranded breaks in the target DNA (Makarova et al., 2015; Zetsche et al., 2015). Among the so far known CRISPR-Cas systems, the type V seems to constitute the most minimalistic components.

1.3. Definition of the problem

In 2010, the area of CRISPR-Cas immune system was at its beginning stage when the author joined the lab and not much information was available about its functioning. In fact, the adaptable and heritable immunity provided by CRISPR-Cas system was evidenced only in 2007. The CRISPR-Cas system machinery operated in three stages involving adaptation, expression and maturation, and target degradation. Amidst this, the mature CRISPR RNA played the pivotal role in CRISPR-Cas mediated immunity, by serving as a guide for the Cas proteins to target the invading nucleic acid. The sequence complementarity between the target and the guide RNA, triggered its degradation. Therefore, the generation of the crRNA held prime importance for the functioning of the CRISPR-Cas system. The mandate to generate crRNA was fulfilled by the pre-crRNA specific endoribonucleases, that processed the pre-crRNA transcript to form a mature guide RNA and thereby had an indispensable role in the CRISPR-mediated immunity. The repeat sequences of the long pre-crRNA were cleaved by the endoRNase Cas6 or its homologs, resulting in a library of crRNAs, each containing a unique spacer sequence flanked by fragments of the adjacent repeats. Interestingly, the repeats in various CRISPR types and subtypes differ in their sequence and structure, which suggests the variations in the mechanism of substrate recognition among different subtypes. Moreover, the repeat can attain a structured architecture in the form of stem-loop or remain unstructured. Thus, the specific recognition of a repeat RNA requires a distinct mechanistic solution for substrate discrimination by the endoRNase. Therefore, the maturation of crRNA captured the interest.

Among the different types of CRISPR-Cas system, the type I was the most widespread system in the prokaryotes, encompassing the highest number of subtypes. This prompted the author to investigate the crRNA maturation in the most prevalent CRISPR-Cas

defense system. Interestingly, a close inspection of the type I subtypes revealed that type I-C system was a paradox among the type I systems, owing to the absence of Cas6 or its homolog, which were involved in crRNA maturation. This raised the question that in the absence of Cas6, how the system compensates the function to make the CRISPR-Cas machinary functional. Another intriguing feature of this system was the presence of only three subtype specific genes (Figure 1.13), which seemed to be a minimal set for then known functional CRISPR-Cas machinery. Since the likely theme, commonly opted by the type I systems was the formation of multiprotein complexes during maturation/interference, the exploration of how these three form the Cascade complex spawned profound interest. Therefore, the type I-C CRISPR-Cas system was chosen to investigate the crRNA maturation and the following objectives were framed.

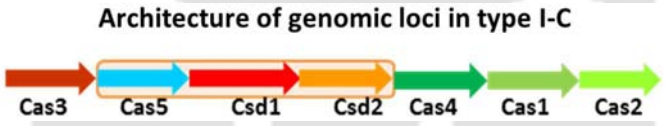


Figure 1.13 ***The architecture of the genomic locus in type I-C CRISPR-Cas system.*** The three subtype specific protein, which might form the surveillance complex of type I-C are boxed in orange.

1.4. Objectives

1. To identify the CRISPR RNA processing endonuclease in type I-C system.
2. To characterize the subtype specific proteins *viz.*, Cas5d, Csd1 and Csd2.
3. To investigate the formation of type I-C Cascade and probe its functionality.

These objectives are addressed in the subsequent chapters.

2.1. Introduction

The crRNA plays a key role in CRISPR-Cas mediated immunity. Among the type I systems, the maturation of crRNA is well characterized in type I-E system (Gesner et al., 2011; Niewoehner et al., 2014; Sashital et al., 2011), where the pre-crRNA is processed by Cas6e, a type I-E subtype specific endoRNase, as a part of Cascade. The I-E Cascade is built by a single 61 nt long crRNA and eleven subunits from five different Cas proteins (Cse1, Cse2, Cse4 and Cas5e), resulting in a total crRNP mass of 405 kDa (Brouns et al., 2008; Jore et al., 2011; Makarova et al., 2013; Wiedenheft et al., 2011a). Further, the molecular details of protein-protein and protein-RNA interactions were revealed by the high resolution crystal structures of the type I-E Cascade, that became available over time (Hayes et al., 2016; Jackson et al., 2014; Mulepati et al., 2014; van Erp et al., 2015; Zhao et al., 2014). The 3.05 Å crystal structure of type I-E Cascade bound to crRNA from *E. coli* (PDB ID: 4U7U) (Zhao et al., 2014) is shown (Figure 2.1).

In contrast to this, type I-C system seems to utilize only three Cas proteins *viz.*, Cas5d, Csd1 and Csd2 to form the nucleoprotein surveillance complex (Figure 2.1). Therefore, this encouraged us to explore how just three proteins will form a fully functional Cascade complex (Chapter 5). Also, the surprising absence of the known pre-crRNA processor Cas6 or its ortholog in type I-C system prompted us to investigate the crRNA maturation. To address this, we characterized each of the three subtype specific proteins namely Cas5d, Csd1 and Csd2 from *B. halodurans* and explored the possibility of being a potential nuclease. The characterization of Cas5d is presented in this chapter.

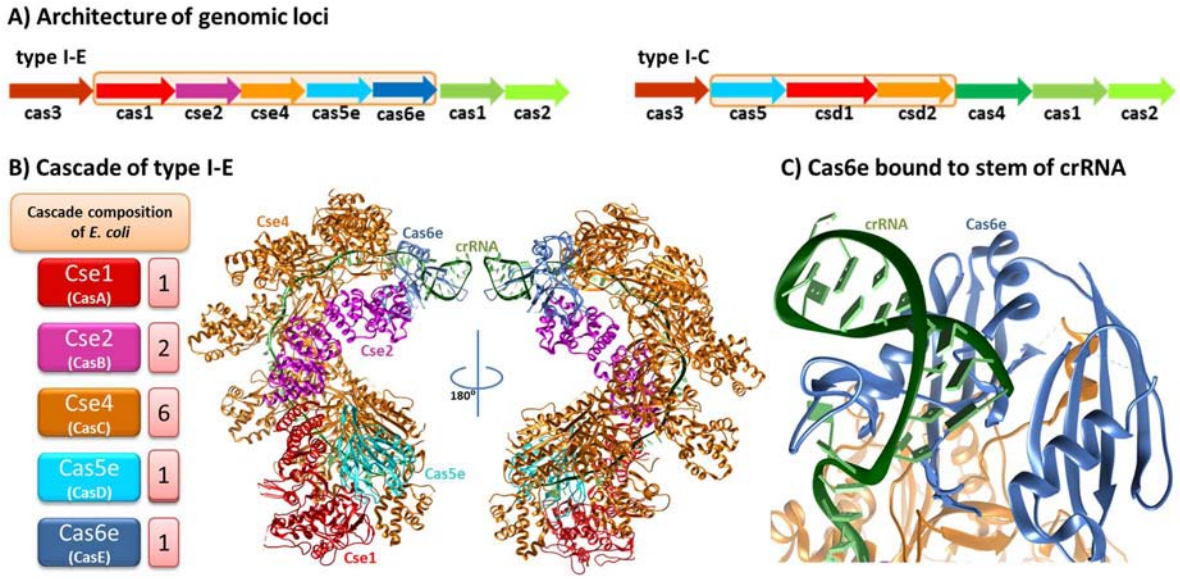


Figure 2.1 ***The Cas proteins of type I-E and type I-C systems.*** (A) Architecture of the genomic loci of type I-E and type I-C system is shown. The components enclosed in the box form the multiprotein assembly in the respective system. (B) Subunits of type I-E Cascade are shown with their copy number present in Cascade (PDB ID: 4U7U). Cse1 is shown in red, Cse2 in purple, Cse4 in orange, Cas5e in cyan and Cas6e in blue. (C) The type I-E Cas6e bound to crRNA is shown. The pre-crRNA in type I-E is processed by Cas6e as a part of Cascade. The Cas6e recognizes the stem-loop of the pre-crRNA and after processing remains bound to the stem of crRNA. The crRNA is shown in green and Cas6e in blue. This figure was rendered using Chimera (Pettersen et al., 2004).

2.2. Materials and methods

2.2.1. Cloning, expression and purification

Gene encoding *cas5d* was amplified from *B. halodurans* genomic DNA using gene specific primers with Pfu DNA polymerase (Fermentas). Amplicon of *cas5d* was cloned into pQE2 to create pCas5d, using the restriction sites for NdeI and PstI (New England Biolabs). Point mutants (Y35F, K39A, H169A, W47F, W187F, Y46F, K116A, H117A) of Cas5d were generated by mega primer based PCR method. All mutants were cloned in pQE2 which encodes an N-terminal $(His)_6$ tag, except H117A that was cloned in LIC vector (a kind gift from Scott Gradia Addgene ID: 29717) having a pET backbone and encodes an N-terminal

Strep-tag II. The cloned constructs were verified by sequencing. Two random mutations (G158R and Y162H) have been identified in Cas5d wild type construct; however, these point mutations are located distantly from the catalytic triad and have no apparent effect on the nuclease activity. All the reported point mutants were generated on this genetic background. Expression was performed in *E. coli* BL21(DE3) by growing the cells in LB medium supplemented with ampicillin (100 µg/ml) at 37°C until OD at 600 nm reached 0.7. The temperature was then reduced to 20°C for 20 min and protein expression was induced by the addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) followed by incubation at 20°C overnight. The cells were harvested by centrifugation and resuspended in buffer A containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6 mM β-mercaptoethanol (β-ME) and 1 mM phenylmethanesulfonyl fluoride (PMSF). After sonication, the lysate was clarified by centrifugation at 36,500g for 30 min. The supernatant was treated with RNase to remove any bound RNA and then loaded onto a 5 ml HiTrap IMAC HP column or StrepTrap HP column (GE Healthcare) pre-equilibrated with buffer B containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6 mM β-ME. After washing the column with buffer C containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6 mM β-ME and 40 mM imidazole, the bound protein was eluted using a linear gradient of imidazole (upto 500 mM) in buffer C. For strep-tagged proteins, washing buffer contained 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6 mM β-ME and the elution was carried out with buffer C that contained 2.5 mM D-Desthiobiotin in place of imidazole. The eluted protein was incubated with 10 mM EDTA for 1hr to remove the bound metal ions if any and then dialyzed against buffer D containing 20 mM Tris-HCl (pH 8.0), 200 mM NaCl and 6 mM β-ME. Subsequently, the proteins were aliquoted, snap frozen in liquid nitrogen and stored at -80°C until required.

2.2.2. Preparation of substrates

Pre-crRNA containing only the repeat sequence was chemically synthesized and differently end-labelled with HEX at 5'-end and 6-FAM at the 3'-end (IDT). The DNA corresponding to the mutated repeat sequence of *B. halodurans* and also repeats belonging to different types were chemically synthesized along with the minus strand of T7 promoter (IDT). These were used as templates for the *in vitro* RNA synthesis using T7 polymerase. The RNA, thus synthesized was treated with DNase to remove the remains of template and pelleted by phenol-chloroform extraction method and resuspended in RNase-free water. It was further PAGE purified to maintain homogeneity. Fluorescein-5-thiosemicarbazide (FTSC) was used to label the repeat RNA at 3'-end, wherever required. The integrity of the synthesized RNA was checked in 15% (w/v) denaturing urea PAGE. The RNA was stored at -80°C until required.

2.2.3. Nuclease activity assays

All pre-crRNA processing reactions were performed at 37°C for 1hr. Time dependent studies were done at room temperature. 0.2 µM of pre-crRNA repeat labelled with HEX at 5'-end or 6-FAM at 3'-end was incubated with Cas5d (2 µM) in 20 mM Tris-HCl (pH 8), 100 mM KCl and 6 mM β-ME. RNase activity was also tested in the presence of 10 mM Mg^{2+}. Cleavage products were analyzed on 15% (w/v) denaturing urea PAGE. Similar assay was conducted for the constructs labelled with FTSC at 3'-end and unlabelled repeats. To visualize the processing pattern, the gel was stained with SYBR gold in cases where the unlabeled RNA was used.

2.2.4. Intrinsic tryptophan fluorescence assay

Intrinsic fluorescence emission spectrum of the protein was measured at 26°C by using a Fluoromax spectrofluorometer (HORIBA Jobin Yvon). To probe the tryptophan environment, the excitation wavelength used was 280 nm and emission was monitored from 285 nm to 500 nm. The concentration of protein used was 10 μM in 20 mM Tris-Cl (pH 8.0). The spectrum generated is an average of three scans after baseline correction. The slit width used for excitation was 1 nm and 9 nm for emission.

2.3. Results and Discussion

2.3.1. Investigating the RNase activity of Cas5d

2.3.1.1. Cas5d processes CRISPR repeat RNA

To test the nuclease activity, *cas5d* was cloned and purified and subsequently employed for activity assays (Figure 2.2). For activity assays, the substrate repeat RNA was chemically synthesized and labelled either at 5’ end with HEX or 3’ end with 6-FAM. Interestingly, when the 3’ 6-FAM and 5’ HEX labelled repeat RNA were incubated with Cas5d, we observed appearance of a band that was smaller in size as compared to the respective substrates, suggesting that Cas5d cleaves the repeat RNA in both the cases (Figure 2.3).

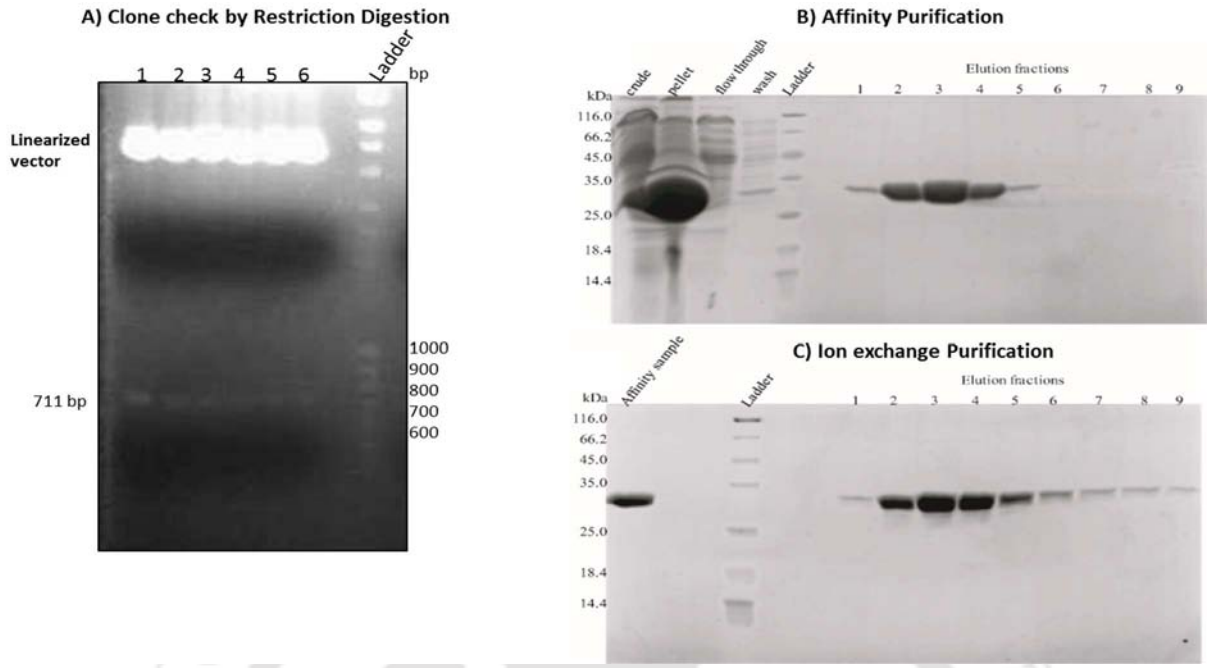


Figure 2.2 ***Cloning and purification of Cas5d from B. halodurans.*** (A) The restriction digestion showed that the insert in the vector corresponded to gene size (711 bp) thus confirming the clone. The clone was further verified by sequencing. (B-C) SDS-PAGE of Cas5d from affinity and ion exchange purification is shown.

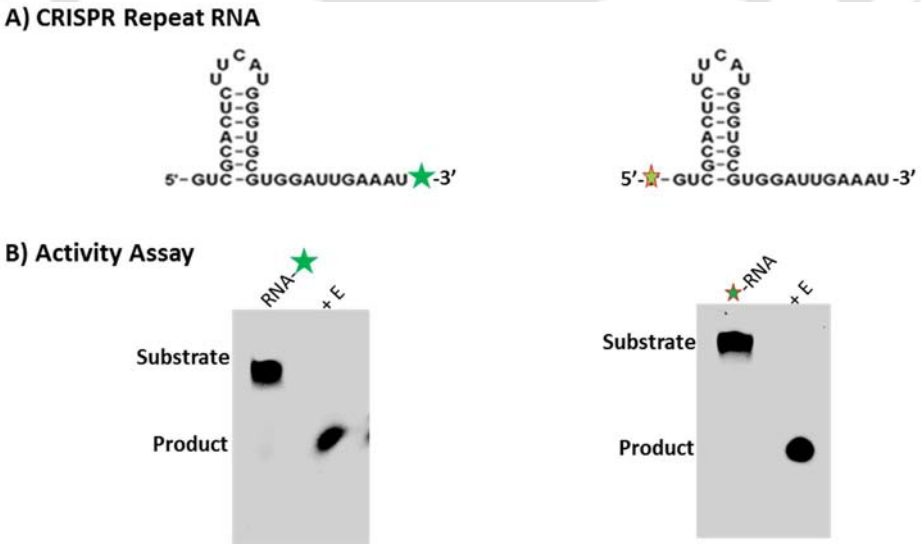


Figure 2.3 ***Cas5d in B. halodurans processes CRISPR repeat RNA.*** (A) The folding of the repeat RNA is shown and the position of the label at 3' and 5' is indicated with star. (B) Activity assay performed with 3' 6-FAM and 5' HEX labelled repeat RNA is shown. E represents Cas5d.

To assess the requirement of metal for nuclease activity, we performed the experiment in the presence and absence of metal. For this, we incubated 3' 6-FAM labelled repeat RNA

with Cas5d in presence of metal for 20 minutes, the resultant product seemed to be similar to the product in the absence of metal (Figure 2.4). Moreover, the addition of EDTA had no effect in the product formation (Figure 2.4). Thus, Cas5d seems to possess metal independent RNase activity.

Further, the experiments were conducted at two different time points to trace the trajectory of product formation. When we incubated the 3'-end labeled repeat RNA with Cas5d for longer time intervals, we observed differences in the position of the band relative to the T1 digest of the repeat (formed from the treatment with RNase T1, a single stranded G-specific nuclease), which in turn indicated the differences in the size of the product over time. The band that appeared within 20 minutes of incubation, migrated above the T1 digested product, while the band that was observed after 60 minutes had relatively lower size and migrated along with the T1 digested product (Figure 2.4).

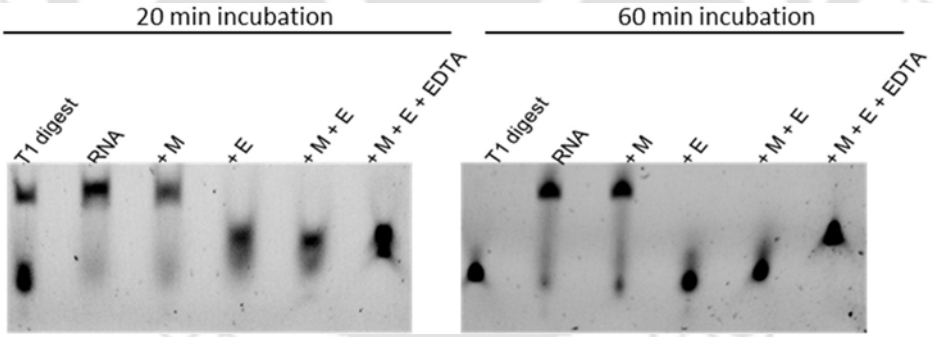


Figure 2.4 ***The differences in the product size of repeat processed over time.*** A shorter incubation of 20 min with repeat RNA produced a fragment larger than the T1 digest. A longer incubation of 60 min resulted in a fragment that was similar in size as that of the T1 digest. M represents metal, Mg^{2+} and E represents the presence of Cas5d. The presence of EDTA in respective lane is indicated.

This raises the possibility of the initial product getting converted to a smaller size product over time. To determine the extent of processing, we performed time dependent assays, where we incubated 3' 6-FAM labelled repeat RNA with Cas5d for different time intervals. The initial band appeared within 5 minutes of incubation, which migrated above the

T1 digested product. This with longer incubation got converted into the lower size band that migrated along the T1 digest (Figure 2.5). This led to the inference that the initial product was getting converted or further processed to a smaller size product and also prompted us to map the products due to extended processing (Figure 2.5).

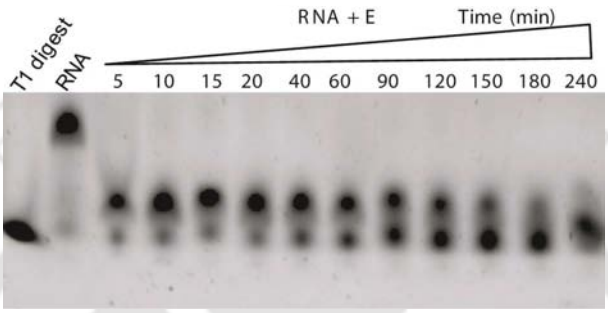


Figure 2.5 ***Time dependent RNase activity of Cas5d.*** The lane containing the T1 digest and RNA is indicated**.** E represents Cas5d. Incubation time period (5-240 min) is shown by a triangle. It may be noted that formation of smaller product occurs with time. For the assay, 0.2 µM of 3'-FAM labeled RNA was incubated with 2 µM of Cas5d.

2.3.1.2. Product mapping of Cas5d

After testing the RNase activity, we intended to map the cleavage site using RNase T1 and alkaline hydrolysis. RNase T1 is a single stranded G-specific nuclease that cleaves after every G in single stranded region. In the 3'-end of the repeat RNA, which appears to be single stranded, three potential cleavage sites, *viz.*, G23, G24 and G28 were noted for T1. Complete cleavage occurring at both G23/24 and G28 is expected to produce two fragments of about 4 nt each, of which G28 fragment is likely to be visible on the gel since it is fluorescently tagged at the 3'-end (Figure 2.6).

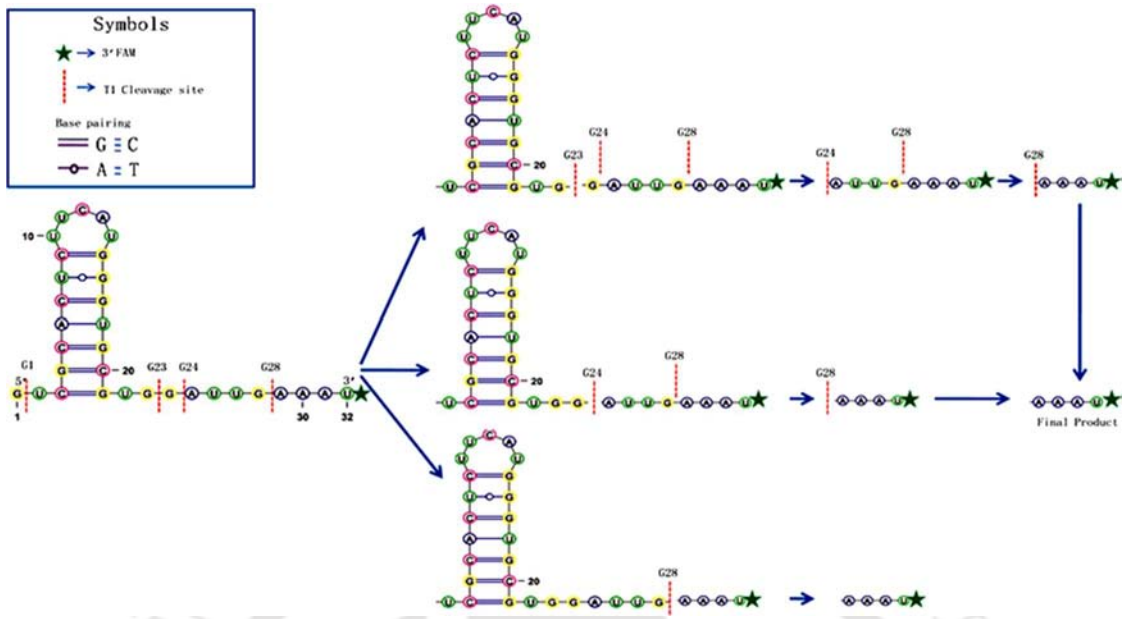


Figure 2.6 ***Schema of T1 digestion.*** The base Adenine is shown in blue, Guanine in yellow, Cytosine in purple and Uracil in green. The cleavage site is indicated by red dotted line. The fold was predicted using MFOLD (Zuker, 2003) and figure was prepared using VARNA (Darty et al., 2009). The potential T1 cleavage sites are indicated. Cleavage occurring at one or all the sites may produce products of different sizes. However, the fragment that is attached to the 3'-end labelled fluorescein alone will be visible on the gel. Complete T1 cleavage resulting from the subsequent processing of all the potential cleavage sites is expected to produce a 4 nt visible band on the gel.

We also used alkaline hydrolysis to generate ladder comprising of the partially hydrolyzed substrate fragments having single nucleotide difference (Figure 2.7). For this, we subjected the 3' 6-FAM labelled repeat RNA to alkaline hydrolysis to generate the OH ladder, schematically shown in Figure 2.7.

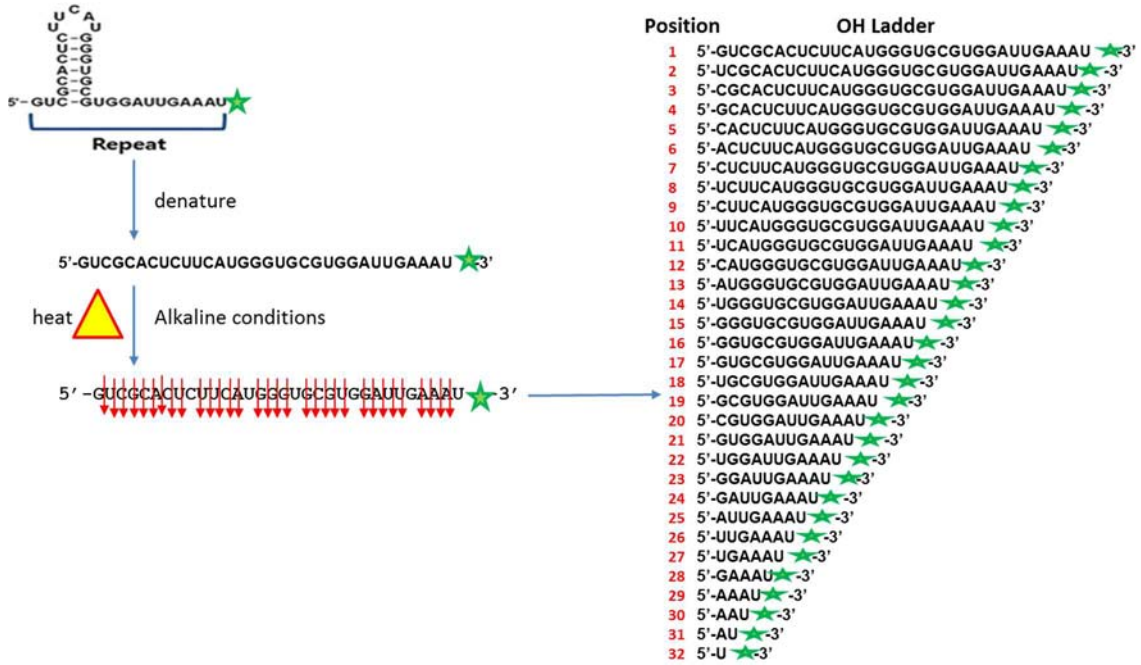


Figure 2.7 ***Schema of Alkaline hydrolysis.*** The 3' 6-FAM labelled repeat RNA was subject to partial alkaline hydrolysis to get single nucleotide resolution between fragments. The position of the label is indicated by green star. The position of the fragments obtained after hydrolysis is shown.

The above two strategies were employed to map the processing pattern. For this, we incubated the 3' 6-FAM labelled repeat RNA with Cas5d for different time intervals and electrophoresed the resulting products, along with the generated T1 digest and OH ladder for the repeat. Initially, a single band was observed which could have been resulted from the cleavage between G21 and U22 forming a fragment of 11 nt (Figure 2.8). A longer incubation ensued the appearance of a smaller size band via two intermediates (presumably by cleavage between G23 and G24, resulting in 9 nt fragment and other between G24 and A25, resulting in 8 nt fragment). The smaller size band was comparable to G28 fragment size from T1 digestion, therefore it can be inferred that Cas5d also cleaves at G28 producing a 4 nt fragment (Figure 2.8). Thus, Cas5d seems to processes the CRISPR repeat via intermediates, which further get converted into a final product of about 4 nt in size with time.

Under this circumstance, the crRNA will have a 4 nt extension of repeat sequences at the 5' end followed by the spacer element.

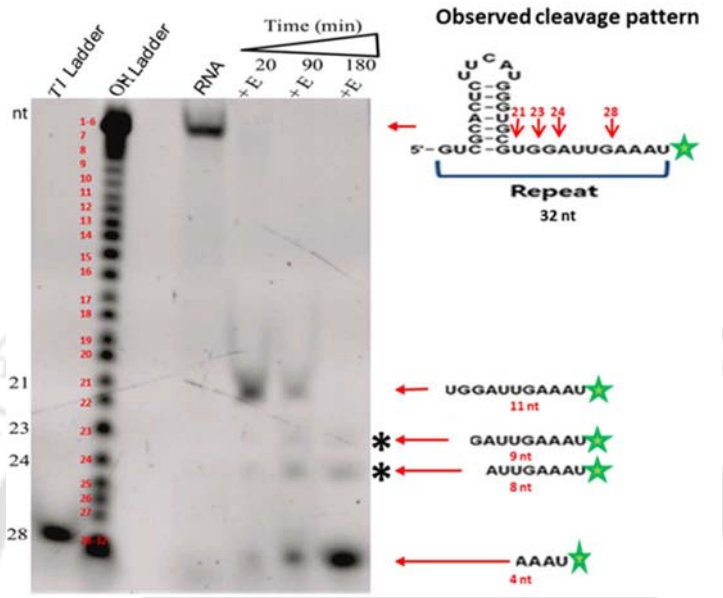


Figure 2.8 ***Mapping of the product formed from the repeat processing by Cas5d.*** Shorter incubation of 20 min with repeat RNA at room temperature produces a fragment of about 11 nucleotides but longer incubation of about 180 min produces a fragment of about 4 nucleotides that is similar in size as that of the T1 digest. * denotes the intermediate products.

Thus, Cas5d seems to process after G in the sequence, based on our product mapping of the processed repeat RNA. This raised the possibility of Cas5d being a single stranded G-specific nuclease, so to address this we designed two constructs with sequence mutated at the corresponding cleavage site. Construct1 had all G mutated to C in the single stranded region while the construct2 included C mutated to G in the loop of the repeat in addition to all G to C mutations in the single stranded region (Figure 2.9). The idea was to obtain a clear pattern of cleavage by converting all free G to C, thereby avoiding extended processing. It was anticipated that in first case, only a single band of 11 nt size should appear on gel as a result of cleavage after G21 of the 3' 6-FAM labelled RNA and no extended processing of repeat should be observed. In the second case, a partial digest was expected to show 2 fragments due

to cleavage at single stranded G12 in the loop region which is in addition to the band from the cleavage after G21 of the 3' 6-FAM labelled RNA as shown schematically in Figure 2.9.

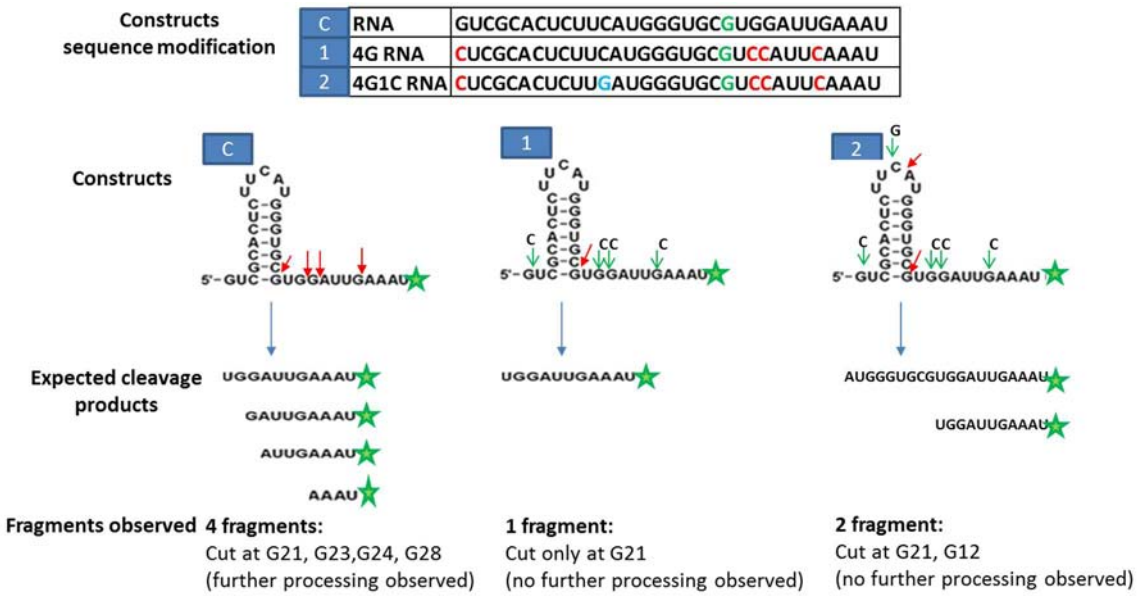


Figure 2.9 ***Effect of mutations of the substrate in the extended processing by Cas5d.*** The repeat RNA used as control is represented by C. While the construct1 is denoted by 1 which has all G mutated to C in the single stranded region and construct2 is denoted by 2 and has C12 mutated to G in the loop region in addition to G to C in single stranded region. In the sequence information of the constructs, G21, the first cleavage point is shown in green. The red and blue coloured bases indicate the mutation of accessible G to C and C to G respectively. In the repeat structure, the cleavage positions are marked by red arrow. Green arrows indicate the position of the mutated bases. The expected cleavage products from the respective constructs are shown.

The DNA constructs for the RNA mutants were synthesized and obtained from IDT and the corresponding RNA was synthesized *in vitro* followed by FTSC labelling at 3'-end (Figure 2.10). To observe the band corresponding to the construct RNA, the labelled RNA were loaded along with their serially diluted RNA preparations, obtained from *in vitro* synthesis.

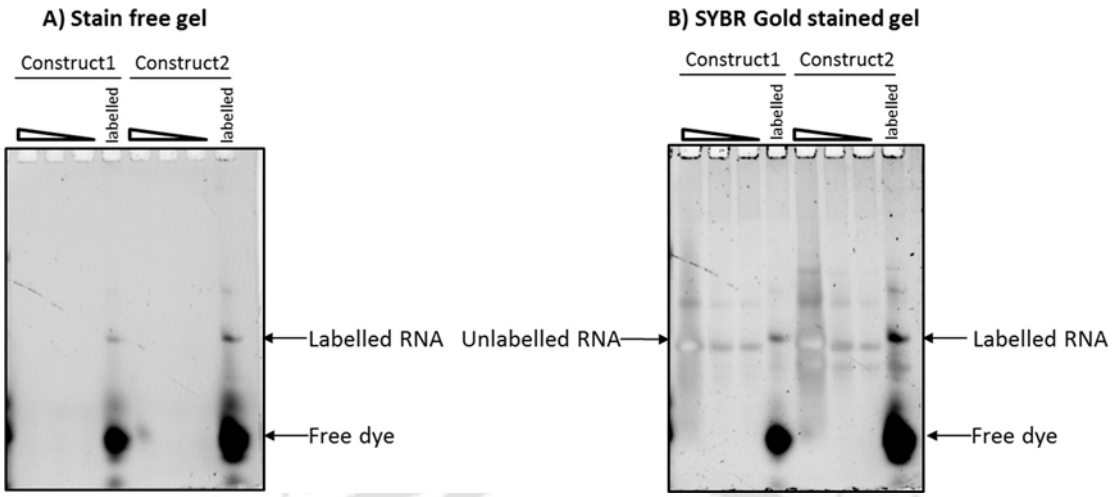


Figure 2.10 ***The in vitro synthesized RNA labelled with FTSC at 3'-end.*** The unlabelled RNA preparation for both the constructs are serially diluted to mark the position of the corresponding band which is shown by triangle with its tapering end representing the drop in concentration of RNA with increasing dilution. The lanes containing the labelled RNA is indicated. (A) The labelled RNA is visualized in a stain free gel (B) The gel in (A) is stained with SYBR gold to locate the position of corresponding unlabeled RNA.

When the 3' FTSC labelled mutant RNA were subjected to Cas5d cleavage, we observed the difference in the cleavage pattern from the control repeat RNA (as shown in Figure 2.11 with red arrows). But we were unable to map the cleavage products as the labelling efficiency was not sufficient to map the corresponding cleavage pattern. This resulted in low resolution of the labelled products. Moreover, the high intensity of the free/unbound dye interfered with the substrate signal.

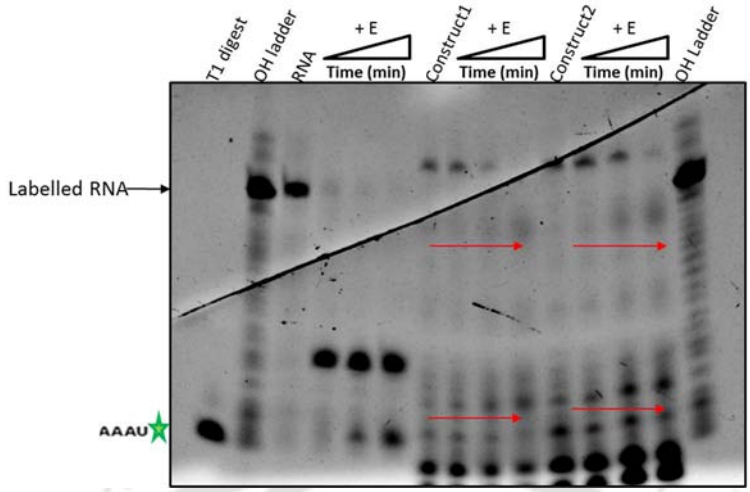


Figure 2.11 ***Cas5d activity assay with 3' FTSC labelled mutant RNA.*** The RNA mutants were incubated with Cas5d for 20 min, 90 min and 180 min at room temperature and the reaction products were run in 20% urea PAGE. T1 digest, OH ladder and the control RNA in the corresponding lanes are indicated. E represents Cas5d. The difference in the product pattern is shown by red arrows. The location of the fragment obtained from the complete T1 digestion of the RNA is shown with green star indicating the label position.

To troubleshoot this problem of signal interference with unbound dye in labelled RNA substrates, we tried to use membrane cut off filters to remove the unbound dye. Since the molecular weight of the free dye is 421.4 Da (FTSC) and the labelled RNA mutants are approximately 10.9 kDa, so we used Amicon Ultra 3000 Da centrifugal membrane filter to remove the undesired lower molecular weight contaminants. The labelled substrates were passed through this filter and eluted after appropriate washing (Figure 2.12). But even after filtration we were unable to get rid of the unbound dye. The plausible reason can be the unbound dye might associate or aggregate together at the membrane and thus could not be removed from the product.

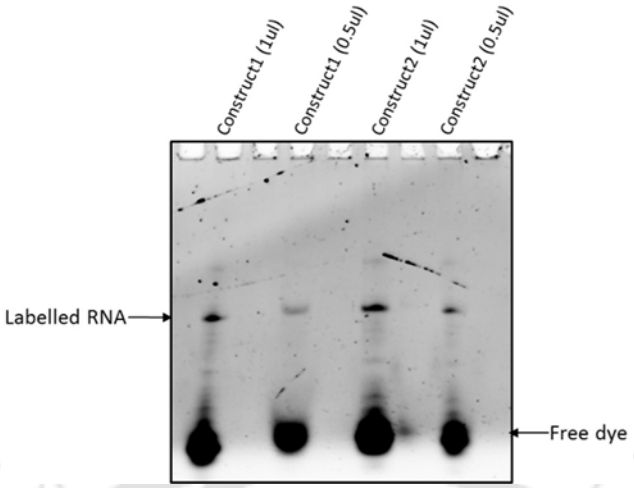


Figure 2.12 ***The integrity of the filter purified labelled RNA.*** The lanes containing the filter purified labelled RNA for the constructs are shown. The position of labelled product and the unbound dye is indicated by black arrow. The contamination with the unbound dye was still present even after filter purification. The intensity of the unbound dye was much high as compared to the signal of the labelled product and also the integrity of RNA seems to be compromised during filtration resulting in slight fragmentation of RNA.

Though we observed differences with the cleavage pattern using the mutant constructs, our attempt wasn't met with success, as the labelling efficiency was not sufficient to map the cleavage pattern. Therefore, we tried to address the specificity of the Cas5d RNase activity against repeats from other CRISPR subtypes.

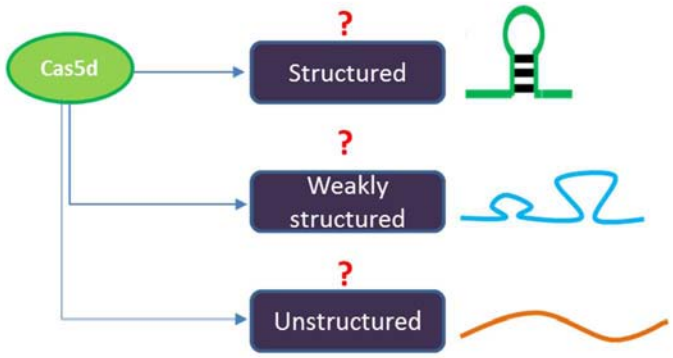
2.3.3. Investigating the Cas5d specificity

The CRISPR array in type I-C contains structured form of repeats, *i.e.,* the repeats attain a stem-loop architecture in pre-crRNA transcript, which are recognized and processed by Cas5d leading to specific product formation. This raises a question, whether Cas5d is able to recognize other structured repeats as well?

The repeats within the various types of CRISPR-Cas system can be structured, weakly structured or completely unstructured. Among the type I systems, the subtypes I-E, I-F and I-

C form structured repeats, while type I-A and type I-B are associated with unstructured and weakly structured repeats respectively. The repeats of type II and type III systems are also unstructured. This prompted us to test the activity of Cas5d with unstructured repeats as well. So, we chose both structured and unstructured repeat from various type I subtypes and also an unstructured repeat from type III system. The variability between the chosen repeats is shown in Figure 2.13.

A) Schematic representation of variability existing within CRISPR repeats



B) Differences in repeats among the CRISPR types

S.No.	Type	Organism	Repeat Size (nt)	Repeat Folding
1	Type I-A	*Aeropyrum pernix*	24	unstructured
2	Type I-B	*Haloarcula marismortui*	30	weakly structured
3	Type I-C	*Bacillus halodurans*	32	structured
4	Type I-E	*Escherichia coli*	29	structured
5	Type I-F	*Pseudomonas aeroginosa*	28	structured
6	Type III-A	*Mycobacterium tuberculosis*	36	unstructured

Figure 2.13 ***The repeat variability within CRISPR-Cas systems.*** (A) The schema showing the differences in the architecture of repeats existing within the CRISPR-Cas system, which was taken into account while selecting the substrates for Cas5d. (B) The differences in the repeats among the selected types are shown in tabular form. The selected repeats showed differences in size and folding.

The RNA of the selected repeats were subjected to fold prediction using MFOLD (Zuker, 2003) which predicts minimum free energy structures. The folding differences among the repeats were evident as some showed well-structured stem loop architecture, while others showed weakly structured stem. Moreover, differences were observed among the structured repeats in having different lengths of stem, the loop region and variable lengths at 3' and 5'-

end, *i.e.,* the stem-loop was either present towards 3'-end or 5'-end of the sequence (Figure 2.15A). Thus, the CRISPR repeats show differences in sequence, size and shape. The corresponding DNA for the selected repeats were synthesized from IDT, which were subsequently used for the *in vitro* RNA synthesis of the repeat. The *in vitro* synthesized repeat RNA were gel extracted followed by electro-elution to maintain homogeneity in the substrates (Figure 2.14). To test the activity, we incubated Cas5d with the repeat substrates belonging to type I-A, type I-B, type I-C, type I-E, type I-F and type III-A and analyzed the pattern in urea PAGE after staining with SYBR gold (Figure 2.15B).

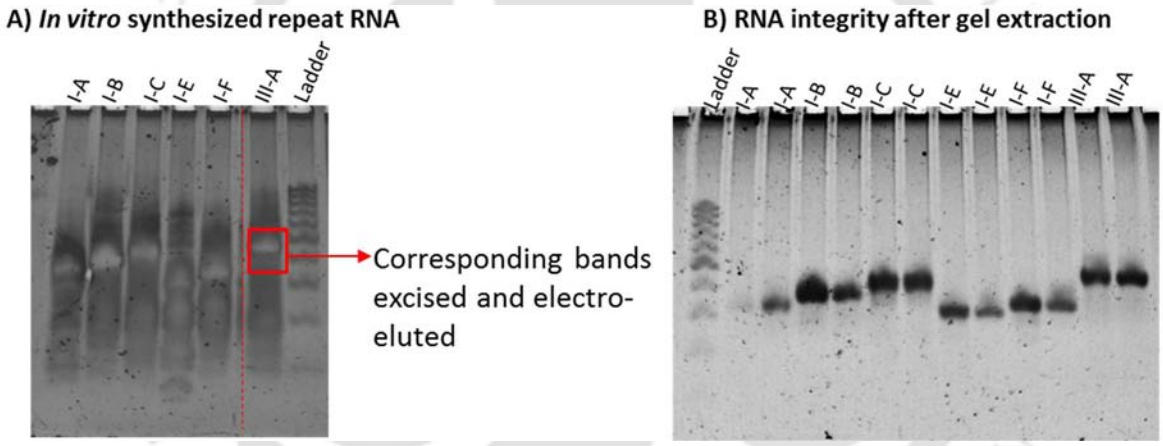


Figure 2.14 ***The in vitro synthesized CRISPR repeat RNA.*** (A) The *in vitro* synthesized repeat RNA belonging to various types are shown. The band corresponding to the repeat was excised and electro-eluted to obtain a homogenous substrate, as indicated by red box for a case. (B) The RNA integrity of the electro-eluted bands were checked in 15% urea PAGE after staining with SYBR Gold for visualization.

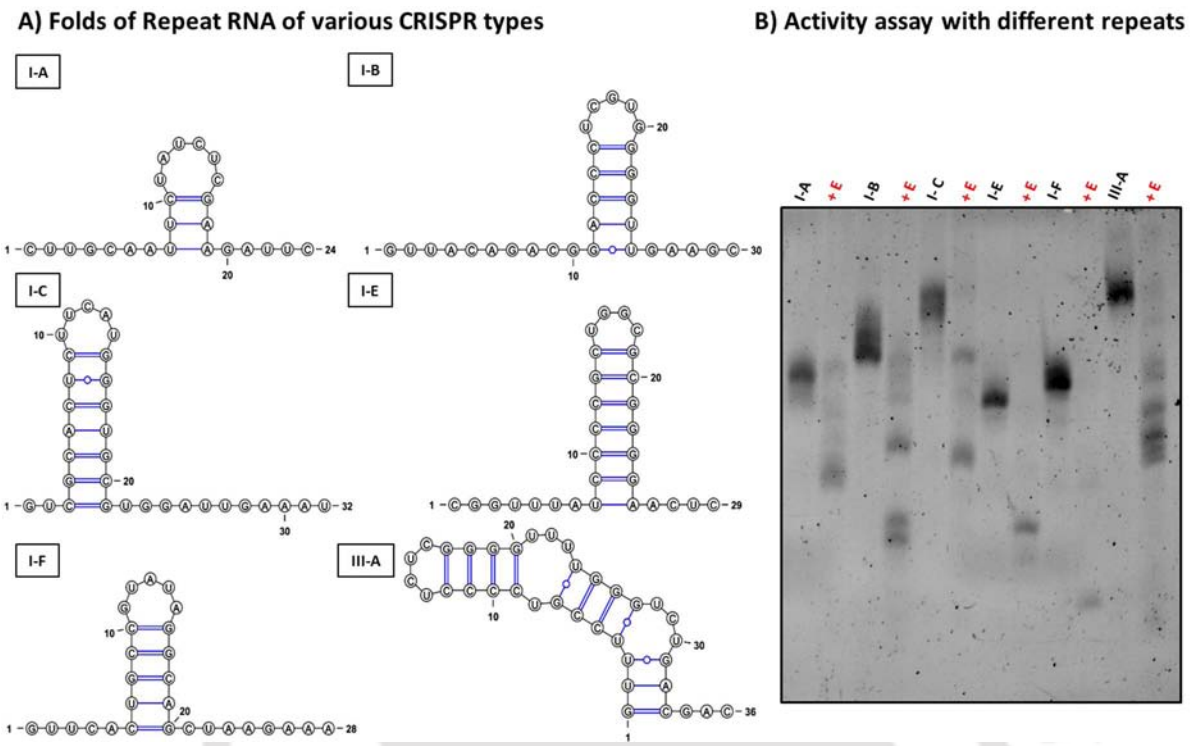


Figure 2.15 ***Cas5d activity against various repeats.*** (A) The folding pattern of CRISPR repeat RNAs used as substrates are shown. (B) Cas5d activity assay with different repeat RNA is shown. The lanes with particular repeat RNA is indicated by its respective type. E, represents the presence of Cas5d.

Surprisingly, we observed smaller sized bands for all the repeats in the presence of Cas5d, though the sizes of these bands differed among the repeats, which is suggestive of processing of all the repeat substrates by Cas5d. Thus, Cas5d was able to process CRISPR repeat within type I systems and also across types, as it cleaved CRISPR repeat RNA belonging to type III-A. It was interesting to see the repeats with varied sequence, size and folding pattern were still recognized and processed by Cas5d. It motivated us to explore the features in the sequence of the CRISPR repeats that might be responsible for their recognition by Cas5d. So, we aligned the repeat RNA sequences from various CRISPR types including the above repeats chosen for the activity assay and looked for the conservation of the residues (Figure 2.16). Interestingly, we observed that in different types of CRISPR repeat, the

residues that aligned corresponding to the position of cleavage site in type I-C showed a G-conservation. Thereby, explaining the processing of other repeats by Cas5d. Thus, Cas5d seems to be a G-specific nuclease.

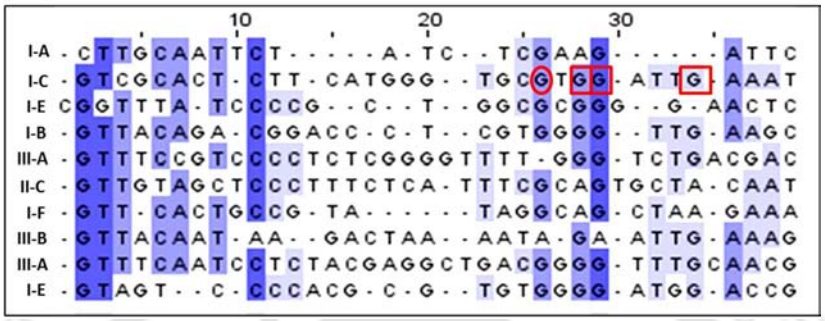


Figure 2.16 ***Multiple sequence alignment of the CRISPR repeat from different types.*** The identified cleavage site of Cas5d are shown. The initial cleavage site at G21 of type I-C repeat is shown in red circle and the extended processing site at G23, G24 and G28 are shown in red boxes. In the alignment of the repeats from various CRISPR types, the residues corresponding to the cleavage site in the type I-C repeat show a G-conservation.

2.3.4. Probing the active site residues of Cas5d

Cas5d repeat processing pattern revealed that it cleaved at the base of the stem and ensued extended processing in the 3' single stranded region. We inspected Cas5d structure for the residues that might be involved in processing. The crystal structure of Cas5d reveals a ferredoxin fold (also known as RRM), with structural features distinct from other known pre-crRNA processing factors (Nam et al., 2012).

2.3.4.1. Identification of the active site residues

Cas5d displays a single RRM N-terminal domain with a distinct C-terminal β-sheet extension. Inspection of the structure enabled us to identify the residues positioned in a

favorable conformation to conduct an acid base catalysis. Interestingly, we spotted two triads formed of Y46, K116 and H117 (henceforth referred as triad1) and Y35, K39 and H169 (henceforth referred as triad2). The triad2 was located geometrically in a conformation analogous to triad1 (Figure 2.17A). Remarkably, Cas5d also possessed two tryptophan residues located adjacent to these triads, with W47 located adjacent to the triad1 and W187 in proximity to triad2 (Figure 2.17A). Looking at the sequence conservation profiles of these residues across the type I-C organisms, it appears that the chemical nature of the residues seems to be preserved over their identity (Figure 2.17B). This drove us to hypothesize that the residues from these triads may be involved in catalysis. So, in order to test their involvement in Cas5d nuclease activity, we mutated these residues using site directed mutagenesis. The mutants were cloned and purified for the activity assays (Figure 2.17C).

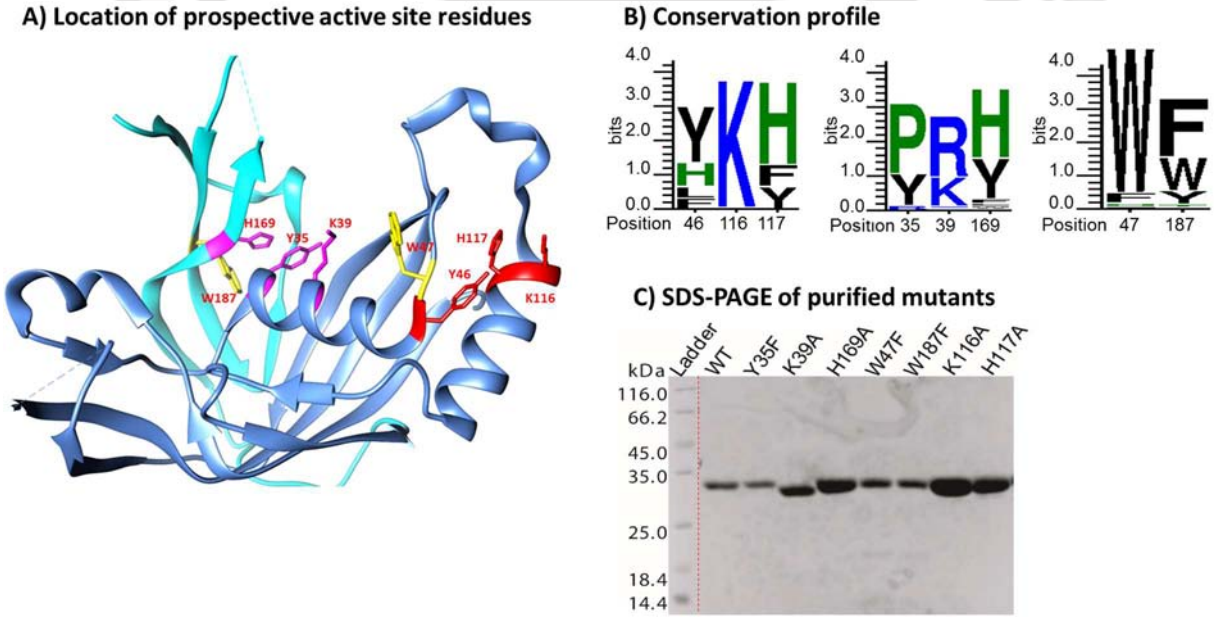


Figure 2.17 ***The prospective active site residues of Cas5d.*** (A) The crystal stucture of Cas5d, PDB ID: 4F3M is shown. The triad1 is shown in red and triad2 in magenta. The sequence number of the corresponding residue is indicated. The N-terminal is shown in blue and C-terminal in cyan. This figure was rendered using Chimera (Pettersen et al., 2004). (B) The sequence logo depicting the conservation of these residues across type I-C orthologs is shown. The residue number is indicated below the logo. The height of a residue (in bits) represents the extent of conservation. Positions 46, 116 and 117 comprise the residues of triad1 and 35, 39 and 169 positions correspond to triad2 and 47 and 187 represents position of the tryptophan residues which are in proximity to the respective triads. (C) The SDS-PAGE of the purified Cas5d (WT) and its mutants is shown.

2.3.4.2. Effect of mutations on Cas5d RNase activity

The generated point mutants comprising of Y46F, K116A, H117A, Y35F, K39A, H169A, W47F and W187F were investigated for their role in the cleavage of the repeat RNA. All pre-crRNA processing reactions were performed at 37°C for 1hr. The 3' 6-FAM labelled pre-crRNA repeat at 0.2 μM was incubated with the mutants of Cas5d (2 μM) in the reaction buffer and cleavage products were analyzed on 15% (w/v) denaturing urea gel. In case of triad1 residues, we found that K116A impaired the activity while H117A abrogated the activity completely but Y46F does not seem to affect the activity (Figure 2.18B). Among the triad2 residues, H169A was as active as wild type Cas5d while Y35F and K39A showed accumulation of the larger fragment, *i.e.,* impacted the activity slightly so that product conversion was slowed down (Figure 2.18C). The mutants of the two tryptophan residues - W47F and W187F produced the smaller fragment as that of the wild type, therefore did not seem to effect the activity (Figure 2.18C).

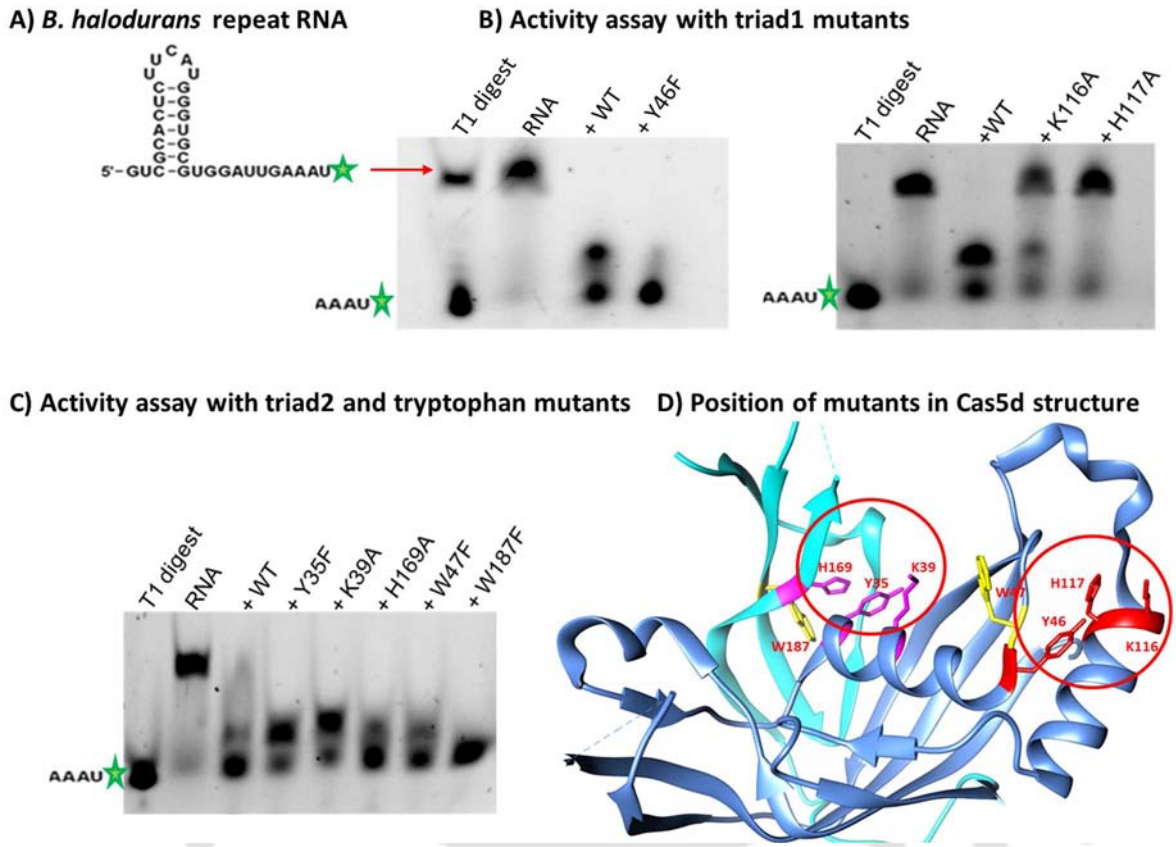


Figure 2.18 ***Effect of mutations on Cas5d nuclease activity.*** WT represents the wild type Cas5d and the respective mutants are shown at the top. The lane that has substrate as control is denoted as RNA. The lane having T1 digest of RNA is shown. (A) The RNA is 3'-end labeled with 6-FAM indicated with green star. (B) Assay to monitor the effect of the point mutants Y46F, K116A and H117A in RNase activity is presented. (C) Assay to monitor the effect of the point mutants Y35F, K39A, H169A, W47F and W187F is shown. (D) The structure of Cas5d (PDB ID: 4F3M) displaying the position of residues that are proposed to be involved in substrate binding and/or nuclease activity is shown. The residue number is indicated. The N-terminal is shown in blue and C-terminal in cyan. This figure was rendered using Chimera (Pettersen et al., 2004).

Among the point mutants tested, H117A mutant abrogated RNase activity of Cas5d while others impaired the activity to various extents. Therefore, we tried to understand the involvement of H117 residue in Cas5d activity. The impaired activity can be a result of direct involvement of H117 in RNA catalysis or due to the involvement in effective binding and positioning of the substrate for the catalysis. To address this, we performed electrophoretic mobility shift assay of H117A with repeat RNA (Figure 2.19). For the reaction, 1 µM of 5'

HEX and 3' 6-FAM labeled repeat RNA was titrated with 0.1-50 µM of H117A at room temperature for 20 minutes. The reaction mixture was loaded onto 12% Native PAGE.

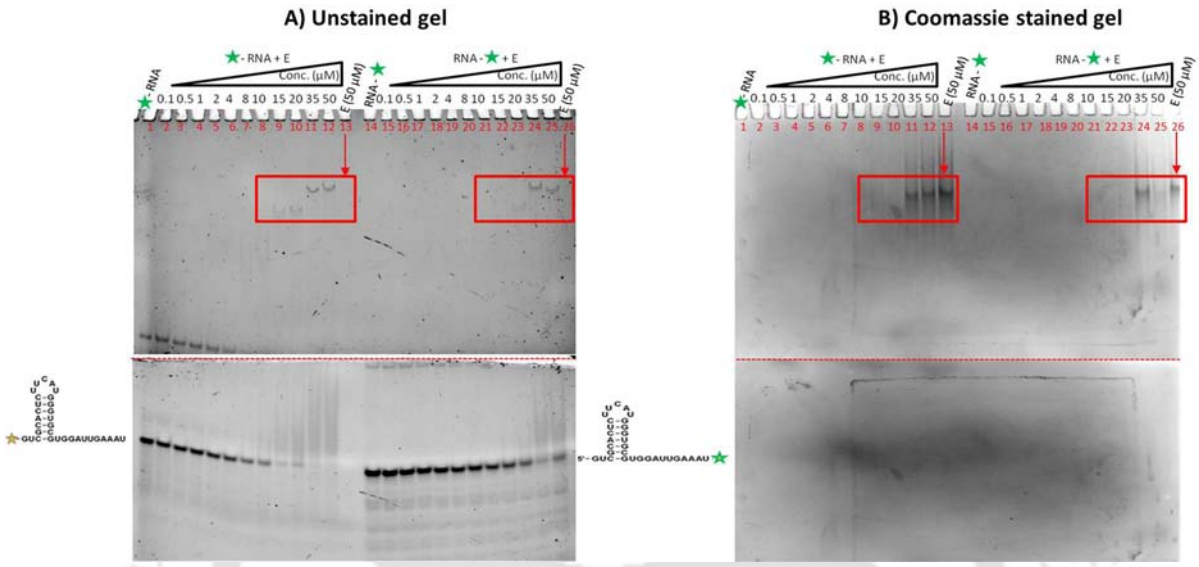


Figure 2.19 ***EMSA with H117A mutant of Cas5d.*** The experiment performed with 5' HEX (lanes 1-12) and 3' 6-FAM (lanes 14-25) labelled RNA is shown. The label position is shown by green star. E represents H117A. Triangle shows increasing concentration of H117A. Lanes 13 and 25 contain H117A loaded alone in 50 µM, to locate the position of the H117A. (A) The unstained gel showing the shift in the position of the labelled substrate with the increasing concentration of the protein. (B) The gel in (A) is stained with Coomassie to locate the position of mutant protein indicated by red arrows.

We observed a clear shift in the positions of the both 5' HEX (loaded in lanes 1-12) and 3' 6-FAM labelled RNA (loaded in lanes 14-25) with increasing concentration of protein under native conditions, which might be due to the H117A binding to RNA. To confirm this the same gel was stained with Coomassie, which showed the presence of protein corresponding to the shifted positions of the labelled RNA in the unstained gel. Moreover, the protein without any substrate (control) was at the same position as that of the bound protein. This confirmed that H117A binds to the substrate. Therefore, the abolishment of RNase activity of H117A could be a result of involvement of H117 residue in catalysis of the substrate. Our work together with the demonstration by Nam et al. (2012) suggests that K116 and H117 participate in hydrolyzing the RNA. Y46, K116 and H117 seem to be attractive candidates to assume the analogous role

as proposed for the equivalent residues in Cas6 (Y31: K52: H46) and tRNA intron splicing endonucleases (Y246: K287: H257) (Calvin and Li, 2008; Carte et al., 2008). The archetypal enzyme RNaseA too exhibits similar triad (H12:K41:H119) in catalysis wherein the Tyr is replaced by His12 (Raines, 1998). This reinforces the notion that the structurally unrelated enzyme may exhibit similar catalytic mechanism by means of convergent evolution (Galperin and Koonin, 2012). Subscribing to this view, it is possible to propose that Y46 in Cas5d may play a role of a base deprotonating the 2'-OH of G21 for inline nucleophilic attack on the scissile phosphate. K116 is a likely candidate to stabilize the negatively charged transition state and H117 may protonate the leaving group akin to K41 and H119, respectively, in RNaseA (Raines, 1998).

2.3.4.3. Effect of metal on RNase activity of Cas5d mutants

Earlier, as we have observed that Cas5d repeat processing was independent of the presence of metal, we indented to test whether the same holds for the point mutants. To address this, we employed two differently end labelled RNA substrates to monitor the effect of metal on RNase activity of the Cas5d point mutants. When we incubated 5' HEX and 3' 6-FAM labelled repeat RNA with Cas5d mutants in presence of metal, we found that the mutant H169A showed drastic reduction in activity, while the other mutants showed no significant alteration in the cleavage pattern, although the product conversion seems to be retarded in presence of metal (Figure 2.20). This suggests the possibility of a metal binding site nearby these residues, as the metal binding could induce localized conformational change, thereby affecting their activity. Since Cas5d does not require metal cofactors to exhibit the RNase activity, it raises question on the utility of binding of metal to Cas5d. We have tried to address this interesting aspect in Chapter 4.

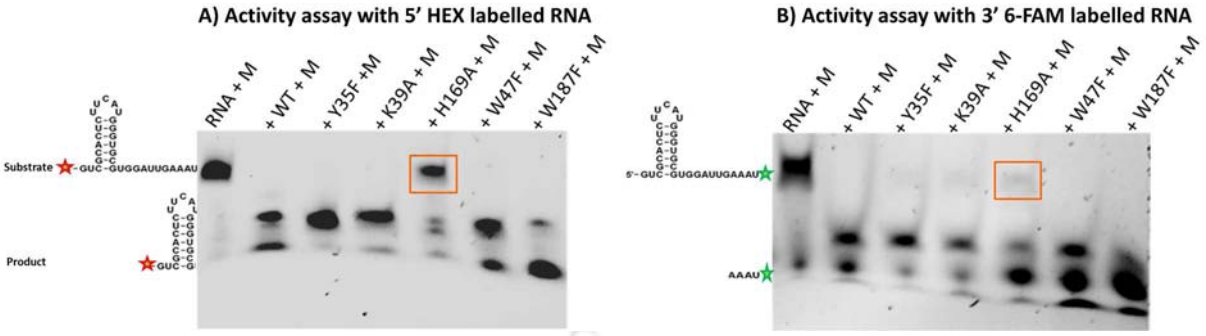


Figure 2.20 ***Effect of metal on RNase activity of Cas5d mutants.*** WT represents the wild type Cas5d and lanes containing the respective mutants are indicated. The lane that has labelled substrate as control is denoted as RNA. The presence of metal is indicated by M, Mg^{2+}. (A) The activity assay with the 5' HEX RNA substrate is shown. The red star indicates the position of label in the RNA. (B) The activity assay with the 3' 6-FAM RNA substrate is shown. The green star indicates the position of label in the RNA. The presence of metal seems to retard the product conversion in Y35F and K39A mutants while H169A mutant showed drastic reduction in the activity.

2.3.5. Probing the nature of active site using intrinsic tryptophan fluorescence

We further tried to probe the active site using the two tryptophans that were present in the vicinity of the active triads of Cas5d (Figure 2.21A). Both the tryptophans (W47 and W187) had varied extent of exposure. W47 was exposed to solvent (relative surface accessibility = 39.9%) and W187 was largely buried (relative surface accessibility = 10.2 %). For probing the nature of active site, we monitored the effect on intrinsic tryptophan fluorescence of the Cas5d in presence of RNA and in presence of acrylamide, using 280 nm excitation wavelength and collecting the emission scan from 285 nm to 500 nm. The addition of increasing amounts of RNA to Cas5d resulted in quenching of tryptophan fluorescence suggesting that RNA is binding to Cas5d, resulting in exposure of one of the tryptophans to the solvent or the RNA binding site is situated closer to it (Figure 2.21B). The buried tryptophan (perhaps W187) getting exposed to the solvent is possible if it undergoes conformational changes upon RNA binding. In other words, it may be inferred that RNA

binding may induce a localized conformational transitions in Cas5d structure around W187 that makes it accessible to solvent. This is evidenced by the Stern-Volmer plot of F_0/F vs. RNA concentration, where F_0 is the fluorescence intensity without RNA and F is the fluorescence intensity at a particular concentration of RNA. The quenching pattern was linear upto 40 μM of RNA and then a steep increase was observed which remains constant above 60 μM (Figure 2.21B). The initial linear quenching was perhaps due to the accessibility of W47 to the RNA and the steep increase in the quenching may be attributed to the exposure of W187 (which is largely buried) to RNA at higher concentration.

In the case of probing with acrylamide, which is a neutral quencher, the Cas5d fluorescence was quenched with increasing concentration of acrylamide. The quenching pattern was analyzed using the Stern-Volmer plot of F_0/F vs. acrylamide concentration, where F_0 is the fluorescence intensity without the acrylamide and F is the fluorescence intensity at a particular concentration of acrylamide. Here too, the pattern of quenching was similar to RNA and hence this may be attributed to the varied exposure of the W47 and W187 to the acrylamide. However, it may be noted that compared to RNA, higher concentration of acrylamide is required to access W187 (Figure 2.21C).

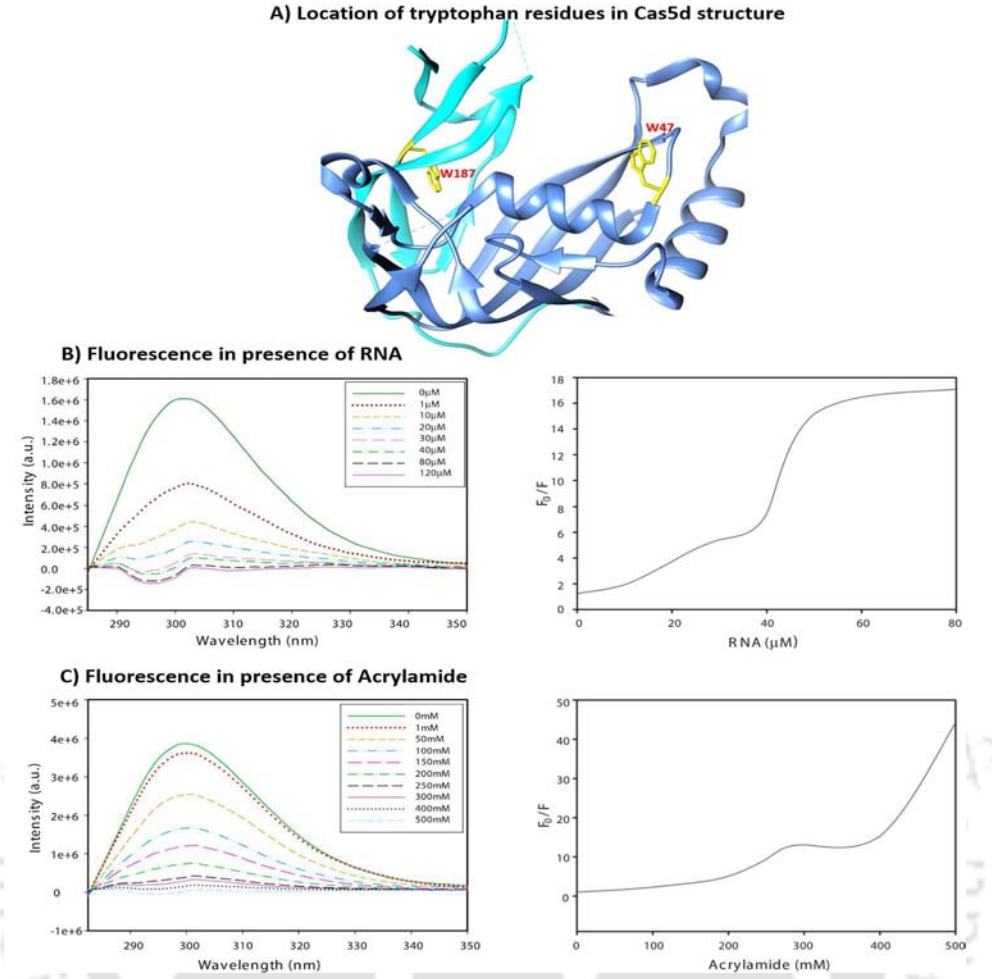


Figure 2.21 ***Fluorescence studies to probe the nature of active site.*** Intensity is shown in arbitrary units (a.u.). (A) The two tryptophans positions are shown in Cas5d structure (PDB ID: 4F3M). The corresponding residue number is indicated. The N-terminal is shown in blue and C-terminal in cyan. This figure was rendered using Chimera (Pettersen et al., 2004). (B) The quenching of tryptophan fluorescence is shown with increasing concentration of RNA (μM), along with the plot of F_0/F vs. RNA concentration, where F_0 is the fluorescence intensity without the RNA and F is the fluorescence intensity at a particular concentration of RNA. (C) The quenching of tryptophan fluorescence with increasing concentration of acrylamide (mM) is shown along with the plot of F_0/F vs. acrylamide concentration, where F_0 is the fluorescence intensity without the acrylamide and F is the fluorescence intensity at a particular concentration of acrylamide.

2.3.6. Impact of transition state inhibitors on Cas5d activity

After identifying the residues involved in Cas5d nuclease activity, we were interested in understanding how these residues were positioned during catalysis to bring about the effective processing. This required the information of the transition state. Typically, enzymes

interact with substrates by introducing strain or distortions, moving the substrates towards the transition state. There are transition state analogs that resemble the transition state of a substrate molecule in an enzyme-catalyzed chemical reaction. These enzyme inhibitors, which resemble the transition state structure, bind more tightly to the enzymes than the actual substrates. Vanadium compounds are good examples of transition state inhibitors for RNases (Crans et al., 2004). So, in order to trap the endonuclease Cas5d in its transition state, we tried the vanadium compounds and studied their effect on activity. 0.2 µM of 3' 6-FAM labelled RNA was incubated with 2 µM of Cas5d for 2 hours at 37ºC in presence of the varying concentration of uridine metavanadate and sodium orthovanadate inhibitors. In case of incubation of Cas5d with end-labelled RNA in presence of uridine metavanadate, we observed increase in intensity of the band at the position similar to the band of control substrate, with the increasing concentration of uridine metavanadate and the smaller sized band that was present at lower concentration of uridine metavanadate disappeared at higher concentrations, suggesting inhibition of Cas5d activity (Figure 2.22A). This effect was more pronounced at 20 mM concentration of uridine metavanadate. In case of incubation of Cas5d with end-labelled RNA in presence of sodium orthovanadate, a similar effect was observed but the smaller sized band did not disappear even at high concentrations of about 155 mM, though showed decrease in intensity (Figure 2.22B). This suggests that sodium orthovanadate is not able to inhibit Cas5d activity completely and still higher concentration of it may be required to observe the complete inhibition. To get further insight, we attempted to solve the structure of Cas5d trapped in the transition state by crystallizing it in presence of transition state inhibitor. For this, we used ammonium sulphate grid to set drops of Cas5d along with sodium orthovanadate using hanging drop vapour diffusion method (Figure 2.22C). Small sized crystals were obtained initially. But despite several attempts to optimize the conditions, we couldn't succeed in obtaining the bigger crystals suitable for diffraction.

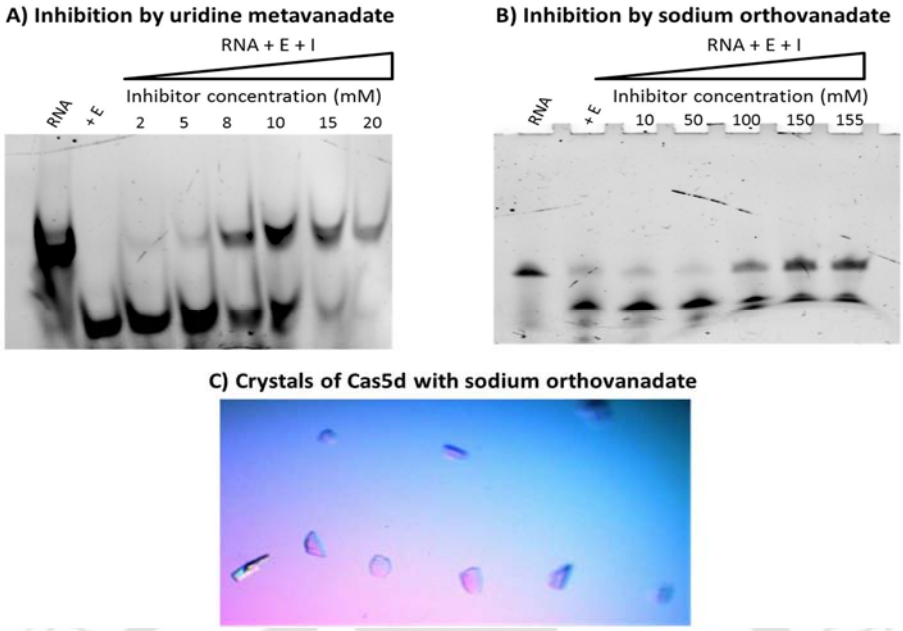


Figure 2.22 ***RNase activity inhibition by vanadate.*** (A) Activity inhibition of Cas5d by uridine metavanadate is shown. (B) Activity inhibition of Cas5d by sodium orthovanadate is shown. I represent the inhibitor. The lanes containing the control RNA is indicted. E represents the presence of Cas5d. The increasing concentration of inhibitor is indicated by triangle. (C) Crystals of Cas5d with sodium orthovanadate are shown.

2.4. Summary

We identified Cas5d as a metal independent endoRNase that processes the CRIPSR repeat RNA. Further, Cas5d promotes extended processing of the repeat in the single stranded region at 3'-end over time, as revealed by our *in vitro* experiments. This extended processing would add 4 nt 5' overhang of repeat sequence to the spacer sequence in the crRNA. We also found that Cas5d nuclease activity is skewed towards non-specificity *in vitro*, as it cleaves various CRISPR repeats irrespective of their shape, size and sequence but the activity seems to be G-specific in nature. We identified two active triads in Cas5d, formed by Y46, K116 and H117 and Y35, K39 and H169 residues that seem to participate in RNA hydrolysis. Thus, we found Cas5d to be an active enzyme in type I-C system which is in contrast to Cas5 in other type I subtypes that are reported to play only the structural role.

3.1. Introduction

It is known that RNA folds as it is being transcribed (Brehm and Cech, 1983; Kramer and Mills, 1981; Lai et al., 2013; Meyer and Miklos, 2004), which is often termed as co-transcriptional folding. RNA transcription is directional, *i.e.*, 5' to 3'-end, due to which the co-transcriptional folding whether *in vivo* or *in vitro* tends to proceed sequentially (Mahen et al., 2005; Mahen et al., 2010). As a result, the base pairs at the 5'-end of the RNA can form first, whereas base pairs involving the 3'-end can only form once transcription is nearing completion. This directionality of transcription influences both the folding pathway and the formation of functional secondary structure of the RNA molecule (Meyer and Miklos, 2004). The intermediate or transient structures formed during transcription may not necessarily be the same as the final structure of the sequence, but they help in the formation of functional secondary structure by suppressing the alternative competing structures. Therefore, it can be said that the folding of a nascent sequence emerging from the RNA polymerase (RNAP) is affected by the fold adopted by the regions that were formed earlier. Thus, they favor certain pathways in the energy landscape over others and guide the folding towards the desired functional secondary structure. The co-transcriptional folding pathways adopted may not necessarily be unique (Jackson et al., 2006), and factors such as transcription speed, flanking sequences (Koduvayur and Woodson, 2004) and tertiary interactions (Chauhan and Woodson, 2008) can influence which pathway dominates, resulting in functional structure.

Studies have shown clear differences between co-transcriptional and post-transcriptional folding of RNA (Lutz et al., 2014; Treiber and Williamson, 2001; Woodson, 2002). During post-transcriptional folding, the lack of constraints might result in a population of non-native like or non-functional structures, which might affect their recognition by other molecules to perform the required function. In other words, it might be recognized differently

from the native co-transcriptionally folded structure (Artsimovitch et al., 2000). All these results suggest that co-transcriptional structure formation play an important role in the correct folding of RNA sequences resulting in functional RNA structure, which in some cases might be recognized as substrates to RNases for being processed to perform varied functions. When misfolded or non-native structures are formed due to the intermediates getting stuck in the kinetic trap, the canonical recognition sites for the RNases can't be formed. It appears that this can be circumvented by the co-transcriptional processing, *i.e.*, coupling of transcription with the processing of the transcript. The co-transcriptional processing is shown to occur in nature in a number of different organisms, which has increased the understanding of how processing and transcription can be intricately intertwined resulting in specific product formation (Khodor et al., 2011; Kornblihtt et al., 2004; Merkhofer et al., 2014; Proudfoot, 2000; Proudfoot et al., 2002).

Abiding by this view, we hypothesized that the CRISPR repeats in CRISPR array might adopt different structures than the conventional stem-loop structure of a stand-alone repeat sequence, if they undergo post-transcriptional folding, which in turn may influence the specificity of RNA processing by RNases. In this chapter, we have attempted to test this hypothesis.

3.1.1. Processing of CRISPR repeat

This study tries to explore the mode of CRISPR RNA processing inside the cell. As discussed in chapter 2, we had observed extended processing of repeat *in vitro*, which resulted in variable length products. Moreover, we found Cas5d activity skewed towards non-specificity, as it cleaved repeats of other CRISPR types as well, which had different

sequences and folds. Thus, the formation of crRNA with homogenous size inside the cell is possible if the CRIPSR RNA processing is coupled with its synthesis thereby releasing individual guide RNA units as and when it is formed. This coupling of transcription and processing is required because repeats can be processed as soon as they are formed. Thus, all repeats will have similar structure which will be subjected to processing resulting in uniform processing pattern. If the entire array is to be transcribed first and then processed, the repeats may possess different structures. Additionally, the complex folding pattern of the array might hinder the accessibility of Cas5d to uniformly cleave all the repeats, thus generating a varied cleavage pattern. So, in order to understand the basis of specificity inside the cell we proceeded to explore the CRISPR RNA processing *in vivo*. This will also address the question of whether the extended processing of repeats takes place *in vivo*, similar to that observed *in vitro*. The two possible modes of processing are discussed below.

3.1.1.1. Post-transcriptional processing

In case of post-transcriptional processing, the CRISPR array is transcribed completely forming the pre-crRNA that attains its final fold without any constraint, which is then subjected to processing by Cas5d. But the complexity in folding of the array increases with the increasing number of repeat-spacer units, which might result in varied folding of the repeats in the array. To understand this effect, we subjected various stretches of the CRISPR array 4 (366106-368458) of *B. halodurans* to fold prediction using MFOLD (Zuker, 2003) and analyzed the structures having the minimum free energy (Figure 3.1). As anticipated, we found that the fold of a single standalone repeat was not the same, when it existed as part of array and also the effect of the surrounding sequence was clearly evident (Figure 3.1).

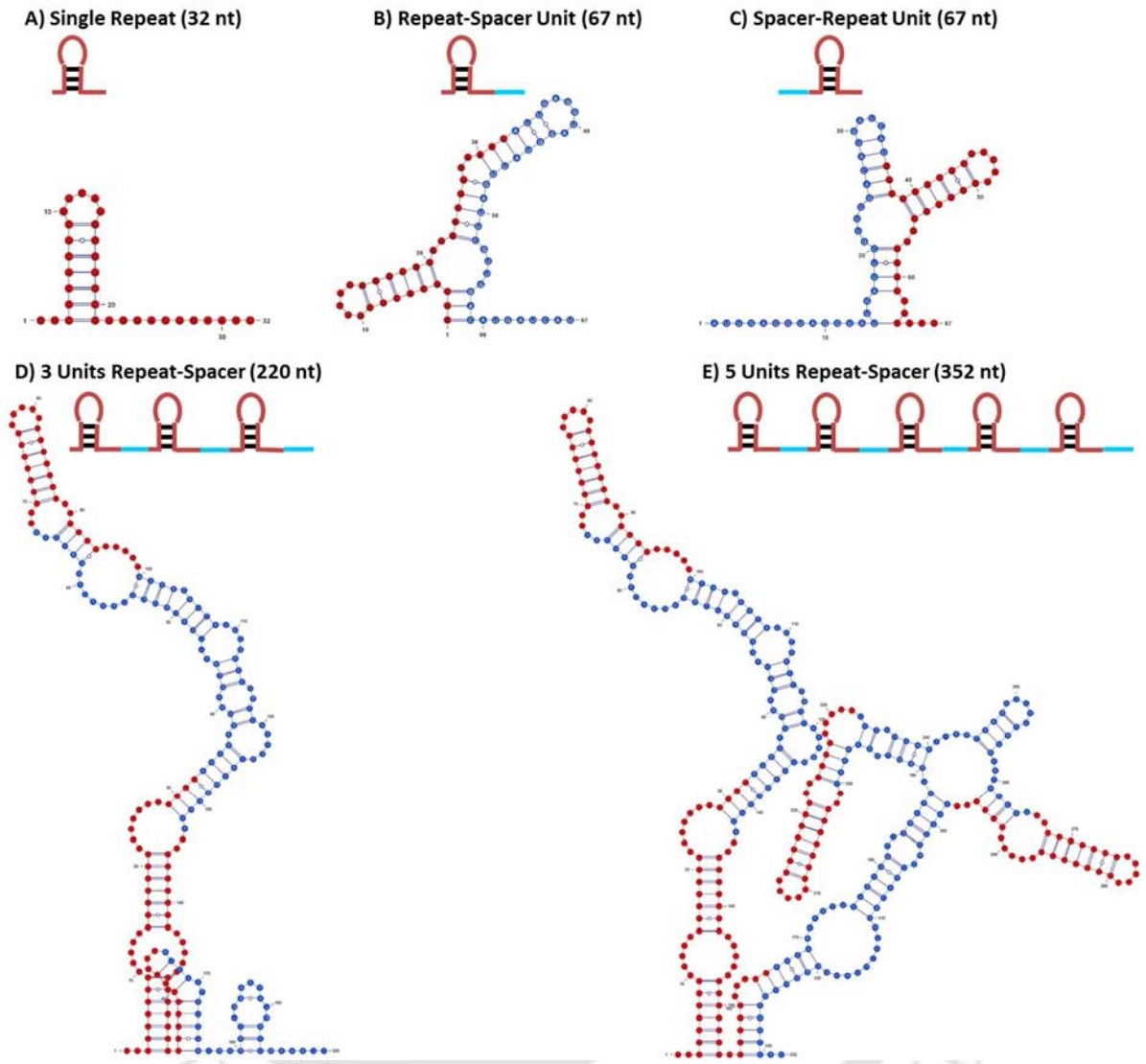


Figure 3.1 ***Folding of various regions of CRISPR array.*** The repeat and spacer units are colour coded. Repeat is shown in red and spacer in blue. Folding of (A) single repeat, (B) single repeat-spacer unit with spacer succeeding the repeat, (C) single repeat-spacer unit with spacer preceding the repeat, (D) three repeat-spacer units and (E) five repeat-spacer units are depicted.

Based on the differences observed in the folding of these constructs it can be speculated that the entire CRISPR array will have more complex folding due to which all the repeats may not be uniformly accessible to Cas5d, which in turn will result in crRNA of varied length (Figure 3.2A). Therefore, it seems that post-transcriptional processing of pre-CRISPR RNA is likely to produce heterogeneity in the maturation of crRNA. Hence we tried to look into another possibility where processing could be coupled with transcription.

3.1.1.2. Co-transcriptional processing

In case of co-transcriptional processing, the CRISPR array will be processed simultaneously during transcription. Since the transcription proceeds in 5' to 3' direction, each repeat-spacer unit is sequentially synthesized and adopts the native structure without the influence of other regions. Therefore, once a repeat emerges from RNAP, it is likely to adopt the cognate stem-loop structure that is recognized by Cas5d. This in turn leads to consistent cleavage pattern of CRISPR repeat by Cas5d producing the mature crRNA of uniform length (Figure 3.2B).

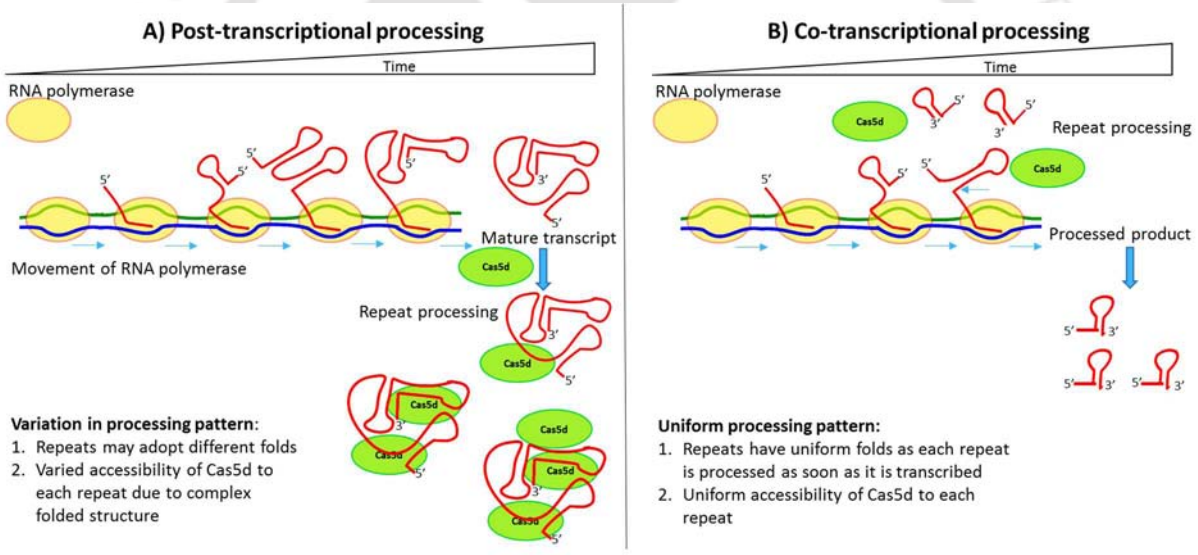


Figure 3.2 ***The differences between post-transcriptional and co-transcriptional processing.*** (A) During post-transcriptional processing, it is likely that each of the repeats takes up different structure due to the influence of the neighbouring region that imposes structural constraints. This in turn is expected to hinder Cas5d to access the respective binding and cleavage site on the repeats leading to variation in the size of the matured crRNA. (B) During co-transcriptional processing, sequential emergence of the RNA region facilitates the adoption of consistent repeat structure, which allows Cas5d to homogenously access the respective binding and cleavage region in each repeat leading to the formation of matured crRNA without any variation in size.

3.1.2. Schema to analyze repeat processing *in vivo*

To explore the mode of repeat processing *in vivo*, we designed constructs containing stretches of CRISPR array of varied repeat-spacer unit composition and co-expressed with Cas5d. The RNA processed inside the cell is then isolated and analyzed for the processing pattern by monitoring the extension of a labelled probe after the reverse transcription reaction (Figure 3.3).

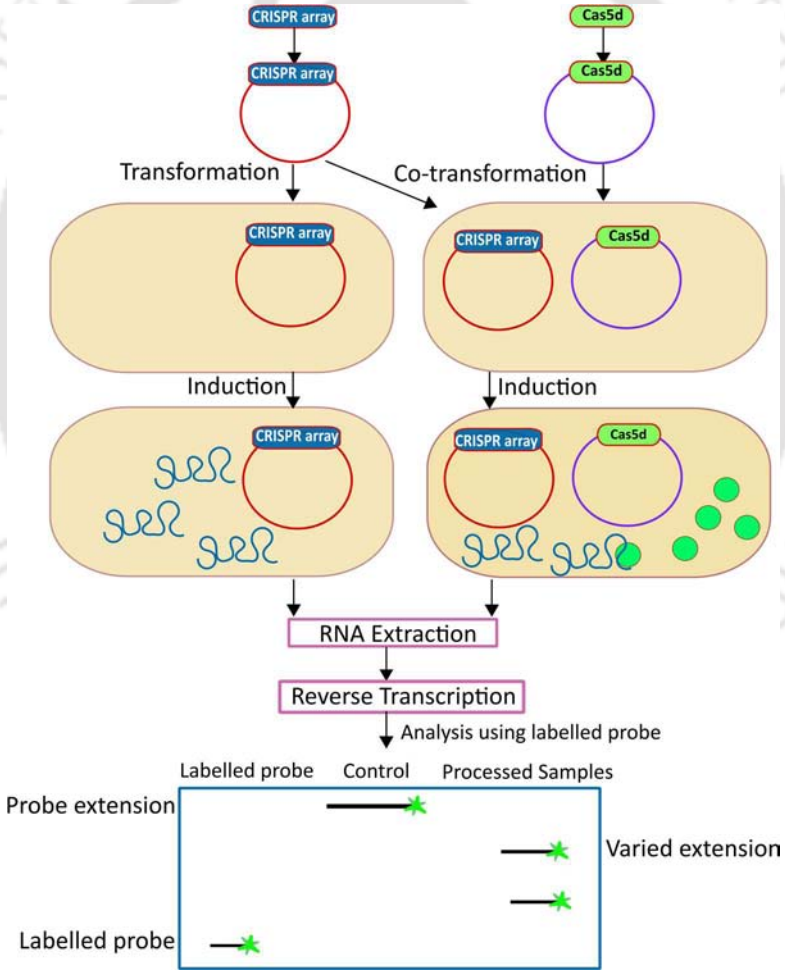


Figure 3.3 ***Schema to analyze in vivo processed repeat.*** Plasmids containing the CRISPR array constructs and Cas5d are shown in red and purple respectively. The CRISPR array transcript is shown in blue and green circles denote Cas5d. The label in the probe is indicated by green star. The lanes containing the labelled probe, control (unprocessed RNA) and processed sample are indicated.

For the unprocessed array, the probe extension will be complete and will correspond to the entire length of the construct RNA, while the processed array will show partial extensions of the probe, till the point of cleavage. The size of the extended probe can be mapped to the construct sequence to know the cleavage point. Thus, the comparison of the processing pattern of constructs, which possess variable repeat-spacer units will reveal the actual mode of processing inside the cell.

3.1.3. Considerations in construct design

We had observed that the flanking sequences of repeat affect its folding (Figure 3.1). In a CRISPR array, the first repeat is flanked by leader at the 5'-end and the first spacer at the 3'-end, while the subsequent repeats have spacers in their proximity (Figure 3.4). Since leader is present at the start of the CRISPR array it can directly affect the folding of first repeat and thereby the folding of the entire array. Therefore, we planned to incorporate a portion of the leader in our constructs to test this possibility.

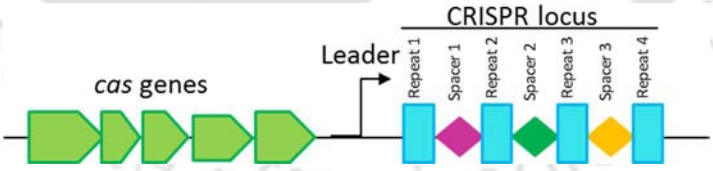


Figure 3.4 ***Schematic representation of the CRISPR locus.*** The invariant repeats are shown by blue rectangles and the diamonds represent the unique spacer sequences. The presence of leader upstream of the CRISPR array is marked. The green pentagons represent the associated *cas* genes.

The designed constructs comprise of variable number of repeat-spacer units together with a short leader sequence. The constructs are as follows – (1) 1SLT – which consists of a short leader sequence of 20 bp with 1 unit of repeat-spacer (L-R_1-S_1) and (2) 3SLT – with a short leader sequence of 20 bp with 3 units of repeat-spacer (L- R_1-S_1-R_2-S_2-R_3-S_3). To assess

the folding of the repeat units in these constructs we predicted the secondary structure using MFOLD (Zuker, 2003) (Figure 3.5). The folding pattern of each repeat in these constructs seems to vary. To analyze the processing pattern of these repeats, we designed 5' FAM labelled DNA probes having sequences complementary to the unique spacer region of the construct RNA. Thus, the cleavage pattern under different conditions can be deciphered by the variable probe extension during reverse transcription.

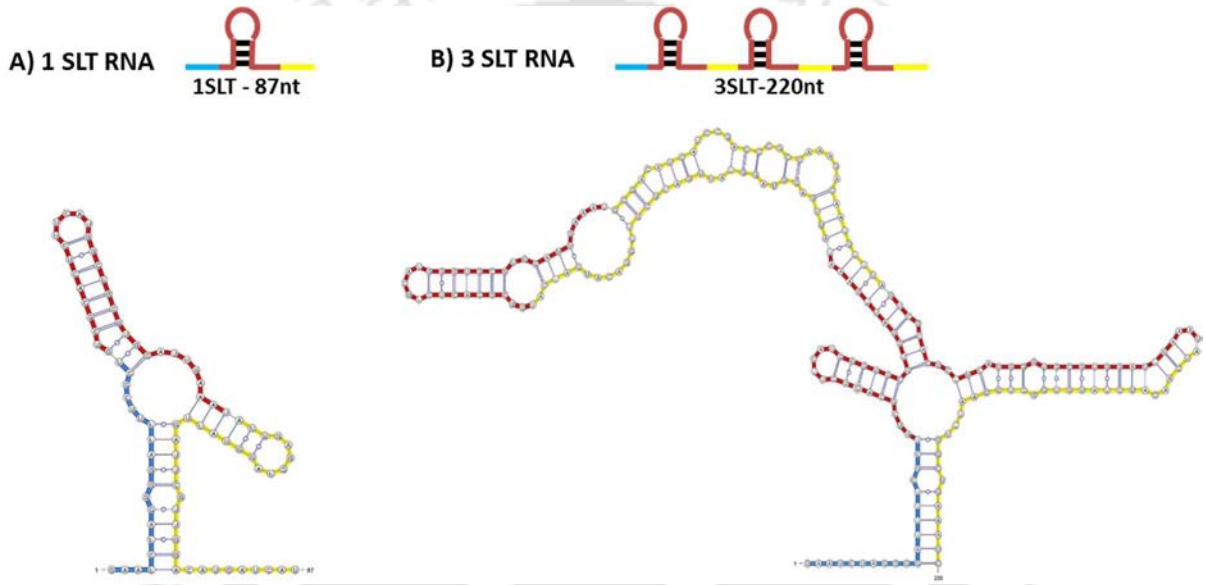


Figure 3.5 ***The differences in RNA folding of 1SLT and 3SLT constructs.*** The repeats in the construct RNA take up different folds as shown. (A) 1SLT construct of 87 nt having 1 repeat-spacer unit with a short leader region is shown. (B) 3SLT construct of 220 nt having 3 repeat-spacer units with a short leader region is shown. The leader is shown in blue, repeats in red and spacers in yellow. The folds were predicted using MFOLD (Zuker, 2003) and figures were prepared using VARNA (Darty et al., 2009).

3.2. Material and Methods

3.2.1. Preparation of substrates

The constructs of CRISPR array were amplified from *B. halodurans* genomic DNA with primers specific to leader and spacer region using Pfu polymerase. These amplified

CRISPR array constructs were cloned in p15A using restriction sites BglII and XhoI (New England Biolabs) to create pCRISPR.

For *in vivo* studies, both pCRISPR and pCas5d (cloning is described in Chapter 2) were used to transform BL21(DE3) and were allowed to co-express by growing the cells in LB medium supplemented with chloramphenicol (20µg/ml) and ampicillin (100µg/ml) at 37°C until OD at 600 nm reached 0.3. The cells were then induced by the addition of 0.2 mM IPTG followed by incubation at 37°C for 2 hours. The RNA constructs generated inside the cell were extracted using TRIzol method (Ambion) followed by RNase free-DNase I (New England Biolabs) treatment to remove any traces of contaminating DNA. The RNA was harvested using phenol-chloroform extraction and ethanol precipitation and resuspended in RNase-free water. The total RNA integrity was checked in 0.8% Agarose gel and stored at -80°C until required.

For *in vitro* studies, pCRISPR was linearized with XhoI and used as template for *in vitro* RNA synthesis using T7 polymerase. The RNA thus formed was DNase treated to remove the remains of template and the RNA was harvested by phenol-chloroform extraction method and resuspended in RNase-free water. The integrity of the synthesized RNA was checked in 12% (w/v) denaturing urea PAGE. The RNA was stored at -80°C until required. The Cas5d was purified as described in Chapter 2.

3.2.2. Nuclease activity assays

For *in vivo* processing of the pre-crRNA, the transformed BL21(DE3) cells having pCRISPR and pCas5d were co-expressed (as described above), which result in the production of RNA and Cas5d together. Upon induction, the Cas5d is allowed to process the transcripts

inside the cells. The resulting processed RNA is extracted from the cells and used for analysis of the processing pattern.

For *in vitro* nuclease activity assays, the *in vitro* synthesized RNA from the constructs (0.2 μM) were incubated with Cas5d (2 μM) in 20 mM Tris-HCl (pH 8), 100 mM KCl and 6 mM β-ME for 20 minutes at room temperature. The cleavage products were frozen and used for analysis of the processing pattern.

3.2.3. Analysis of processing pattern

The analysis of the processing pattern was done using a 5' FAM labelled DNA probe (synthesized from IDT) complementary to the desired spacer regions. RNA processed by *in vivo* and *in vitro* conditions were allowed to anneal with the 5' FAM labelled probe by heating at 65°C for 10 min followed by snap cooling. Subsequently, this was subjected to reverse transcription using the labelled probes and M-MuLV Reverse Transcriptase (New England Biolabs). The reaction was carried out in 1X M-MuLV Reverse Transcriptase buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT) in presence of 4 μM of dNTP, murine RNase inhibitor for 1 hour at 42°C followed by heating at 90°C for 10 min. The extension of the labelled probes was monitored in 15% (w/v) denaturing urea PAGE to analyze the cleavage pattern. The cDNA fragments were also analyzed using denaturing capillary electrophoresis for high-resolution mapping. Liz600 was used as size standard marker (Thermo Fisher Scientific). The .fsa file containing the peaks was visualized using Peak Scanner Software version 2.0 (Thermo Fisher Scientific). The data was plotted using SigmaPlot version 12.5.

3.3. Results and Discussion

3.3.1. Cloning and purification of constructs

The CRISPR array constructs for the study were PCR amplified using primers specific for leader and spacer regions (Figure 3.6A). The amplicons corresponding to specific constructs were further scaled up for being used for cloning (Figure 3.6B). The amplicons corresponding to the specific constructs were cloned in desired vector to produce pCRISPR for each construct (see 3.2 Materials and method). The clones were verified using vector specific primers for PCR (Figure 3.7), which was further confirmed by sequencing.

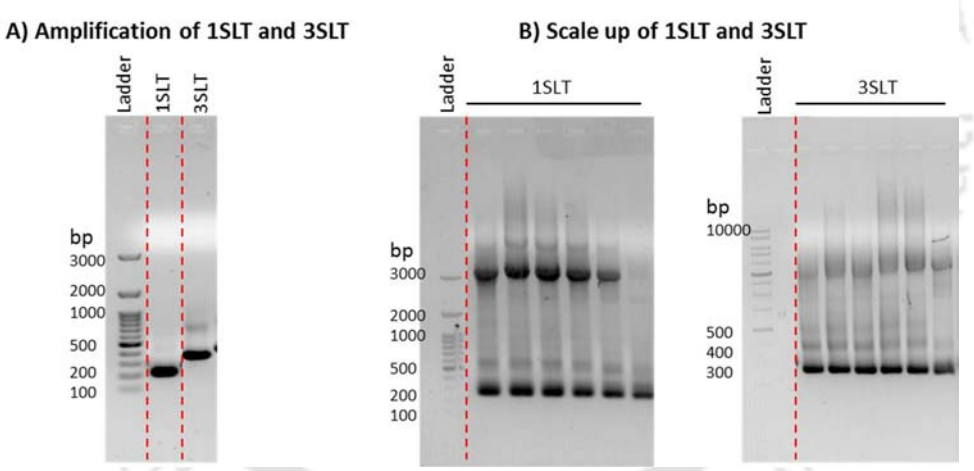


Figure 3.6 ***Amplification of CRISPR array constructs.*** (A) 1SLT and 3SLT constructs amplified from *B. halodurans* genomic DNA are shown. 1SLT construct size is 205 bp and 3SLT is 338 bp. (B) The products for 1SLT and 3SLT were scaled up for further use. The dotted red line indicates the discontinuity in gel.

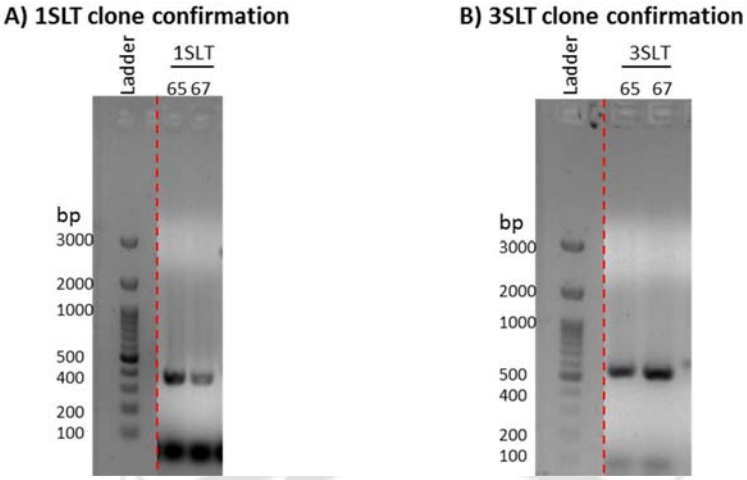


Figure 3.7 ***Clone verification of 1SLT and 3SLT constructs.*** (A) Amplicons of 1SLT (B) Amplicons 3SLT**.** Since the vector specific primers were used for clone check, the amplicons will have extra 200 bp in addition to their actual sizes. Actual sizes of constructs – 1SLT is 205 bp and 3SLT is 338 bp. The two different temperature used for annealing were 65°C and 67°C. The dotted red line indicates the discontinuity in gel.

The constructs were utilized to generate the RNA substrates both *in vivo* and *in vitro.* The *in vivo* RNA substrates were expressed by inducing the cells containing pCRISPR for the constructs, followed by total RNA extraction. In order to obtain the *in vivo* processed RNA, the cells containing pCas5 and pCRISPR were induced to co-express Cas5d and RNA of the constructs respectively. This ensues the RNA processing by Cas5d, which is followed by total RNA extraction from the cells (Figure 3.8A). For obtaining the RNA substrates *in vitro*, the pCRISPR of the constructs was linearized by XhoI to serve as template for *in vitro* RNA synthesis by T7 polymerase. The synthesized 1SLT RNA is 87 nt and 3SLT RNA is 220 nt in length (Figure 3.8B).

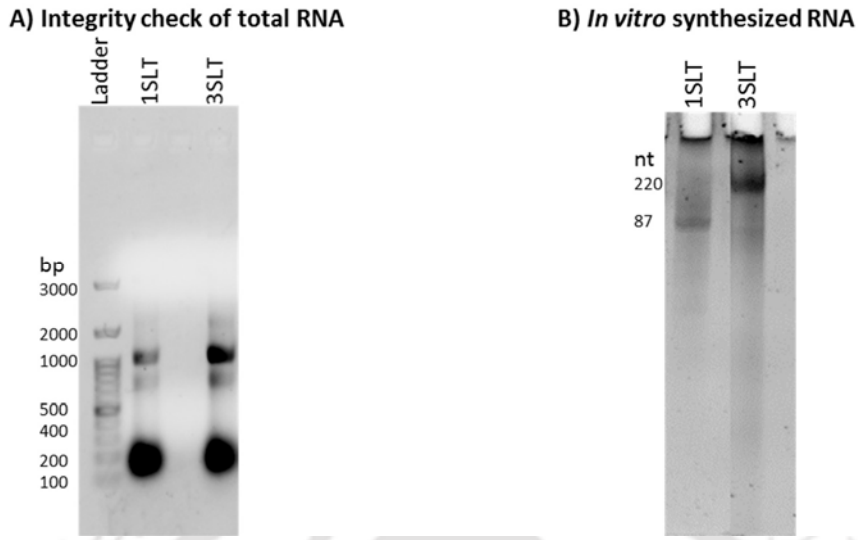


Figure 3.8 ***RNA integrity of the constructs.*** (A) Total RNA extracted from co-transformed cells. Integrity checked in 0.8% Agarose gel. (B) The integrity of *in vitro* synthesized RNA was checked in 12% (w/v) denaturing urea PAGE.

3.3.2. Evidence of co-transcriptional processing of CRISPR repeats *in vivo*

To understand the effect on repeat processing due to differences in the folding of repeats, we analyzed the processing pattern of constructs containing 1 unit of repeat-spacer (1SLT) and 3 units of repeat-spacer (3SLT) by extending a labelled DNA probe by reverse transcription.

The probing of *in vivo* and *in vitro* processed RNA of 1SLT construct using a labelled probe corresponding to spacer 1 is shown in Figure 3.9. The probe extension is monitored following a reverse transcription reaction.

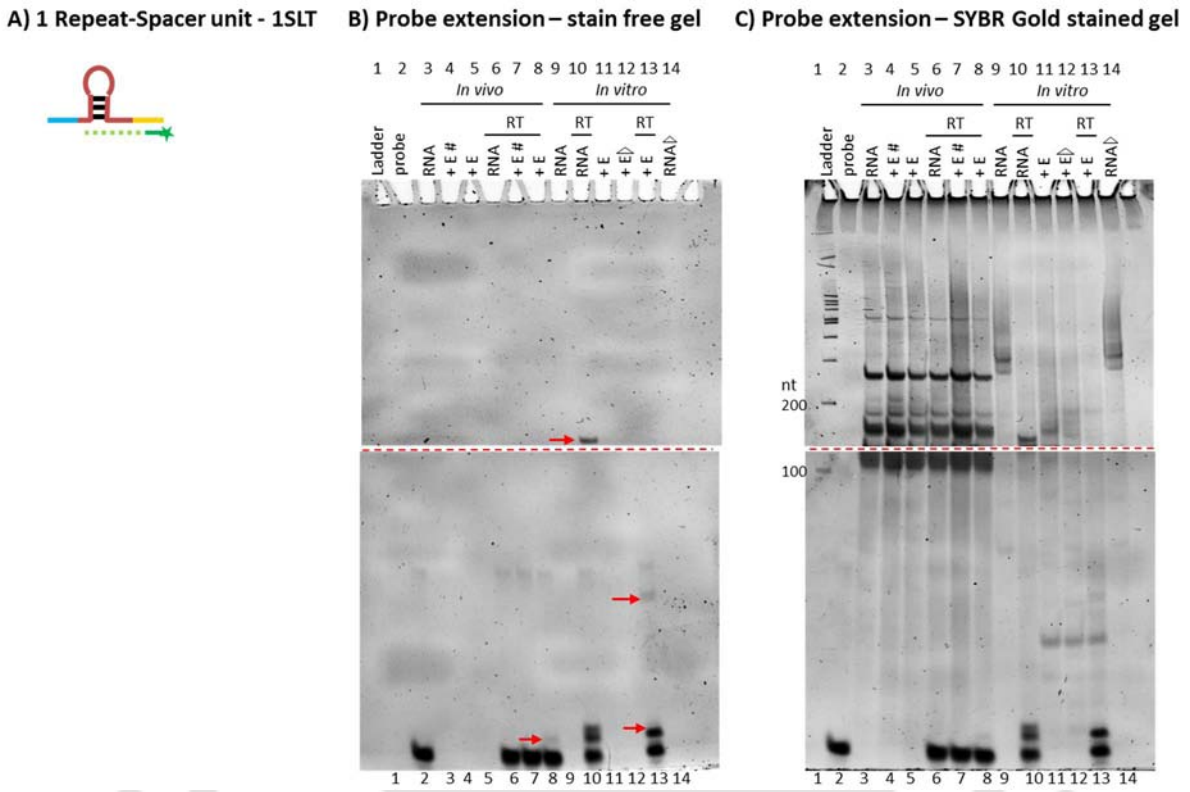


Figure 3.9 ***Probing in vivo and in vitro processed RNA of 1SLT with Spacer1.*** (A) The pictorial depiction of 1SLT construct showing the leader region in blue, repeat in red, spacer in yellow and the probe complementary to spacer region in green. (B) Probe extension under the various conditions is visualized in stain free gel (C) The same gel in (B) is stained with SYBR Gold to visualize the corresponding unlabeled RNA. Lanes 3-8 belong to *in vivo* assay and lanes 9-14 are of *in vitro* assay. To observe the *in vivo* processing, cells harbouring pCRISPR and pCas5 were induced to co-express the RNA and Cas5d (represented by E) followed by total RNA extraction for analysis. Lane 4 and 7 containing the total RNA extracted from the uninduced cells are marked with #. In lanes 6-8 the same samples of lanes 3-5 respectively were used for reverse transcription reaction. To observe *in vitro* processing, the fully formed RNA was incubated with Cas5d (represented by E) for 20 min. Lanes 12 and 14 contain the *in vitro* processed and unprocessed RNA subjected to heat treatment, indicated by empty black triangle. The lanes indicated with RT contain the samples subjected to reverse transcription reaction. The probe extension is shown by red arrows. Lane 2 contains the unextended labelled probe. The processing pattern was analyzed in 15% (w/v) denaturing urea PAGE.

For 1SLT construct, the condition of *in vivo* processed RNA showed the presence of a single band above the labelled probe (lane 8), while the *in vitro* processed sample showed the presence of higher order bands (lane 13). The unprocessed *in vitro* RNA (lane 10) showed a band much higher than all other bands, which can be considered as completely extended probe corresponding to the full length of the construct RNA. Similar reverse transcription

reaction was carried out for *in vivo* and *in vitro* processed RNA using the labelled probe for spacer 1 (Figure 3.10).

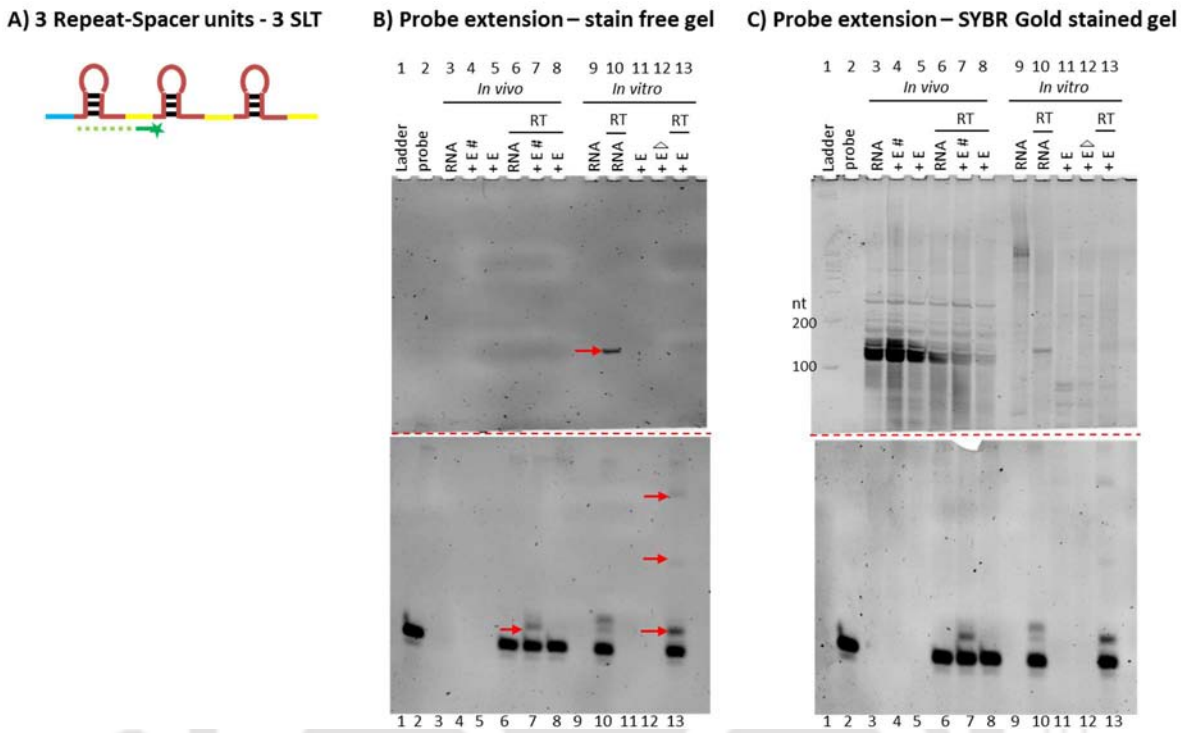


Figure 3.10 ***Probing in vivo and in vitro processed RNA of 3SLT with Spacer1.*** (A) The pictorial depiction of 3SLT construct showing the leader region in blue, repeat in red, spacer in yellow and the probe complementary to spacer region in green. (B) Probe extension under the various conditions is visualized in stain free gel (C) The same gel in (B) is stained with SYBR Gold to visualize the corresponding unlabeled RNA. Lanes 3-8 belong to *in vivo* assay and lanes 9-13 are of *in vitro* assay. To observe the *in vivo* processing, cells harbouring pCRISPR and pCas5 were induced to co-express the RNA and Cas5d (represented by E) followed by total RNA extraction for analysis. Lane 4 and 7 containing the total RNA extracted from the uninduced cells are marked with #. In lanes 6-8 the same samples of lanes 3-5 respectively were used for reverse transcription reaction. To observe *in vitro* processing, the fully formed RNA was incubated with Cas5d (represented by E) for 20 min. Lane 12 contains the *in vitro* processed RNA subjected to heat treatment, indicated by empty black triangle. The lanes indicated with RT contain the samples subjected to reverse transcription reaction. The probe extension is shown by red arrows. Lane 2 contains the unextended labelled probe. The processing pattern was analyzed in 15% (w/v) denaturing urea PAGE.

The probe extension pattern observed for 3SLT *in vivo* RNA sample (lane 7) seemed to be similar to that of 1SLT construct (lane 8 in Figure 3.9). There was an extended band just above the labelled probe in *in vivo* RNA sample (lane 7). The *in vitro* processed sample showed the presence of higher order bands similar to 1SLT construct (lane 13 in Figure 3.9)

but 3SLT showed more bands (lane 13). The unprocessed *in vitro* RNA (lane 10) showed a band much higher than all other bands, which can be considered as the full length RNA of the 3SLT construct.

The results thus obtained from the reverse transcription reactions of the differently processed RNAs (*in vivo* and *in vitro* processed RNA) with labelled probe showed differences in band patterns. These band patterns emerge due to the extension of labelled probe to varied length, suggestive of differences in *in vivo* and *in vitro* processing of RNA, which in turn reveals the differences in the cleavage site of Cas5d. This can be attributed to unconstraint folding of the RNA after transcription, which results in the formation of non-canonical sites in the repeats *in vitro,* that is differently recognized during post-transcriptional processing of RNA. This is evident by the differences in the *in vitro* processed RNA of 1SLT and 3SLT construct, which showed the presence of multiple fragments of different sizes (Figure 3.9 and 3.10). But, interestingly, in all constructs whether 1SLT (1 repeat-spacer unit) or 3SLT (3 repeat-spacer units) the band pattern for *in vivo* processed RNA seems to be the same which is a clear indication that repeats are processed uniformly inside the cell. This implies that all the repeats have a similar fold that is subjected to Cas5d processing. This is possible only when all the repeats are processed as soon as they are synthesized so that their fold is not effected by flanking sequence due to increase in number of the repeat-spacer units. Thus, co-transcriptional processing seems to provide specificity in repeat processing *in vivo*.

3.3.3. Fragment analysis confirmed co-transcriptional processing of repeats

To get a better resolution of the CRISPR repeat processing, we tried to analyze the fragment sizes obtained from *in vivo* and *in vitro* processing of RNA by Cas5d. We used denaturing capillary electrophoresis for high-resolution mapping of fragments by employing Liz600 as size standard marker.

The cDNA obtained from the reverse transcription reaction of *in vivo* and *in vitro* processed RNA with the 5' FAM labelled probe was subjected to fragment analysis using capillary electrophoresis. The results show differences between the sizes of *in vivo* and *in vitro* processed RNA products of 1SLT construct, which has 1 repeat-spacer unit (Figure 3.11). This confirms our finding that there are differences between *in vivo* and *in vitro* processed repeat RNA due to differences in folding of repeats which lead to the differential recognition by Cas5d.

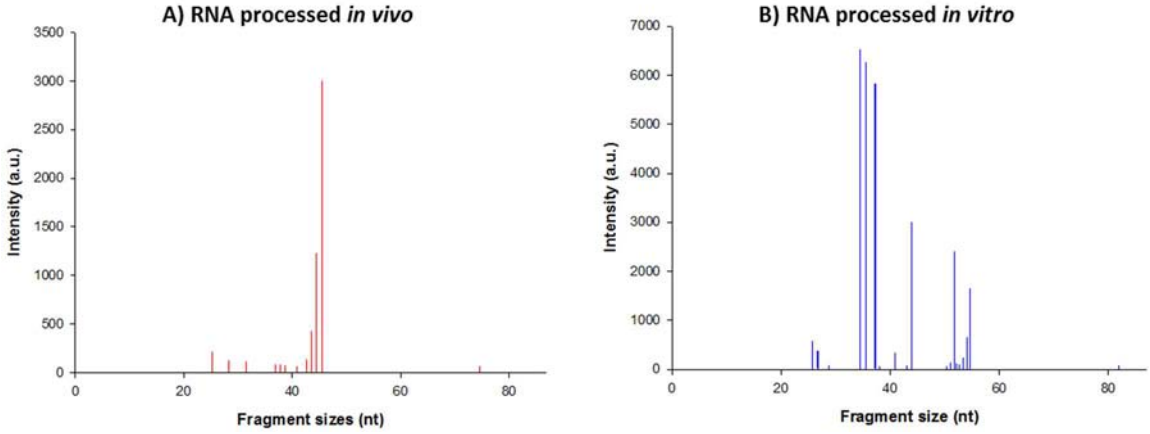


Figure 3.11 ***Fragment analysis of 1SLT probed with Spacer 1.*** (A) The fragments obtained from *in vivo* processed RNA is shown in red. (B) The fragments obtained from *in vitro* processed RNA is shown in blue.

The *in vivo* processed RNA showed a prominent fragment of size 46 nt (Figure 3.11A) and some fragments towards the lower range of negligible intensity, while the *in vitro* processed repeat showed multiple high intensity bands around size of 34-37 nt and another cluster of lower intensity sizing around 51-54 nt, in addition to the 46 nt fragment (Figure 3.11B). The size of these cDNA fragments were mapped onto the sequence of the corresponding RNA to decipher the cleavage site of Cas5d under different conditions (Figure 3.12). The labelled DNA probe can be extended only till the point of cleavage and since the probe was designed complementary to the spacer region, its position in the sequence of construct RNA was known. In case of 1SLT construct, the full length extension of the probe complementary to spacer 1 will result in 87 nt cDNA fragment which comprises of 20 nt short leader region, 32 nt repeat and 35 nt spacer region (Figure 3.12). The fragment size of 46 nt obtained from probing the *in vivo* processed 1SLT RNA with spacer 1 will therefore correspond to the cleavage site at G21 of the repeat (Figure 3.12). This was in agreement with the earlier reports on processing of single repeat by Cas5d *in vitro* (Nam et al., 2012; Punetha et al., 2014), wherein Cas5d cleaved at G21, though longer incubation ensued extended processing (Punetha et al., 2014). For *in vitro* processed RNA, the fragment of 46 nt also shows the cleavage at G21 but the additional multiple bands of size 34 nt, 35 nt, 36 nt and 37 nt correspond to cleavage at A1 of the spacer and U32, A31, A30 of the repeat respectively. The another cluster of lower intensity bands sizing around 51-54 nt corresponds to the cleavage at positions G16, G15, U14, A13 of the repeat respectively (Figure 3.12). Thus, Cas5d shows non-specific cleavage *in vitro*.

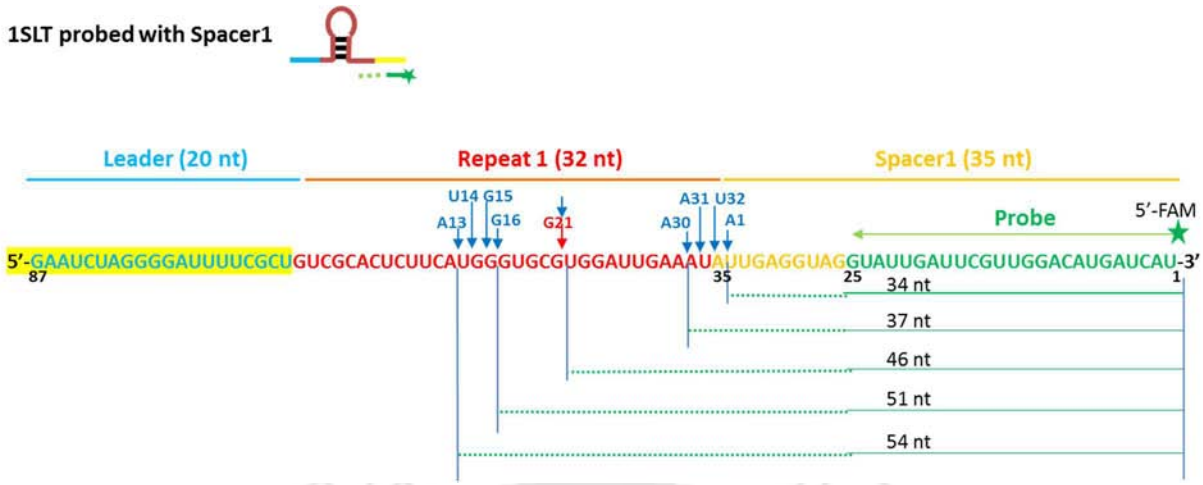


Figure 3.12 ***Mapping the size of the extended probe to decipher the cleavage site in 1SLT.*** In the RNA of 1SLT the leader region is shown in blue, repeat in red, spacer in yellow and the spacer region complementary to probe in green. The labeled DNA probe extension occurs till the point of cleavage. These fragments can be mapped onto the sequence from the position of the probe. The fragment of 46 nt observed for *in vivo* processed RNA corresponds to cleavage site at G21 of the repeat, indicated by red arrow. The fragment clusters of 34-37 nt and 51-54 nt for *in vitro* processed RNA corresponds to cleavage at A1 of the spacer-A30 of the repeat and G16-A13 of the repeat respectively, indicated by blue arrows. Apart from these clusters, there was also a fragment of 46 nt resulting from the cleavage at G21, which was similar to the *in vivo* cleavage site.

We further investigated the case of 3SLT which had 3 repeat-spacer units to observe the effect of added unit length. For *in vivo* processed RNA the fragments sizes were similar to the fragments observed for single repeat-spacer unit, while the fragments obtained for *in vitro* processed RNA clustered around different sizes in 3SLT as compared to 1SLT (Figure 3.13).

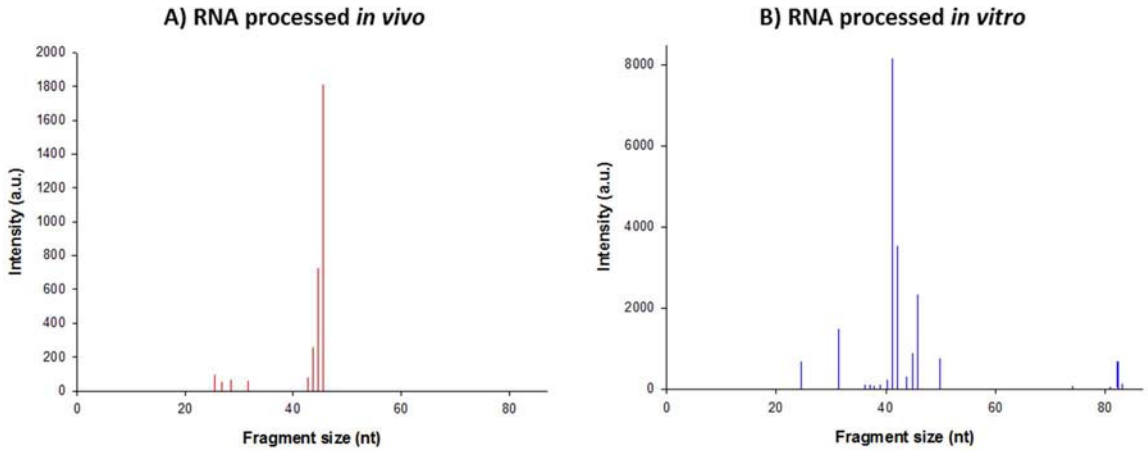


Figure 3.13 ***Fragment analysis of 3SLT probed with Spacer 1.*** (A) The fragments obtained from *in vivo* processed RNA is shown in red. (B) The fragments obtained from *in vitro* processed RNA is shown in blue.

In case of 3SLT, the *in vivo* processed RNA showed a prominent fragment of size 46 nt (Figure 3.13A) with few other fragments in lower range of negligible intensity, while the *in vitro* processed repeat showed multiple high intensity fragments of 31 nt, 41 nt, 42 nt, 45 nt and 46 nt (Figure 3.13B) and also few lower intensity fragments around 82 nt. Since the probe is complementary to spacer 1, the cDNA fragment of 87 nt will correspond to full length extension of the probe. The fragment size of 46 nt indicates the cleavage site to be at G21 of the repeat as shown in *in vivo* and *in vitro* processed 3SLT RNA (Figure 3.14), which is similar to 1SLT (Figure 3.12). The fragment size of 31 nt, 41 nt, 42 nt and 45 nt corresponds to the cleavage at G4 in the spacer, U26, A25 and U22 in the repeat region respectively. Any fragment above 67 nt and below 87 nt may result due to non-specific cleavage in the leader region. Thus, the additional fragments in the repeat region and also in the flanking regions in the *in vitro* processed RNA indicates the non-specific cleavage of RNA by Cas5d due to the formation of non-canonical cleavage sites.

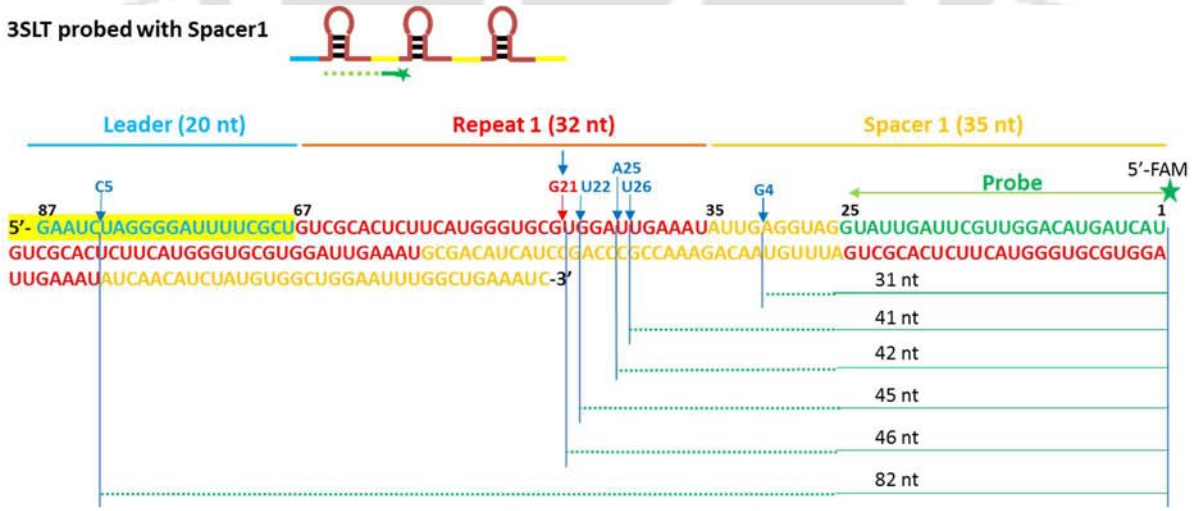


Figure 3.14 ***Mapping the fragment size of the extended probe to decipher the cleavage site in 3SLT.*** In the RNA of 3SLT the leader region is shown in blue, repeats in red, spacers in yellow and the spacer region complementary to probe in green. The labeled DNA probe extension occurs till the point of cleavage. These fragments can be mapped onto the sequence from the position of the probe. The cleavage sites are indicated by coloured arrows – red for *in vivo* processed and blue for *in vitro* processed RNA. The fragment of 46 nt corresponds to cleavage site at G21 of the repeat. In case of *in vitro* processed the fragments of 31 nt and 82 nt indicate cleavage in the flanking region of the repeat. Apart from this, the fragments of 41 nt, 42 nt, 45 nt indicate non-specific cleavages at U26, A25 and U22 respectively in the repeat region.

Interestingly, the comparison of the fragment analysis results of 1SLT and 3SLT *in vivo* and *in vitro* processed RNA probed with spacer 1 showed the same band fragment of 46 nt corresponding to cleavage site of G21 in repeat (Figure 3.15), suggesting it to be the canonical cleavage site of Cas5d. For the *in vitro* processed RNA both the constructs showed the occurrence of non-specific cleavage in the flanking regions of the repeat and also multiple cleavages in the repeat region, resulting in a variable band pattern (Figure 3.15).

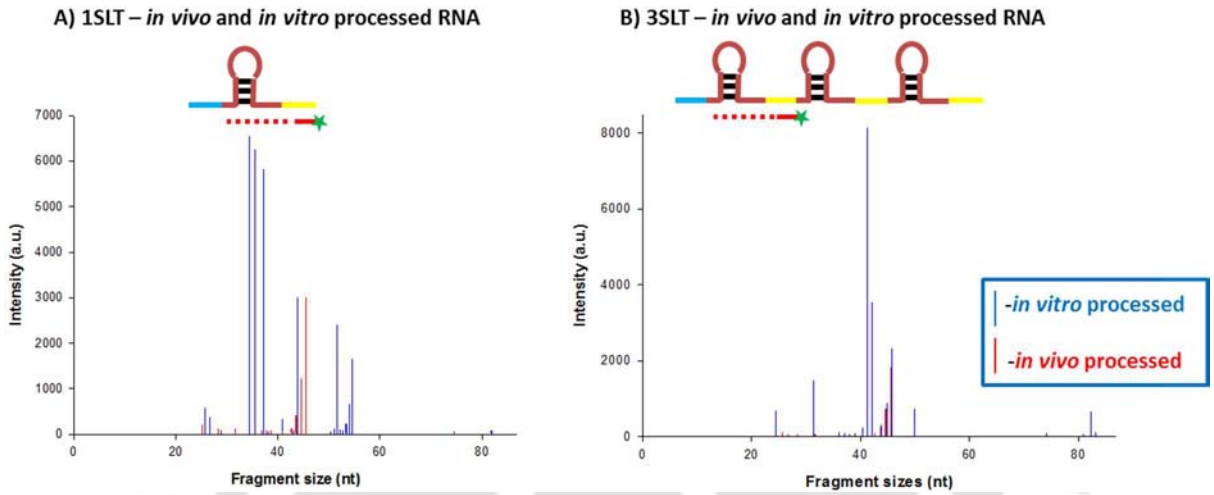


Figure 3.15 ***Comparison of fragment analysis results of 1SLT and 3SLT probed with Spacer 1.*** (A) The size comparison between *in vivo* and *in vitro* processed 1SLT RNA. (B) The size comparison between *in vivo* and *in vitro* processed 3SLT RNA. The *in vitro* fragments size and intensity is shown in blue and the *in vivo* fragments in red. In both the cases, the *in vivo* processed RNA band pattern is similar around 46 nt in size.

This advocates that the repeats of *in vitro* synthesized RNA have different folds, which can be a resultant of the effect of flanking sequence on folding of the repeat. Moreover, the addition of repeat-spacer units in RNA may further increase the complexity of folding thereby imposing constraint on Cas5d to access all the repeats uniformly, generating non-specific processing pattern. Inside the cell, the repeats are present as a part of CRISPR array which has multiple repeat-spacer units that can affect the repeat folding and thereby its processing if processing occurred post-transcriptionally. But the non-specific processing of

repeat as observed *in vitro* does not seem to occur in *in vivo* condition. We observed the specific cleavage site at G21 in the repeat in both the constructs – 1SLT and 3SLT. This is possible if the transcription and processing are coupled, so that similar folds are subjected to processing.

3.4. Summary

We identified the coupling between CRISPR RNA transcription and its processing as a means of getting a specific mature product from non-specific endonuclease. Based on our earlier experiments, we know Cas5d possesses the ability to recognize and process various CRISPR repeats having different sequence and secondary structure and generates product of varying length. This activity of Cas5d *in vitro*, which is tilted towards non-specificity can be tuned towards specificity if the CRISPR repeats of similar folds are provided to Cas5d without any constraint. This is possible in case of co-transcriptional processing where repeats are processed soon after their formation, so that their fold is not effected by any other proximal sequence or the complex folding of CRISPR array. The nature therefore seems to tune a non-specific nuclease towards specificity for driving a particular function. The non-specific function can be required for other processes, which are not yet known. Thus, we discussed the importance of co-transcriptional processing of repeats to generate a homogenous population of mature guide RNA by Cas5d in type I-C CRISPR-Cas system.

4.1. Introduction

In Chapter 2, we have shown that the presence of metal retards the extended processing of repeat RNA by Cas5d (Figure 2.20 in Chapter 2). However, when the metal is absent the enzyme efficiently cleaves the RNA sequentially into smaller products (Figure 2.18 and 2.8 in Chapter 2). The effect of metal on RNA integrity was also tested by incubating only the RNA with metal, but, this did not compromise the RNA integrity. This alludes to the possibility of binding of metal to Cas5d, which might lead to some structural changes thereby effecting the RNA hydrolysis. This motivated us to investigate the utility of metal binding to Cas5d. Interestingly, a report came up that showed Cas5d orthologs from *Streptococcus pyogenes* and *Xanthomonas oryzae* bind DNA but do not cleave it (Koo et al., 2013). This prompted us to examine the factors that typically influence the DNase activity.

The hydrolysis of the nucleic acids reveals an interesting aspect on the requirement of external nucleophile by nucleases. The ribonucleic acid (RNA) has an inbuilt nucleophile in the form of 2'-OH group and therefore the hydrolysis can be initiated in the absence of an external nucleophile. Since deoxyribonucleic acid (DNA) lacks this group, consequently, most of the DNases utilize metal ions as cofactors for the cleavage activity. Therefore, we intended to test whether metal ions could stimulate the DNA cleavage by Cas5d. These attempts are presented in this chapter.

4.2. Materials and methods

4.2.1. Cloning, expression and purification

The Cas5d was purified as described in Chapter 2. The point mutants E4A, D56A, H98A and E100A of Cas5d were generated by mega primer based PCR method. All mutants were cloned in pQE2 except H98A which was cloned in LIC vector (a kind gift from Scott Gradia, Addgene ID: 29717) having a pET backbone and encodes an N-terminal Strep-tag II. The cloned constructs were verified by sequencing. Expression was performed in *E. coli* BL21(DE3) by growing the cells in LB medium supplemented with ampicillin (100 µg/ml) at 37°C until OD at 600 nm reached 0.7. The temperature was then reduced to 20°C for 20 min and protein expression was induced by the addition of 0.2 mM IPTG followed by incubation at 20°C overnight. The cells were harvested by centrifugation and resuspended in buffer A containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6mM β-ME and 1 mM PMSF. After sonication, the lysate was clarified by centrifugation at 36,500g for 30 min. The supernatant was treated with RNase to remove any bound RNA and then loaded onto a 5 ml HiTrap IMAC HP column or StrepTrap HP column (GE Healthcare) pre-equilibrated with buffer B containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6mM β-ME. After washing the column with buffer C containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6 mM β-ME and 40 mM imidazole, the bound protein was eluted using a linear gradient of imidazole (upto 500mM) in buffer C. For strep-tagged protein, washing buffer contained 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6mM β-ME and the elution was carried out with buffer C that contained 2.5 mM D-Desthiobiotin in place of imidazole. The eluted protein was incubated with 10 mM EDTA for 1hr to remove the bound metal ions if any and then dialyzed against buffer D containing 20 mM Tris-HCl (pH 8.0), 200 mM NaCl and 6mM β-ME.

Subsequently, the proteins were aliquoted, snap frozen in liquid nitrogen and stored at -80°C until required.

4.2.2. Preparation of substrates

Pre-crRNA containing only the repeat sequence was chemically synthesized and labelled with a 6-FAM at the 3' end (IDT). The DNA sequences corresponding to the CRISPR repeat with a T7 promoter (-) sequence were obtained from IDT and subsequently were labelled with fluorescein tagged dUTP using deoxy terminal transferase (New England Biolabs) at the 3'-end. pQE2 or pEGFP plasmid was employed as circular DNA and pQE2 linearized with KpnI served as linear substrate. Single stranded circular M13mp18 phage DNA was obtained from New England Biolabs.

4.2.3. Nuclease activity assays

The nuclease activity against both the substrates RNA and DNA, was performed in the presence of 10 mM Mg^{2+} at 37°C for 1 hr. The 3' 6-FAM labelled pre-crRNA repeat at 0.2 μM concentration was incubated with Cas5d (2 μM) in presence of 3'-fluorescein labelled CRISPR DNA in 20 mM Tris-HCl (pH 8), 100 mM KCl and 6 mM β-ME. Cleavage products were analyzed on 15% (w/v) denaturing urea PAGE.

DNase activity assays were performed with double stranded (linear or circular) and single stranded (circular) DNA at 37°C for 1hr in the buffer containing 20 mM Tris-HCl (pH 8.0), 100 mM KCl, 6 mM β-ME, 10 mM $MgCl_2$ and 2 μM Cas5d. Time dependent nuclease activity was performed at 37°C and samples were taken at the indicated time intervals. The

reaction was stopped using 50 mM EDTA (pH 8.0) and the products were analyzed on 0.8% agarose gel and visualized by ethidium bromide staining. Metal dependent DNase activity was measured in the presence of 10 mM divalent cation (Mg^{2+}, Mn^{2+}, Zn^{2+}, Ca^{2+} and Ni^{2+}) and 50 mM EDTA (if indicated).

4.2.4. Analysis of binding sites

The nature of metal binding is probed using Hill plot analysis in the absence of DNA using the following form of the equation:

$$\log\left(\frac{\Delta F}{(\Delta F_{max} - \Delta F)}\right) = n_H \log\left[Mg^{2+}\right] - \log\left(K_d\right) \quad (1)$$

where n_H is the Hill coefficient and K_d is the apparent dissociation constant. $\Delta F = F_0 - F$ where F_0 denotes the fluorescence intensity in the absence of Mg^{2+}, F is the fluorescence intensity at a particular concentration of Mg^{2+} and ΔF_{max} represents the difference between F_0 and F at infinite concentration of Mg^{2+}.

The dependence of the DNA cleavage activity on the concentration of Mg^{2+} was analyzed using the following form of the Hill equation:

$$\log\left(\left(\frac{(100-b)}{(100-x)}\right)-1\right) = n_H \log\left[Mg^{2+}\right] - \log\left(K_d\right) \quad (2)$$

where n_H is the Hill coefficient and K_d is the apparent dissociation constant, b is the activity of enzyme in the absence of Mg^{2+} and x is the percentage of activity in the presence of varying concentrations of Mg^{2+}. The maximal percentage activity of DNA cleavage was taken as 100%. The data were fit using SigmaPlot version 12.5.

Intrinsic fluorescence emission spectrum of the protein was measured at 26°C by using a Fluoromax spectrofluorometer (HORIBA Jobin Yvon). To probe the tryptophan environment, the excitation wavelength used was 295 nm and emission scan was done from wavelength 310 nm to 500 nm. The concentration of protein used was 10 µM in 20 mM Tris-Cl (pH 8.0). The spectrum generated is an average of three scans after baseline correction. The slit width used was 1 nm for excitation and 9 nm for emission.

4.3. Results and Discussion

4.3.1. Investigating the Cas5d DNase activity

We used pQE2 plasmid as substrate to test the nuclease activity. It was observed that while Cas5d cleaved the circular DNA, the cleavage was more dramatic in the presence of a divalent metal ion (Figure 4.1A). The plasmid preparation showed polymorphism in its mobility that is typical for a circular DNA. Cas5d showed no preference for these and all of them were digested with no traces of DNA left behind (Figure 4.1A). This suggested that Cas5d, indeed, exhibits endodeoxyribonuclease activity that is stimulated in the presence of a divalent metal. Encouraged by this, we asked whether it is adept at acting on linear substrate too. Therefore, we linearized the pQE2 and repeated the assay. Here too, the activity was seen prominently in the presence of a divalent metal ion (Figure 4.1B). The aforementioned experiments were performed using double stranded DNA substrates which raised the question on Cas5d activity, that whether it is capable of acting on single stranded DNA substrates or not. Hence, we employed the single stranded circular DNA from M13mp18 phage for the activity assay and here too we noticed the DNase activity that was discernable when a

divalent metal ion was present (Figure 4.1C). Thus, we found that Cas5d is able to recognize and cleave all forms of DNA efficiently in presence of metal.

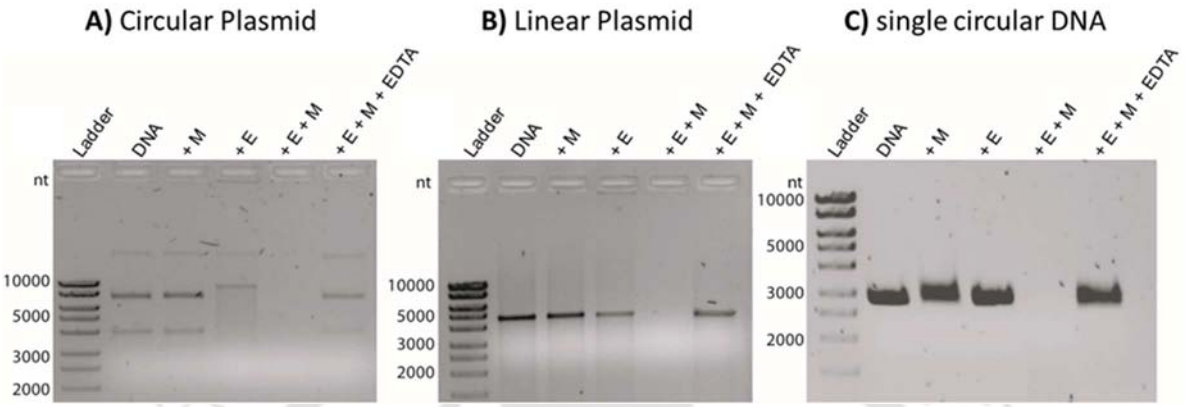


Figure 4.1 ***Cas5d activity against different forms of DNA.*** The DNase activity of Cas5d was assayed in the presence of (A) double stranded circular DNA (B) double stranded linear DNA (C) M13mp18 phage single stranded circular DNA. In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown.

The hallmark of the RNase activity of Cas5d is its ability to recognize the structured form of RNA. Nam et al. (2012) systematically varied the sequences of the repeat RNA and identified that the recognition is stronger at the base of the stem and 3'-end overhang. Prompted by this, we asked whether the DNA recognition is structure specific too. We used the sense, the antisense and the duplex forms of the CRISPR repeat DNA to clarify whether the sequence is recognized and cleaved in a manner similar to RNA. Both the sense and antisense DNA showed the presence of stem and loop when subjected to the fold prediction using MFOLD (Zuker, 2003) (Figure 4.2).

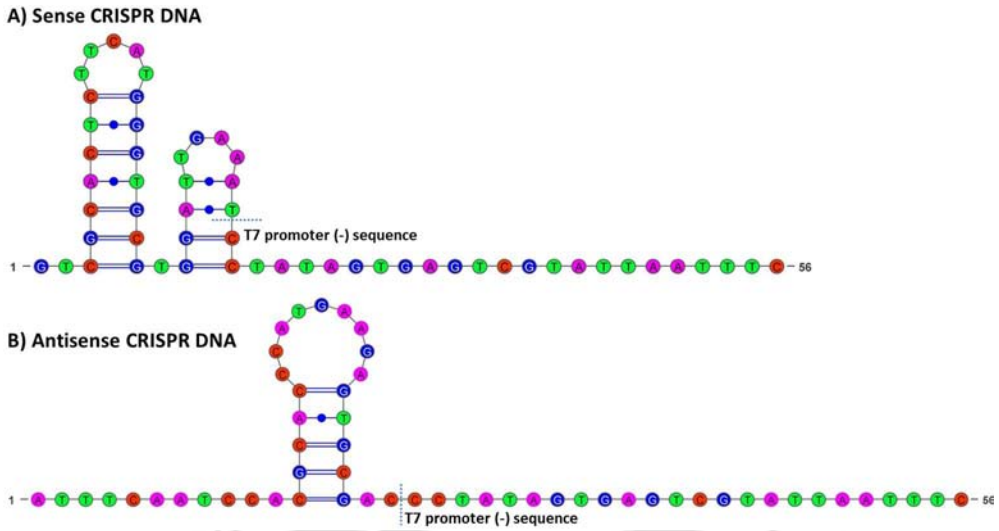


Figure 4.2 ***The folded architecture of CRISPR repeat DNA.*** The (A) Sense and (B) Antisense CRISPR DNA with additional T7 promoter sequence (-) is shown. The base Adenine is shown in purple, Thymine in green, Guanine in blue and Cytosine in red. The folds were predicted using MFOLD (Zuker, 2003) and the figures were prepared using VARNA (Darty et al., 2009).

We observed that Cas5d cleaved sense, antisense and the duplex forms preferentially in the presence of the divalent metal (Figure 4.3). Unlike the CRISPR repeat RNA where a single cleavage is shown to occur between the positions G21 and U22, there seemed to be no such preferential cleavage of the DNA substrates (Figure 4.3). This corroborates that Cas5d displays sequence-independent endonuclease activity against both single and double stranded DNA substrates that is stimulated in the presence of the divalent metal ion.

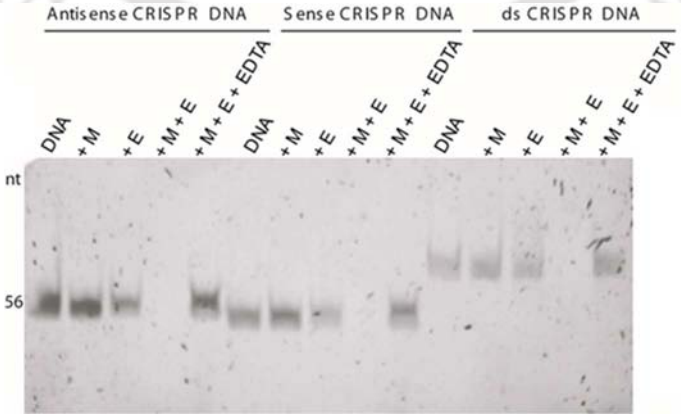


Figure 4.3 ***Cas5d activity against CRISPR DNA.*** The DNase activity of Cas5d was assayed in the presence of the single stranded antisense and sense strands as well as the duplex of sense and antisense CRISPR repeat DNA. In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown.

4.3.2. Factors modulating the DNase activity of Cas5d

4.3.2.1. Effect of metals on the Cas5d DNase activity

Motivated by the nuclease activity of Cas5d in presence of divalent metal ion, we investigated whether Cas5d shows preference towards a particular type of metals. To examine this, the reaction was conducted in the presence of Mg^{2+}, Mn^{2+}, Zn^{2+}, Ca^{2+} and Ni^{2+}, which are often found to be associated with nucleic acid binding proteins. The activity was observed in the presence of all metals except Ca^{2+} (Figure 4.4A). This suggests that perhaps there is a metal selectivity filter within the structure which determines the specificity for a particular kind of metal. Since the size of these metal ions differ, it is possible to attribute the differences in the activity to differences in the size (ionic radius) of the metal ions – Ni^{2+} (83 pm); Mn^{2+} (81 pm); Zn^{2+} (88 pm); Mg^{2+} (86 pm) and Ca^{2+} (114pm). Therefore, it is likely that the active site that harbors the metal may accommodate those metals with ionic radius ranging from 80 to 90 pm. This also explains why Ca^{2+} with an ionic radius well beyond the aforementioned range does not elicit a response. Owing to this opposing nature, it seems probable that Mg^{2+} and Ca^{2+} can compete against each other to regulate the DNase activity (Figure 4.4B).

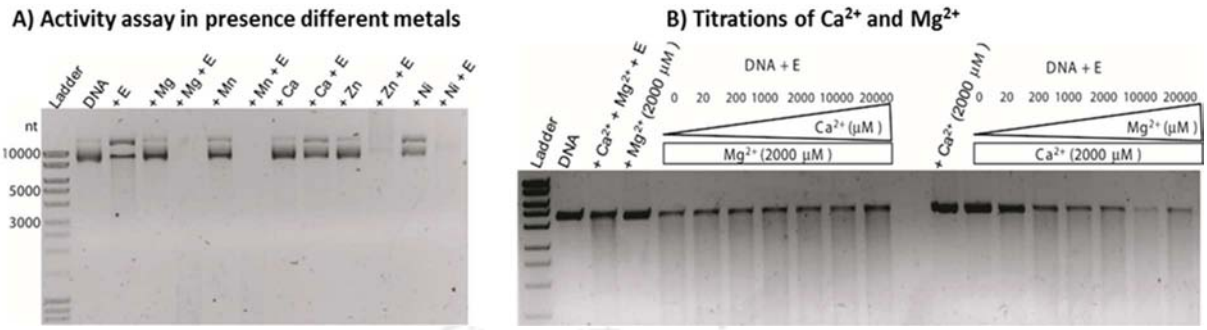


Figure 4.4 ***Cas5d selectivity filter for metals.*** (A) Nuclease activity in presence of 10 mM Mg^{2+}, Mn^{2+}, Ca^{2+}, Zn^{2+} and Ni^{2+} is shown. In all panels, E represents Cas5d. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown. (B) Inhibitory effect of Ca^{2+} on the Cas5d DNase activity is observed in titrations of Ca^{2+} and Mg^{2+}. The rectangle represents the metal ion whose concentration is kept constant whereas the triangle depicts the metal ion whose concentration varies. The lanes containing the DNA along with the respective metals are shown. The lane 3 (from left) containing equal concentration of Ca^{2+} and Mg^{2+} in presence of enzyme is indicated. The experiment shows that in the presence of 2 mM Mg^{2+}, the inhibitory effect of Ca^{2+} is apparent only at 20 mM concentration. On the other hand, when Ca^{2+} concentration is kept constant at 2 mM, rapid cleavage is perceptible even in the presence of 200 µM Mg^{2+}.

4.3.2.2. Effect of salts on the Cas5d DNase activity

It is known that salts affect the electrostatic interactions between protein and nucleic acid. So, we tested the influence of the nature and concentration of salts on the nuclease activity of Cas5d. Hence the assay was conducted in the presence of 50, 100 and 200 mM of NaCl, KCl and NH_4Cl. It was observed that the activity remained unaffected at 50 and 100 mM respectively, for all type of salts, which suggests that the nature of the salt does not have any apparent effect on the activity. However, when the concentration was increased to 200 mM, in all three salts, the activity got reduced to a significant extent (Figure 4.5). This also suggests that the DNA recognition could involve significant electrostatic interaction.

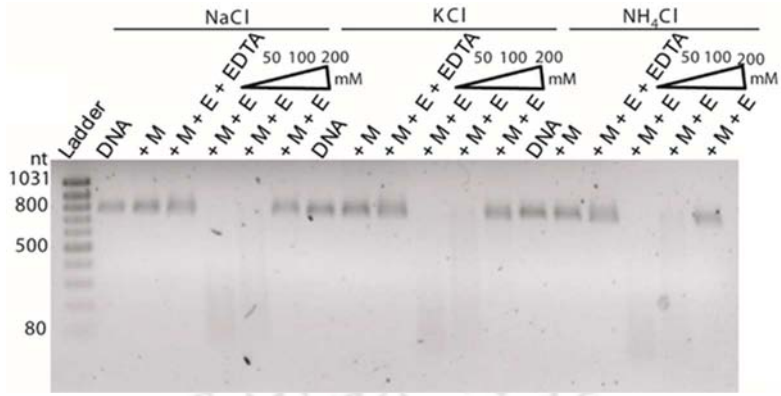


Figure 4.5 ***Effect of salts on Cas5d DNase activity.*** Assay in the presence of NaCl, KCl and NH_4Cl is shown and the corresponding salt in the lanes is indicated. The increasing concentration (50, 100 and 200 mM) of the salt in the corresponding lanes is depicted as a triangle. In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown.

4.3.2.3. Effect of substrate length on the Cas5d DNase activity

Next we tried to understand whether there is any dependence towards the DNA substrates. To address this, we used linear DNA substrates of varied sizes (260, 795, 2500 and 4800 bp) and sequences. Incision of the DNA substrates was observed prominently when Mg^{2+} was included (Figure 4.6). This suggests that the Cas5d nuclease activity against DNA seems to be promiscuous and independent of the length and sequences of the substrates.

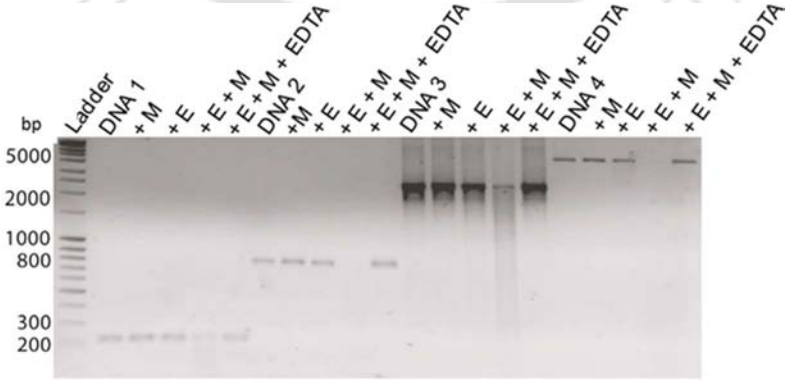


Figure 4.6 ***Effect of substrate length on Cas5d DNase activity.*** Activity in presence of the substrate with varied length is shown. DNA 1 (260 bp), DNA 2 (800 bp), DNA 3 (2500 bp) and DNA 4 (4800 bp) are indicated. In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated.

4.3.3. Probing the nature of metal binding

The metal-dependent DNase activity instigated us to probe the metal binding sites using the intrinsic tryptophan fluorescence. Cas5d structure encompasses two tryptophan residues in the vicinity of each of the active triads (Figure 2.17A in Chapter2). Their nature seems to be conserved across the type I-C organisms as shown in the conservation profile (Figure 2.17B in Chapter 2). Both the tryptophan residues showed varied extent of exposure. W47 is exposed to solvent (relative surface accessibility = 39.9%) while W187 is largely buried (relative surface accessibility = 10.2 %). The addition of increasing amounts of Mg^{2+} ensued quenching of tryptophan fluorescence. This suggests that one of them might get exposed to the solvent or the metal binding site was probably situated closer to it (Figure 4.7A). The buried tryptophan (perhaps W187) getting exposed to the solvent is possible if Cas5d undergoes conformational changes upon metal binding. In other words, it may be inferred that metal binding may induce a localized conformational transitions in Cas5d. To probe the metal binding sites further, we resorted to the Scatchard plot analysis that showed a concave-up curve, suggesting that either there are multiple classes of metal binding sites or the presence of cooperativity (Figure 4.7B).

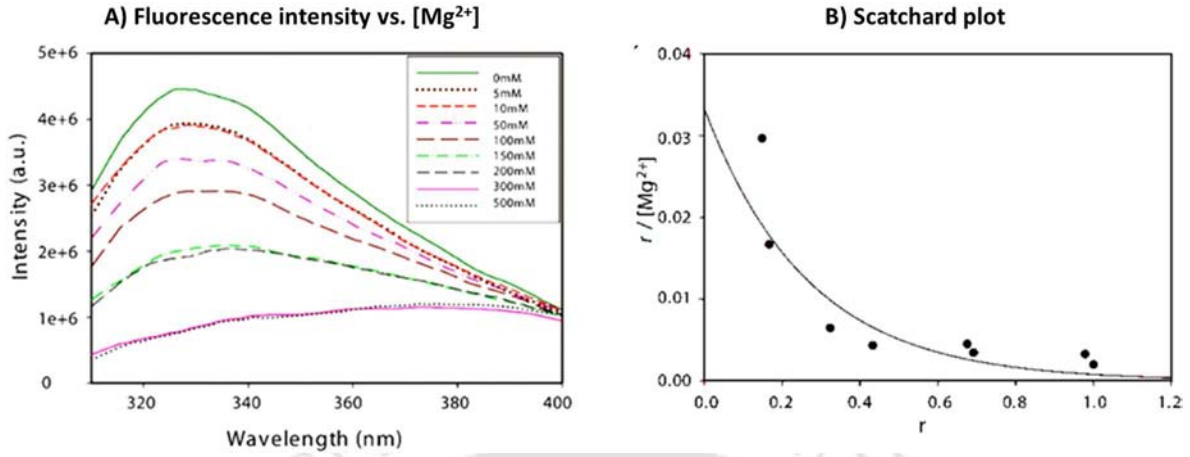


Figure 4.7 ***Fluorescence studies to probe the metal binding.*** (A) Cas5d tryptophan fluorescence undergoes quenching with increasing Mg^{2+} concentration (0-500 mM). The concentration of Mg^{2+} for the corresponding curve is indicated in the inset. The fluorescence intensity is shown in arbitrary units (a. u.). (B) Scatchard plot shows a concave-up curve. Here, r is represented as the ratio of ΔF and ΔF_{max} where ΔF_{max} is the differential intensity at the infinite concentration of Mg^{2+}.

To clarify this, we employed Hill plot analysis that showed a linear trend with the Hill coefficient (n_H) of 0.7 and the apparent dissociation constant (K_d) in the absence of DNA of 35.79 mM (Figure 4.8A). The Scatchard and Hill plot together with Klotz plot analysis allows us to negate the presence of multiple classes of metal binding sites and to propose the possibility of negative cooperativity in binding the metal (Mg^{2+} in this case) in Cas5d (Figure 4.7 and 4.8). Though Cas5d binds metal, in the absence of DNA but the affinity towards metal seems to be weak (k_d = 35.79 mM).

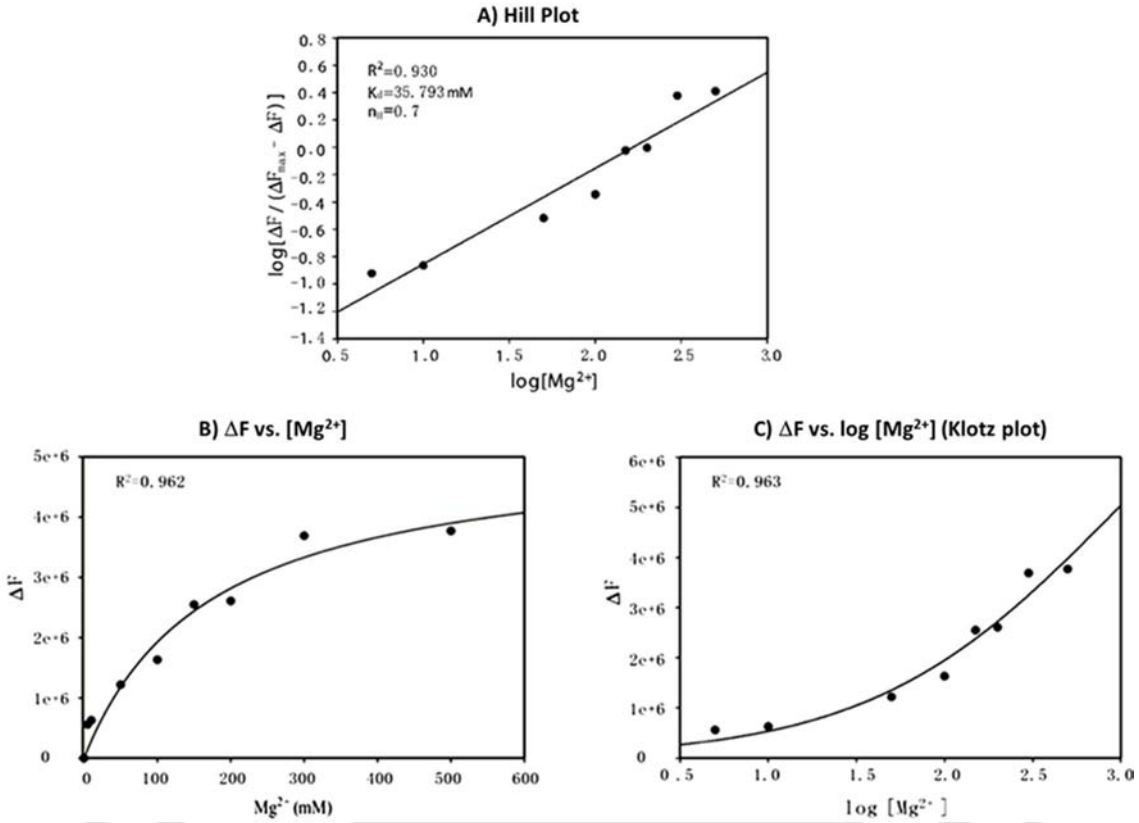


Figure 4.8 ***Analysis on the mode of metal binding to Cas5d.*** (A) Hill plot in the absence of DNA. The linear trend suggests the existence of cooperativity. R^2 represents the square of the goodness of fit, (n_H) denotes the Hill coefficient and K_d indicates the apparent dissociation constant. (B) The plot of ΔF vs. [Mg^{2+}] is shown where ΔF denotes the difference in the fluorescence intensity in the absence of Mg^{2+} (F_o) and presence of varied concentration of Mg^{2+} (F). The pattern is indicative of a hyperbola that tends towards x-asymptote. Here, the rise seems to require wider range of ligand concentration. (C) Klots plot displaying the plot of ΔF vs. log [Mg^{2+}]. Here, the curve depicts a wide sigmoidal pattern.

4.3.4. Probing the link between metal binding and DNase activity

We further tried to investigate the effect on Cas5d-metal binding in presence of DNA. For this, we incubated the DNA with Cas5d in the presence of increasing concentration of metal. The cleavage was assessed and compared with the standard, having DNA of known quantity (Figure 4.9A). From the band intensities, the percentage cleavage activity was calculated and used for Hill plot analysis. Interestingly, the Hill plot analysis in the presence of DNA revealed the affinity for Mg^{2+} (K_d) to be 1.33 μM with the Hill coefficient (n_H) of

0.37 (Figure 4.9B). This in turn, implies that the DNA binding increases the affinity for metal or vice versa.

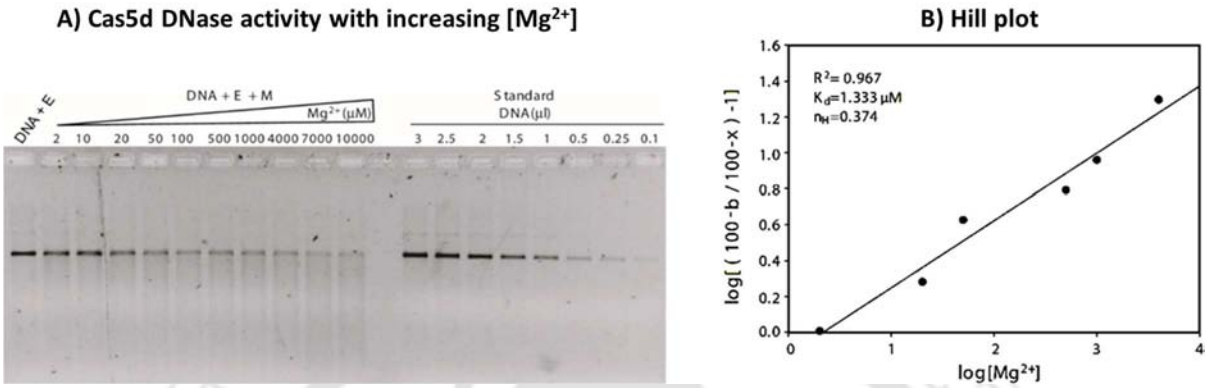


Figure 4.9 ***Probing link between metal binding and DNase activity.*** (A) Effect on Cas5d DNase activity with increasing concentration of metal. In all the lanes Cas5d is represented by E and M represents the metal, Mg^{2+}. Lane 1 represents Cas5d activity in the absence of Mg^{2+}. DNA is incubated with Cas5d with increasing concentration of Mg^{2+} as indicated by a triangle. The known quantity of DNA loaded to make the standard curve is shown on the right. The quantification of the band intensity was made using Kodak Molecular Imaging Software Standard Edition v.5.1.0.27. (B) The dependence of DNA cleavage on Mg^{2+} was analyzed by Hill plot. The maximal activity is taken as 100%. The percentage activity in the absence of Mg^{2+} is b, and x is the percentage activity in the presence of varying concentrations of Mg^{2+}. R^2 represents the square of the goodness of fit, (n_H) denotes the Hill coefficient and K_d indicates the apparent dissociation constant.

Thus, when DNA is present, the affinity seems to be strong (k_d = 1.33 μM), suggesting that the DNA too contributes to metal binding. It is possible that one of the ligands that coordinate the metal may be contributed by the DNA itself as seen in several DNA binding metalloenzymes (Yang et al., 2006). The cellular concentration of magnesium is around 30 mM (Maguire and Cowan, 2002) and hence when the DNA is not in the vicinity of Cas5d, the very low affinity for metal would render it to be an RNase, thus enabling it to facilitate the crRNA maturation. However, proximity to DNA, which might be brought about during the other stages of CRISPR immunity, is likely to enhance the metal affinity (k_d = 1.33 μM) and thus transforming it to be a DNase too.

4.3.5. Tunable DNase activity of Cas5d

Cas5d was shown to selectively process the pre-crRNA leading to the crRNA maturation in a metal-independent manner. Here, we have shown that it also exhibits a metal dependent DNase activity. Because it seems to act against both RNA and DNA substrates, it presents an interesting question whether there is any preference between the two. This was tested against a mixture of RNA and DNA substrates under two different conditions – (1) non-specific substrates and (2) cognate CRISPR repeat RNA and DNA. When metal was not present, Cas5d acted on RNA alone, and when metal was added, DNA cleavage was observed, while the RNase activity remained unaffected. Thus, DNA and RNA were simultaneously acted on in presence of metal, which arises the possibility of harboring a single active site with tunable target selectivity (Figure 4.10). This suggests that the metal bestows a new functionality to Cas5d (DNase activity) even in the presence of RNA.

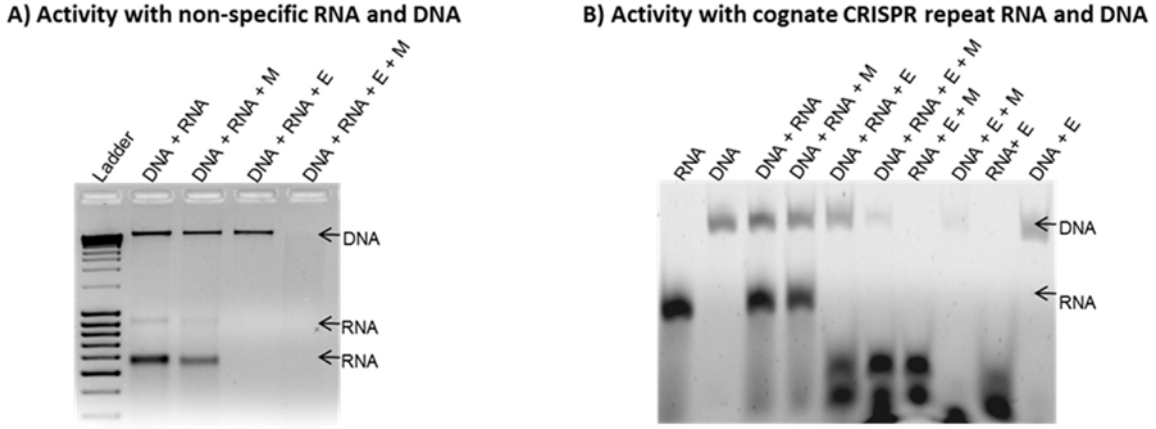


Figure 4.10 ***Metal tunable DNase activity of Cas5d.*** (A) Activity was tested with a mixture of non-cognate RNA and DNA and in the presence and absence of the metal. Two populations of RNA of varied sizes were used and their locations are indicated by an arrow. (B) Nuclease activity on the cognate CRISPR repeat RNA and DNA, respectively, is shown. The DNA substrate has additional T7 promoter (-) sequence and hence its size is larger. This enables to distinguish the DNA and RNA substrates. M indicates the presence of metal, Mg^{2+} and E represents Cas5d.

Because Cas5d exhibits metal-dependent DNase activity even in the presence of RNA, we did time dependent studies on both the substrates to probe its RNase and DNase activity further. We observed that Cas5d showed extended processing of the CRISPR repeat RNA over time (Figure 2.8 in Chapter 2). This sequential processing of repeat RNA provoked us to assess the DNA processing over a long time period. To test this, we performed time dependent assay with DNA in presence of metal and found no accumulation of distinct intermediate or formation of any specific size product, even with longer incubation (Figure 4.11). The Cas5d seems to cleave the DNA to single nucleotide levels.

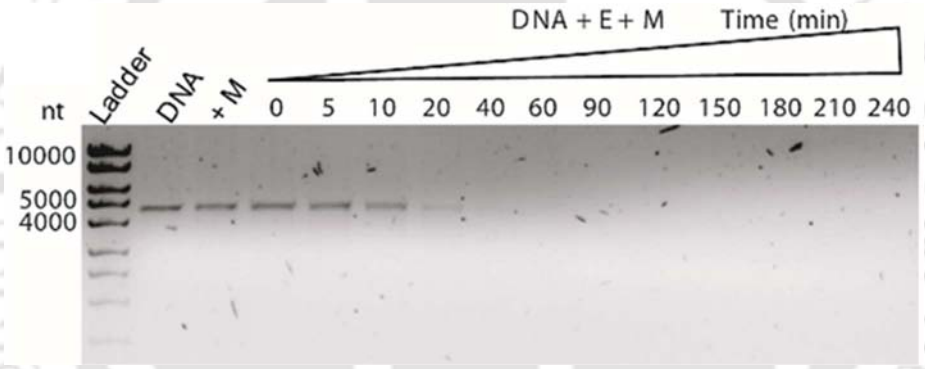


Figure 4.11 ***Time dependent DNase activity of Cas5d.*** In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA is indicated. The lane where the ladder is loaded is shown. Nuclease activity in presence of 10 mM Mg^{2+} with increasing time interval is shown. Almost the entire DNA is cleaved in 40 minutes by Cas5d.

4.3.6. Probing the active site residues of Cas5d involved in DNase activity

To further understand the DNA processing we tried to identify the residues that might be involved in the DNase activity. It was interesting to explore if the residues that were involved in RNase activity (discussed in Chapter 2) can also play a role in DNase activity. Therefore, the mutants of Cas5d encompassing active site triad residues – Y46, K116 and H117 (traid1) as well as the geometrically analogous triad residues – Y35, K39 and H169

(traid2) along with the W47 and W187 residues, which were present adjacent to the respective triad, were inspected for the involvement in DNase activity. When we incubated Cas5d and its mutants Y46F, K116A and H117A with plasmid DNA in presence of metal, we found Y46F which did not have any effect on the RNase activity, seems to be involved in DNase activity, as the mutation abolished the DNase activity of Cas5d. While K116 and H117 residues, in addition to their participation in the RNA hydrolysis, also abrogated the DNase activity, as their mutation (K116A and H117A) rendered Cas5d inactive (Figure 4.12B). Thus, K116 and H117 residues seem to be involved in both RNA and DNA catalysis. For endonuclease activity, the residue Y46 in Cas5d may play a role of a base, deprotonating the 2'-OH of G21 for inline nucleophilic attack on the scissile phosphate. K116 can be the likely candidate to stabilize the negatively charged transition state and H117 may protonate the leaving group akin to K41 and H119, respectively, in RNase A (Raines, 1998). In line with this, the roles of K116 and H117 seem to be apt in satisfying the requirements for hydrolyzing the DNA too. However, for the DNA hydrolysis, the role of a nucleophile activator, Y46, may be taken over by a metal ion since the nucleophile is most likely a water molecule here and not an intrinsic 2'-OH group (Yang et al., 2006). Therefore, this drives us to hypothesize that the active site that promotes RNA hydrolysis may also have the potential to participate in hydrolyzing the DNA too.

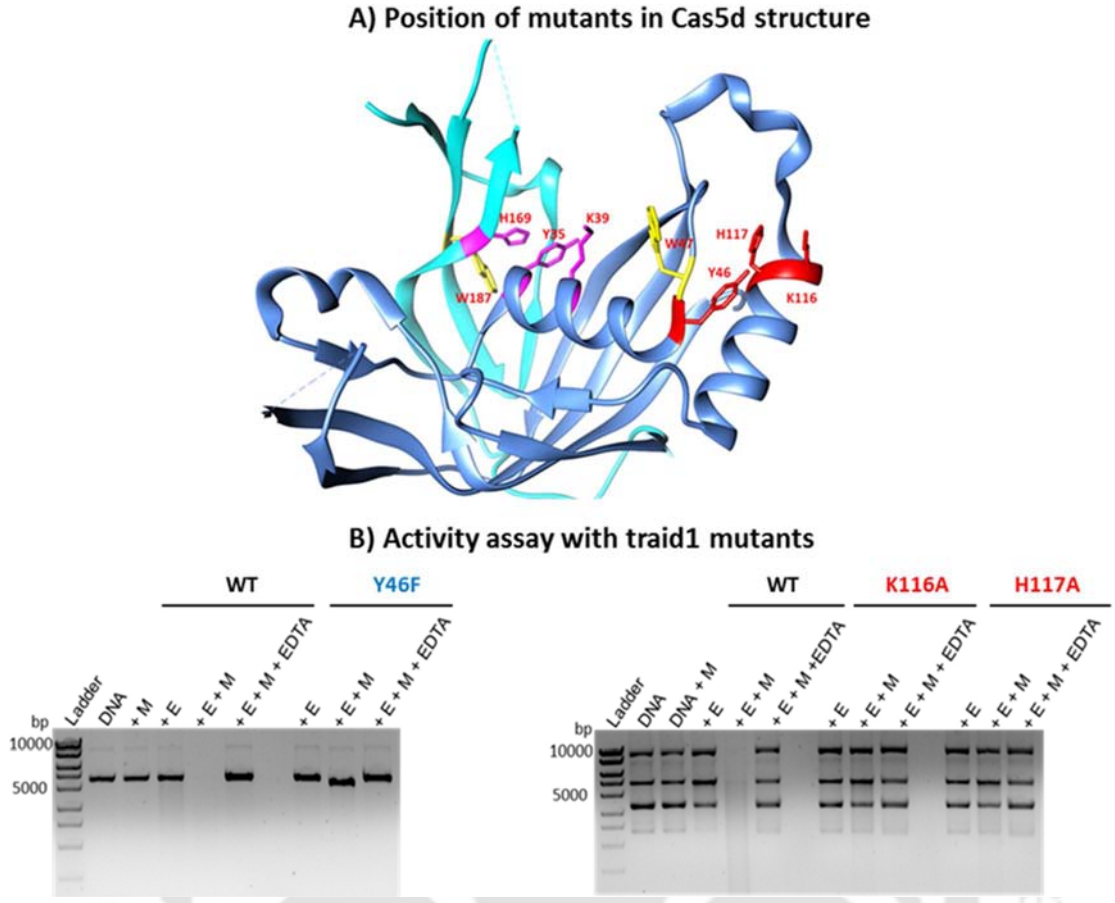


Figure 4.12 ***The effect of mutations in triad1 on the DNase activity of Cas5d.*** (A) The position of all the mutants in Cas5d structure is shown and the N-terminal domain is shown in blue while the C-terminal domain in cyan (PDB ID: 4F3M). (B) The effect of mutations in triad1 on the DNase activity of Cas5d is shown. In all panels, M denotes Mg^{2+} and E denotes the presence of enzyme. WT represents the wild type Cas5d and the respective mutants are shown at the top. The lanes containing the control DNA and EDTA are shown. The residues labelled in red show their involvement in RNase activity too. The residues impacting only DNase activity are shown in blue.

An interesting phenomenon observed earlier with H169A mutant of the active triad2 of Cas5d comprising of Y35, K39 and H169 residues was that it showed drastic reduction in RNase activity when metal was added, which otherwise was equally active as wild type Cas5d (Figure 2.20 in Chapter 2). Whereas the other mutants have shown the retardation in extended processing of RNA substrates in presence of metal. This prompted us to examine the phenomena as it hinted towards the possible involvement in DNase activity. In order to test this, we incubated the plasmid DNA with Cas5d mutants Y35F, K39A and H169A and

found reduced activity with K39A while Y35F was as active as the wild type. Interestingly, the H169A mutant showed complete loss of DNase activity even in the presence of metal (Figure 4.13). This suggests the involvement of H169A in the DNase activity and also provides the plausible reason for the drastic reduction that was observed in the RNase activity in presence of metal. The metal thus seems to tune the activity of Cas5d towards DNA substrates.

We also investigated the role of W47F and W187F mutants present in the vicinity of the active triads of Cas5d. W47F mutant, which had slightly retarded the product conversion in case of RNA, also seems to affect the DNase activity. The impairment in DNase activity is clearly visible (Figure 4.13B). The W187F mutant that had no effect on RNase activity seems to affect DNase activity slightly (Figure 4.13B).

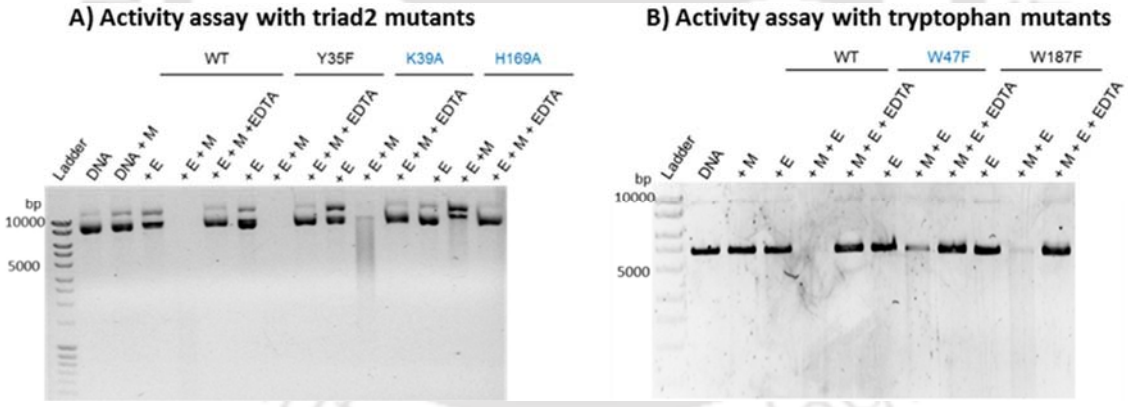


Figure 4.13 ***The effect of mutations in triad2 and the two tryptophan residues on the DNase activity of Cas5d.*** In all panels, M denotes Mg^{2+} and E denotes the presence of enzyme. WT represents the wild type Cas5d and the respective mutants are indicated. The lanes having the control DNA and EDTA are shown. A) DNase activity assay with point mutants Y35F, K39A, H169A of triad2 is shown. B) DNase activity assay with the tryptophan mutants W47F and W187F is shown. The residues impacting DNase activity are shown in blue.

While inspecting the Cas5d structure for DNA binding and metal coordinating residues, we spotted E4, D56, H98 and E100 as prospective candidates owing to the propensity of these residues to coordinate the metal ligands and their clustered location

(Figure 4.14A). Among the identified residues, D56, H98 and E100 seem to be highly conserved across the Cas5d orthologs (Figure 4.14B). So, we performed alanine scan mutation of these candidates and tested the impact on the DNase activity of Cas5d. The selected mutants were cloned into a suitable vector and purified for the DNase activity assays (Figure 4.14C).

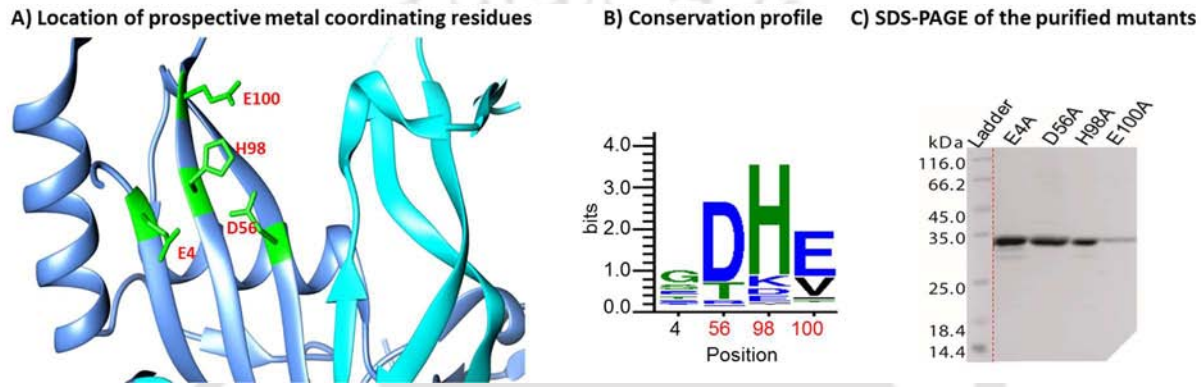


Figure 4.14 ***Prospective metal coordinating residues of Cas5d.*** (A) Cas5d structure displaying prospective residues involved in metal coordination is shown (PDB ID: 4F3M). The N-terminal is shown in blue and C-terminal in cyan. The residue positions are indicated. (B) The sequence logo depicting the conservation of these residues across type I-C orthologs is shown. The residue number is indicated below the logo. The height of a residue (in bits) represents the extent of conservation. (C) The purified mutants are shown. The discontinuity in gel is indicated by red dotted line.

The prospective metal binding residues E4A, D56A, H98A and E100A were incubated with plasmid DNA and the effect on DNase activity was monitored. The DNase activity of E4A remained unaffected while H98A showed marginal reduction in the activity (Figure 4.15A). However, D56A and E100A mutations drastically affected the activity suggesting their involvement in coordinating the metal or binding the DNA. Inquisitively, we also tested these residues for their effect on RNase activity using 3' 6-FAM labelled repeat RNA as substrate. While E4A and D56A had no effect on RNase activity, H98A showed reduced activity, which was further inhibited in the presence of metal. E100A exhibited RNase activity which was severely impaired in the presence of metal (Figure 4.15B). This

phenomenon was similar to the case of H169A mutant which had shown drastic reduction in RNase activity in presence of metal. Thus, it can be inferred that the residue function is tuned by metal to bestow additional functionality.

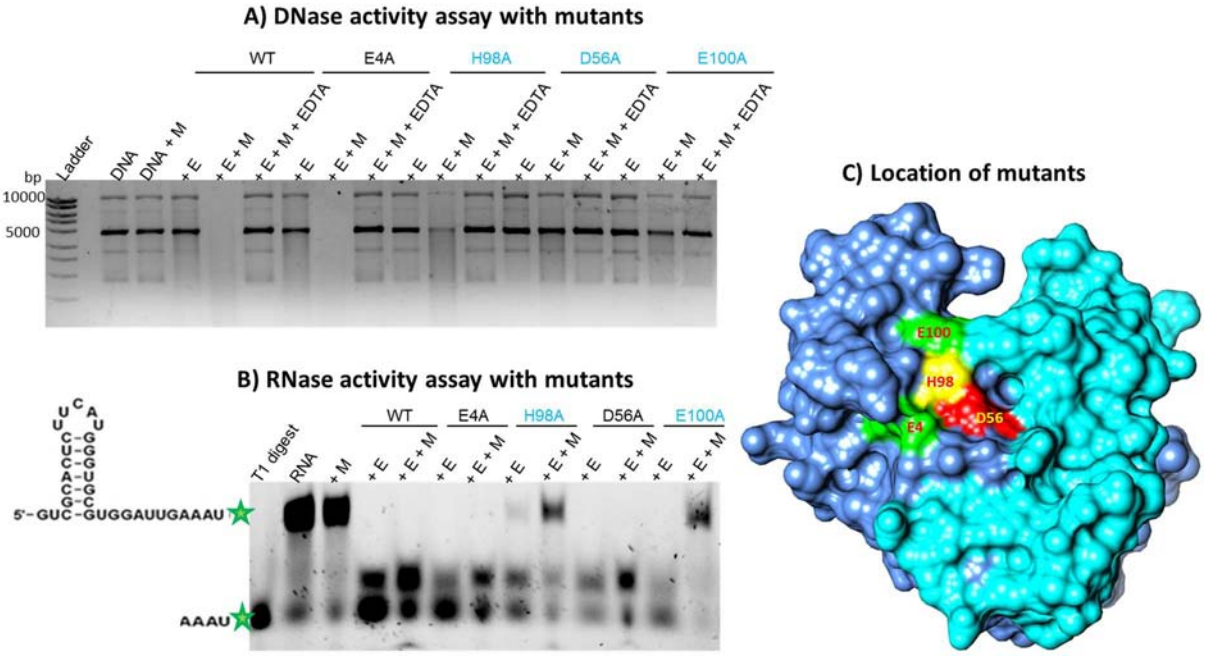


Figure 4.15 ***Effect of mutation of metal coordinating residues on Cas5d nuclease activity.*** In all panels, M denotes Mg^{2+} and E denotes the presence of enzyme. WT represents the wild type Cas5d and the respective mutants are indicated. The lane that has substrate as control is denoted as DNA or RNA. The lane containing EDTA is indicated. (A) DNase activity assay of mutants E4A, D56A, H98A and E100A. The residues impacting the DNase activity are shown in blue. (B) Assay to monitor the effect of these mutants on the RNase activity. The lane containing T1 digest is indicated. (C) Surface of Cas5d (PDB ID: 4F3M) displaying the position of residues that are proposed to be involved in DNA binding or metal coordination is shown. The residue positions are indicated. The N-terminal is shown in blue and C-terminal in cyan. This figure was rendered using Chimera (Pettersen et al., 2004).

4.3.7. Segregation of mutants based on their nuclease activity

The various activity assays performed with the aforementioned mutants allowed us to segregate their involvement in either DNA or RNA hydrolysis, or in some cases, their involvement in both. Remarkably, the mutational analysis of K116A and H117A that affects

both DNase and RNase activity led us to suggest the possibility that the catalytic site for processing the DNA and the RNA is interrelated (Figure 4.16C).

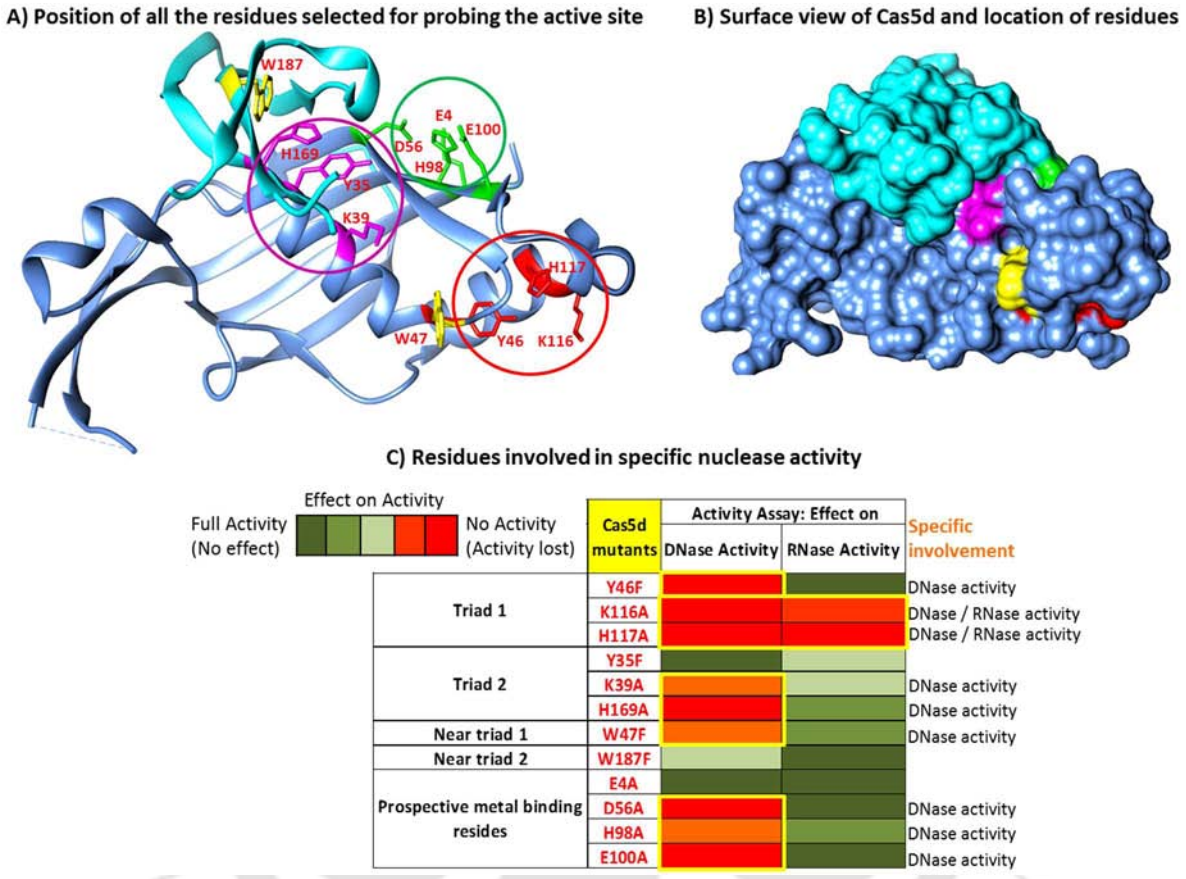


Figure 4.16 ***Segregation of Cas5d mutants based on their involvement in nuclease activity.*** (A) The position of all the residues used for probing the active site of Cas5d is shown (PDB ID: 4F3M). The N-terminal is shown in blue and C-terminal in cyan. The triad1 is shown in red, the triad2 in magenta, the metal coordinating residues in green and the two tryptophans in yellow. (B) The position of the residues on Cas5d surface is shown. Most of the residues seem to cluster around the N-terminal region in Cas5d structure. (C) The specific nuclease activity of the residues is compared in tabular form. The extent to which the activity is affected is shown using the colour code as shown.

The mutational analysis of Cas5d indicates that D56, H98, E100 and H169 residues are involved in the DNase activity either by contributing to DNA binding or by coordinating the metal ion. Though the mutations D56A, E100A and H169A render Cas5d incompetent against DNA, the activity against RNA is retained. Thus, these represent the separation-of-function mutants. Intriguingly, when metal is provided, the RNase activity of H98A, E100A

and H169A is drastically reduced. While the functional roles of these residues in processing the RNA and DNA targets remain to be examined, this mutational analysis raises the possibility of a considerable functional overlap of the residues involved in processing the DNA and RNA substrates.

4.3.8. Investigating the conservation of active site residues across type I systems

After finding the residues involved in Cas5d nuclease activity, we tried to investigate the presence of these active residues in Cas5 across the other type I systems. We aligned Cas5d with Cas5e, Cas5a, Cas5t and Cas5h representing the respective subtypes – type I-C (Cas5d from *B. halodurans*), type I-E (Cas5e *E. coli*), type I-A (Cas5a from *A. pernix*) and type I-B (Cas5t from *P. furiosus* and Cas5h from *T. maritima*) (Figure 4.17). The residues involved in nuclease activity of Cas5d seem to be absent or substituted in other Cas5 across type I CRISPR-Cas systems. This explains why Cas5d is active, while other Cas5 (Cas5e, Cas5a, Cas5t and Cas5h) play only the structural role.

Figure 4.17 ***Sequence analysis of Cas5d across type I subtypes.*** Sequence alignment of Cas5 from subtypes – type I-C (represented by Cas5d from *B. halodurans*), type I-E (represented by Cas5e from *E. coli*), type I-A (represented by Cas5a from *A. pernix*), type I-B (represented by Cas5t from *P. furiosus*) and type I-B (represented by Cas5h from *T. maritima*) is shown. The secondary structure is assigned on the basis of Cas5d structure (PDB ID: 4F3M). The Cas5d mutants that have been employed to probe the nuclease activity are shown in red box along with the equivalent residues in other subtypes. This figure was prepared using JALVIEW (Waterhouse et al., 2009).

4.4. Summary

We found Cas5d to exhibit a moonlighting function by harbouring a metal dependent DNase activity. This activity is in addition to its endoRNase activity. Interestingly, when both RNA and DNA substrates are provided, Cas5d shows DNase activity in the presence of metal, irrespective of the RNA present together. The metal seems to serve as a switch to its DNase activity. The DNase activity is non-specific and leads to complete degradation of substrate, which is in contrast to RNase activity that generates particular cleavage product. Another interesting feature shown by Cas5d is revealed by the mutational analysis, which sheds light on the interplay of residues in both RNase and DNase activity.

Thus, Cas5d in type I-C system that has adopted the role of endoRNase in absence of Cas6 (that processes the CRISPR RNA in other type I systems), also possess additional functionality of being a metal dependent DNase. This hints at the possibility of its involvement in other stages of CRISPR-Cas immunity. Both adaptation and interference stages can recruit DNase having promiscuous restriction. It can be either tuned to fragment the invader genome for spacer uptake or can be employed to degrade the target effectively during interference. Thus, Cas5d seems to have evolved to play multifarious roles in CRISPR-Cas immunity.

5.1. Introduction

The CRISPR-Cas system of type I, III and IV appear to utilize multi-protein assemblies to facilitate the crRNA maturation and/or target interference, while in type II and V the stand-alone nucleases, Cas9 and Cpf1, respectively, are responsible for target cleavage (Brouns et al., 2008; Fonfara et al., 2016; Hale et al., 2009; Jore et al., 2011; Makarova et al., 2015; Wiedenheft et al., 2011b; Zhang et al., 2012). In type I, III and IV CRISPR systems the multi-protein assemblies along with guide RNA form the ribonucleoprotein surveillance complex, which scans the cell for target and ensues the degradation. Not much is known about the recently identified type IV CRISPR system (Makarova et al., 2015), but type I and type III systems have been explored to an appreciable extent. Though there seems to be no significant sequence similarity between the type I and III Cascade-like complex, the genome architecture and synteny of the *cas* operon suggests that both these systems are recognized with contextual semblance. They are found to have minimal components comprising of a large subunit (Cas8/Cse1 in type I and Cas10/Csm1 in type III), a small unit (Cse2 in type I and Csm2 in type III), a backbone subunit (Cas7/Cse4 in type I and Csm3 in type III), Cas5 and Cas6 (Makarova et al., 2011a). Among the type I systems, the type I-E is most extensively characterized both in terms of functional and structural aspects (Brouns et al., 2008; Hayes et al., 2016; Horvath and Barrangou, 2010; Jackson et al., 2014; Jore et al., 2011; Makarova et al., 2013; Mulepati et al., 2014; Wiedenheft et al., 2011a; Zhao et al., 2014). The type I-E Cascade complex in *E. coli* comprises of five types of proteins that participate in the crRNA maturation and target interference *viz.*, Cse1, Cse2, Cse4, Cas5e and Cas6e. It utilizes Cse1 to stabilize the 5'-end of crRNA and Cse2 to further assist the binding to crRNA. Cse4 is used in forming the backbone to bind and display the crRNA and Cas5e further assists in anchoring of 5'-end of crRNA. Cas6e is employed for crRNA processing,

which after processing remains bound to the 3' handle of crRNA (Figure 2.1 in Chapter 2). In contrast to type I-E system, the Cascade complex of type I-C comprises of only three types of proteins *viz.*, Csd1, Csd2 and Cas5d. Thus, it seems to have minimal components to form the Cascade, which raises questions on the functionality of each Cascade components. Out of these three type I-C Cascade proteins, we have already characterized the function of Cas5d, which we identified to be a nuclease with dual functionality – metal independent RNase and metal dependent DNase. This was again in contrast to type I-E system, where Cas5e seems to play only the structural role, which can be attributed to the absence of the catalytic residues of Cas5d at the equivalent position in Cas5e (Figure 4.17 in Chapter 4). Thus, there seems to be considerable adaptation in Cas5d to offset the absence of Cas6e in type I-C system. Since the type I-C Cascade has two other subtype specific proteins namely Csd1 and Csd2, in addition to Cas5d, we were prompted to investigate their role in type I-C system. Therefore, in this chapter, we have tried to characterize Csd1 and Csd2 and also explored the effect on the individual activities when they form the Cascade.

5.2. Materials and methods

5.2.1. Cloning, expression and purification

The Cas5d was purified as described in Chapter 2. Genes encoding *csd1* and *csd2* were amplified from *B. halodurans* genomic DNA using gene specific primers with Pfu DNA polymerase (Fermentas). Amplicon of *csd1* were cloned into pQE2 to create pCsd1 using the restriction sites for NdeI and PstI and that of *csd2* in modified pET23a to create pCsd2 using the restriction sites for NdeI and EcoRI (New England Biolabs). Constructs that were cloned in pQE2 harbor an N-terminal $(His)_6$ tag and the one in modified pET23a harbor a C-terminal

strep-tag II. The cloned constructs were verified by sequencing. Expression was performed in *E. coli* BL21(DE3) by growing the cells in LB medium supplemented with ampicillin (100 µg/ml) or kanamycin (50 µg/ml) at 37°C until OD at 600 nm reached 0.7. The temperature was then reduced to 20°C for 20 min and protein expression was induced by the addition of 0.2 mM IPTG followed by incubation at 20°C overnight. The cells were harvested by centrifugation and resuspended in buffer A containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6 mM β-ME and 1 mM PMSF. After sonication, the lysate was clarified by centrifugation at 36,500g for 30 min. The supernatant was treated with RNase to remove any bound RNA and then loaded onto a 5 ml HiTrap IMAC HP column or StrepTrap HP column (GE Healthcare) pre-equilibrated with buffer B containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6 mM β-ME. After washing the column with buffer C containing 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 6 mM β-ME and 40 mM imidazole, the bound protein was eluted using a linear gradient of imidazole (upto 500 mM) in buffer C. For strep-tagged proteins, washing buffer contained 20 mM Tris-HCl (pH 8.0), 500 mM NaCl and 6 mM β-ME and the elution was carried out with buffer C that contained 2.5 mM D-Desthiobiotin in place of imidazole. The eluted protein was incubated with 10 mM EDTA for 1 hr to remove the bound metal ions if any and then dialyzed against buffer D containing 20 mM Tris-HCl (pH 8.0), 200 mM NaCl and 6 mM β-ME. Subsequently, the proteins were aliquoted, snap frozen in liquid nitrogen and stored at -80°C until required.

5.2.2. Reconstitution of type I-C Cascade

The *in vitro* reconstitution was done by incubating the purified proteins Cas5d, Csd1 and Csd2 for 30 minutes in ice and then passed through 100 kDa membrane cut off filter

(Merck). The unbound or unassociated proteins passed out through the centrifugal filter as the molecular weight of Cas5d, Csd1 and Csd2 are 27, 72 and 31 kDa, respectively. The retentate was then washed several times with buffer containing 20 mM Tris-HCl (pH 8.0), 200 mM NaCl and 6 mM β-ME, so that unbound and loosely bound protein gets filtered out. The retentate was then concentrated and eluted in buffer containing 20 mM Tris-HCl (pH 8.0), 200 mM NaCl and 6 mM β-ME. Aliquots were snap frozen in liquid nitrogen and stored at -80°C until required.

Alternatively, Cascade-crRNA complex was co-expressed and purified from *E. coli*, which was generously provided by Siddharth Nimkar, a PhD student in the lab. To achieve this, the polycistronic region encoding Cas5d, Csd1 and Csd2 were inserted into LIC vector 1R (a kind gift from Scott Gradia, Addgene ID: 29664) having a pET backbone and encodes a TEV cleavable Strep-tag II at the N-terminus of Cas5d. Six copies of identical repeat-spacer units, synthesized from GeneScript, were inserted into 13S-R plasmid (a kind gift from Scott Gradia, Addgene ID: 48328) using XhoI and KpnI restriction sites. Both the plasmids were used to transform *E. coli* BL21(DE3) cells. Co-expression was performed by growing the cells in LB medium supplemented with spectinomycin (100 µg/ml) and kanamycin (50 µg/ml) at 37°C until OD at 600 nm reached 0.7. The temperature was then reduced to 18°C for 20 min and protein expression was induced by the addition of 0.2 mM IPTG followed by incubation at 18°C for 14 hr. The cells were harvested by centrifugation and resuspended in buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1mM Dithoithreitol (DTT) and 1 mM PMSF. Cells were lysed by sonication and the lysate obtained was clarified by centrifugation at 25,000g for 30 min. The supernatant was loaded onto a StrepTrap HP column (GE Healthcare) pre-equilibrated with buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM NaCl and 1 mM DTT. The column was washed with same buffer and the bound complex was eluted in a buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM

DTT and 2.5 mM D-Desthiobiotin. Concentration and removal of unbound Cas5d was performed using 100 kDa cut off membrane filter (Merck). The aliquots of eluted Cascade were snap frozen in liquid nitrogen and stored at -80°C until required.

5.2.3. Preparation of substrates

Pre-crRNA containing only the repeat sequence was chemically synthesized and differently end-labelled with a 5' HEX and 3' 6-FAM (IDT). The DNA sequences corresponding to the CRISPR repeat with an additional T7 promoter (-) sequence were obtained from IDT and subsequently labelled with fluorescein tagged dUTP using deoxy terminal transferase (New England Biolabs) at the 3'-end. pQE2 or pUC19 plasmid was employed as circular DNA and pQE2 linearized with KpnI served as linear substrate. Single stranded circular M13mp18 phage DNA was obtained from New England Biolabs.

5.2.4. Nuclease activity assays

All pre-crRNA processing reactions were performed at 37°C for 1 hr. The 5' HEX or 3' 6-FAM labelled pre-crRNA repeat at 0.2 µM concentration were incubated with 2 µM of any of the required proteins including Csd1, Csd2, Cas5d and Cascade in 20 mM Tris-HCl (pH 8.0), 100 mM KCl and 6 mM β-ME. The cleavage products were analyzed on 15% (w/v) denaturing urea PAGE.

DNase activity assays were performed with double stranded (linear or circular) and single stranded (circular) DNA at 37°C for 1hr in the buffer containing 20 mM Tris-HCl (pH 8.0), 100 mM KCl, 6 mM β-ME, 10 mM $MgCl_2$ and 2 µM Cas5d. Time dependent nuclease

activity was performed at 37°C and samples were taken at the indicated time intervals. The reaction was stopped using 50 mM EDTA (pH 8.0) and the products were analyzed on 0.8% agarose gel and visualized by ethidium bromide staining.

5.3. Results and discussion

5.3.1. Characterization of Csd1

Csd1 is identified to be the large subunit of Cascade/type I-C. While comparing the sequence length of Csd1, we noted that it was longer than the Cse1 of type I-E system (627 vs. 502 aa). This prompted us to probe whether the additional region at the C-terminus can exist as a separate domain in Csd1. Therefore, we subjected this region to fold recognition using FFAS server (Jaroszewski et al., 2011) which predicted that this region indeed could be a separate domain and showed similarity to Cse2 in *T. thermophilus* (Figure 5.1). We predicted the secondary structure using JPRED (Cole et al., 2008) which suggested that the C-terminal region (554-627 aa) is largely α-helical as seen in Cse2. This points at the likelihood that Csd1 could be a fusion protein of its functional counterparts in *E. coli viz.*, Cse1 and Cse2.

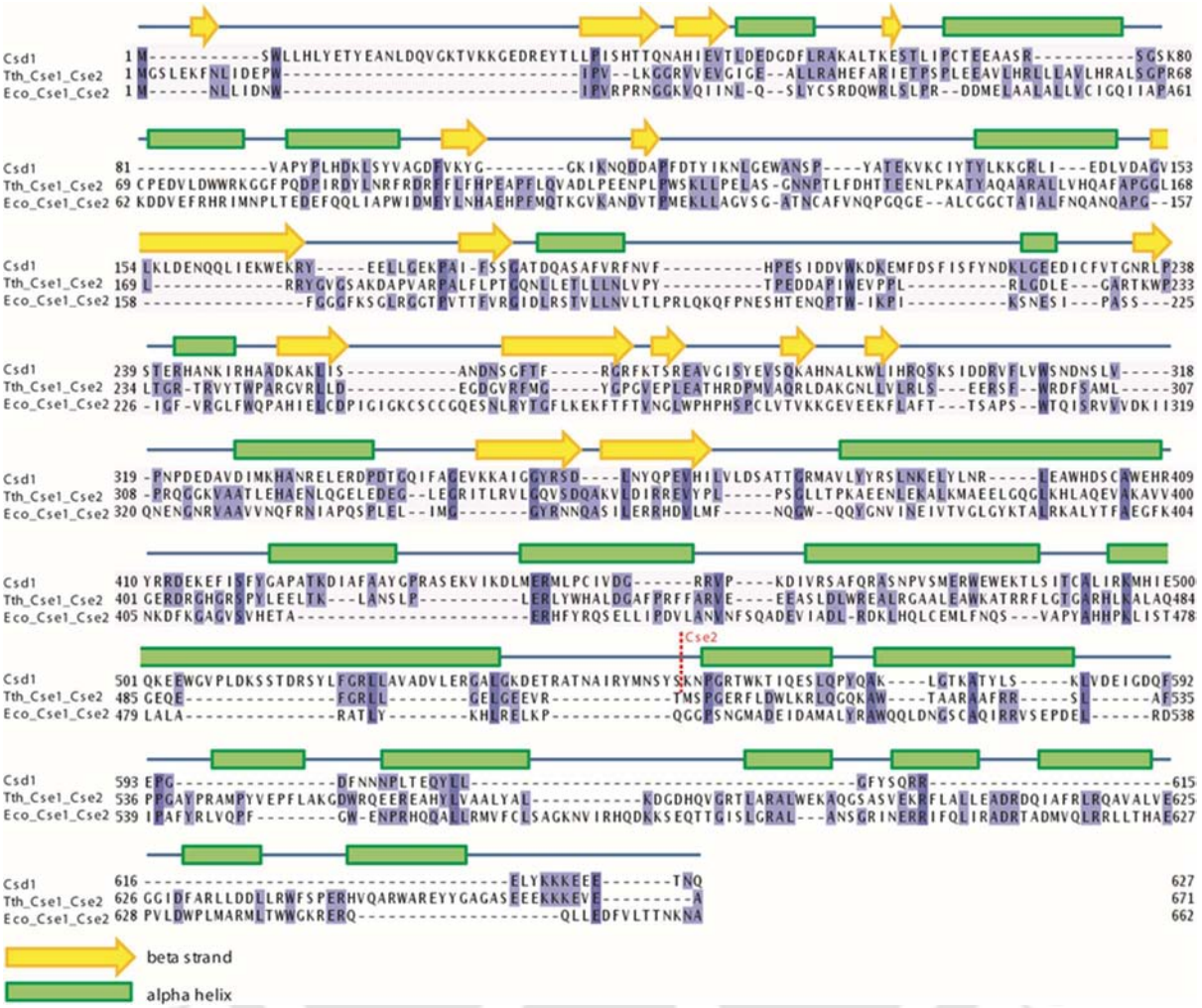


Figure 5.1 ***Sequence analysis of Csd1.*** The multiple sequence alignment of Csd1 from *B. halodurans* with Cse1 and Cse2 from *T. thermophilus* and *E. coli* is shown. For clarity, the Cse1 and Cse2 sequences are merged and the start of the *T. thermophilus* Cse2 is indicated by a dotted red line. The secondary structure is assigned based on the 3D-structure of *T. thermophilus* Cse1 (PDB ID: 4AN8) and Cse2 (PDB ID: 2ZCA). The conserved residues are highlighted in blue. This figure was prepared using JALVIEW (Waterhouse et

To understand the function of this large subunit in type I-C, Csd1 was cloned into suitable vector and purified to carry out functional studies (Figure 5.2A). Interestingly, the gel filtration profile of Csd1 showed two peaks, both of which corresponded to Csd1 as showed in SDS-PAGE (Figure 5.2B and 5.2C). Based on the elution volume the second peak corresponds to ~72 kDa, which seems to contain the monomeric form of Csd1 while the first peak corresponds to ~160 kDa which seems to be an oligomeric form of Csd1 encompassing

two units. This suggests that Csd1 can exist in dimeric form as well, in contrast to the preliminary reports of Cascade composition where it is shown to present in single copy (Nam et al., 2012).

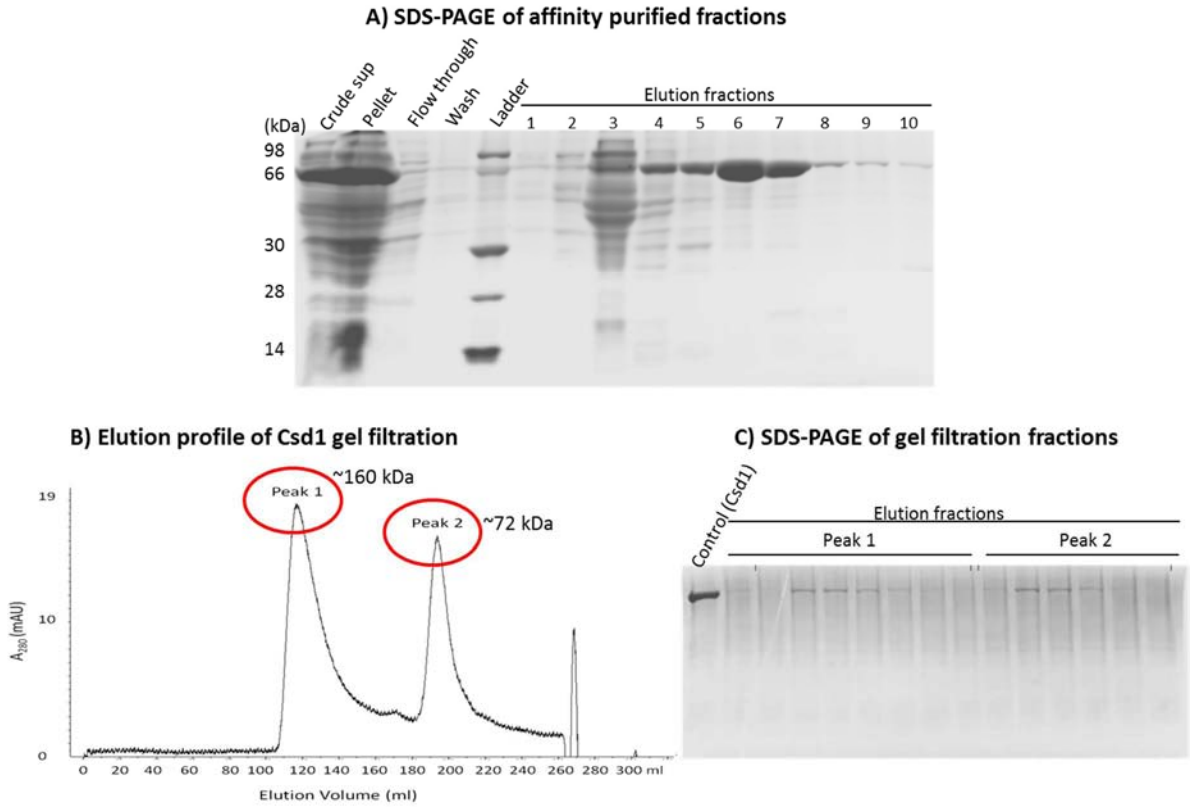


Figure 5.2 ***Csd1 purification.*** (A) The SDS-PAGE of the affinity purified fractions of Csd1 is shown. The fraction 6 and 7 were pooled and taken for further purification. The lanes containing the ladder and control samples are indicated. (B) The gel filtration profile of Csd1 obtained from calibrated HiLoad 26/60 Superdex 200 prep grade column (GE healthcare) is shown. The two peaks observed are marked with approximate molecular weight of the eluted fraction, based on the elution volume. (C) The SDS-PAGE containing fractions from both the peaks is shown. Fractions obtained from both the peaks contained Csd1. The lane containing Csd1 as control is indicated.

To understand the functionality of Csd1, we further queried its sequence against Conserved Domain Database (CDD), which showed it to be a multidomain protein belonging to Cas8 superfamily. The other members of Cas8 family include Zinc Finger Nucleases (ZFNs) and PALM domain polymerases. We also inspected the sequences of the large subunit in type III, which seems to harbor motifs that are reminiscent of a palm domain, that

are typically found in polymerases, but the equivalent subunit in type I shows inactivated polymerase domain (Makarova et al., 2011a). While comparing the sequence of Csd1, we came across Nar71 (MTH1090), a Csd1 ortholog in *Methanothermobacter thermoautotrophicus*, which was reported to have nuclease activity (Guy et al., 2004). But when we compared the sequence of Csd1 with its active ortholog Nar71 to locate the corresponding position and conservation of the active site residues, we found that the active residues of NAR71 were not present in Csd1, suggesting considerable evolution (Figure 5.3). Owing to the highly dynamic nature of the CRISPR-Cas system, the Cas proteins usually show high diversity in their sequence. Assuming that Csd1 might show functional conservation with NAR71, we proceeded to investigate the potential of Csd1 for being a nuclease and tested its activity against both RNA and DNA substrates.

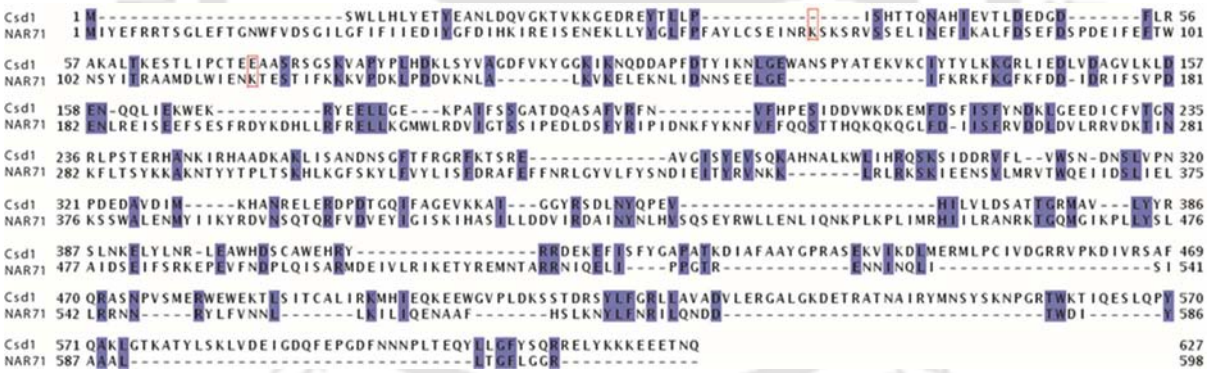


Figure 5.3 ***Sequence comparison of Csd1 with its ortholog Nar71.*** The residues shown to be involved in the nuclease activity of Nar71 (Guy et al., 2004) are shown in red box. It may be noted that K68 that is shown to be important for the ATP hydrolysis and not for the nuclease activity in Nar71 is absent in Csd1 and K117 which has been shown to be required for both ATP hydrolysis and nuclease activity in Nar71 is substituted by a Glu in Csd1. The conserved residues in both the proteins are highlighted in blue. This figure was prepared using JALVIEW (Waterhouse et al., 2009).

5.3.1.1. Investigating the RNase activity of Csd1

To test the RNase activity, we incubated Csd1 with 3' 6-FAM labelled repeat RNA at 37°C for 1 hr and monitored the band pattern using T1 digested substrate as reference. The

RNase T1 which is a single stranded G-specific nuclease cleaves all available Gs in the single stranded region of the repeat RNA. Thus, complete digestion will result in cleavage at G28 position of 32 nt repeat RNA, which will correspond to a length of 4 nt on gel as the repeat RNA is 3'-end labelled (similar to mapping in Chapter 2). The incubation of Csd1 with the end-labelled RNA substrate resulted in a band that was smaller in size than the substrate, as shown by the differences in the migrated position in the denaturing PAGE. This suggests the possibility of substrate getting processed by Csd1. The band seemed similar to T1 digested product of the labelled repeat RNA that corresponds to 4 nt (Figure 5.4). Thus, Csd1 seems to exhibit RNase activity.

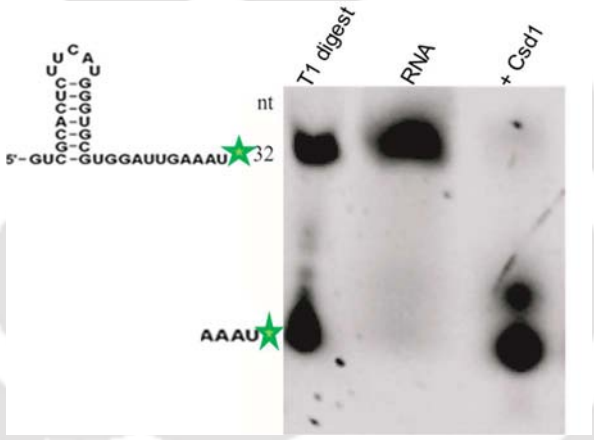


Figure 5.4 ***Csd1 exhibits RNase activity.*** Csd1 is incubated with 3'-end labelled repeat RNA in the absence of metal for 1 hr at 37°C. The lanes containing the control RNA and T1 digest are indicated. The folded repeat RNA with 3' 6-FAM label indicated by star is shown.

Motivated by this, we set out to ask if the trajectory of the product formation resulted from the RNA processing by Csd1 was similar to that of Cas5d. For this, we tested the activity of Csd1 against the 3'-end labelled repeat RNA for different time intervals and mapped the products using RNase T1 and alkaline hydrolysis ladder. When we incubated Csd1 with 3' 6-FAM repeat RNA for 20 minutes faint bands appeared initially, which became prominent over time (Figure 5.5). We observed three bands, two migrated above and

one migrated along the product of T1 digestion. The initial band corresponded to the cleavage at G23, followed by a band that seems to be a result of cleavage at G24 and the smaller band that migrated along with T1 digest corresponded to the cleavage at G28 of the repeat. Since we observed the smaller size band becoming prominent with longer incubation, it can be inferred that Csd1 processes the repeat sequentially overtime, resulting in conversion of longer product to a smaller sized product. The pattern of the mapped cleavage products seemed similar to that obtained for Cas5d, suggesting similarity in the point of cleavage (Figure 2.8 in Chapter 2). Thus, Csd1 seems to be another endoRNase in type I-C system that is able to process the CRISPR repeats.

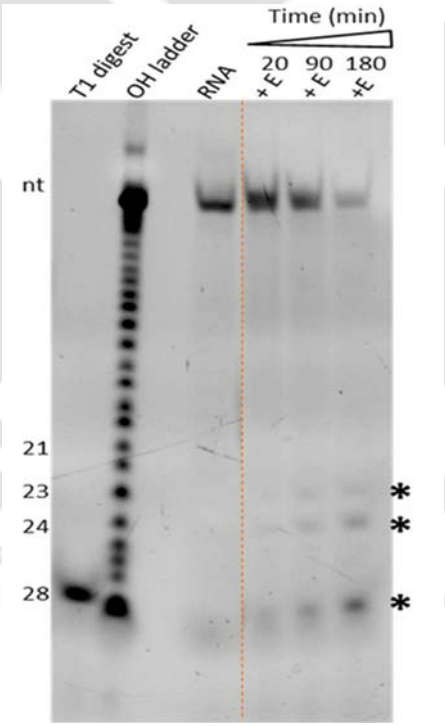


Figure 5.5 ***Mapping the Csd1 cleavage products of repeat RNA.*** In all panels, E represents Csd1. The presence of 3' 6-FAM labelled RNA, T1 digest and OH ladder in the respective lanes is indicated. When incubated with the labelled repeat RNA at 37°C for 20 min, only faint bands are observed. The intensity of the three bands increased with time. Two bands migrated above and one along with the T1 digest. The larger fragments seem to be a result of cleavage at G23 and G24 resulting in 9 nt and 8 nt fragments, respectively, while the shorter fragment seems to be a result of cleavage at G28 producing 4 nt fragment. * denotes the cleavage products. Red dotted line indicates the discontinuity in gel.

5.3.1.2. Investigating the DNase activity of Csd1

To test the DNase activity of Csd1, we employed various forms of DNA as substrates and incubated with Csd1 in presence and absence of metal. To start with we utilized linearized plasmid DNA as substrate and incubated with Csd1 in presence and absence of metal at 37°C and monitored the activity. Surprisingly, the band corresponding to DNA in presence of Csd1 showed drastic decrease in intensity, while there was no band in presence of metal but when EDTA was added to the reaction, no change in the intensity of the band was observed. This suggests that Csd1 seems to possess the DNase activity that is enhanced in presence of divalent metal ion (Figure 5.6). Next we utilized circular DNA with Csd1 and observed decrease in the band intensity, though some portion seems to be retarded in the well, possibly by binding to Csd1. The observed decrease in the intensity and also the trail below the band, suggests the degradation of DNA. Interestingly, there was no band observed in the lane that contained circular DNA with Csd1 in presence of metal ion but when EDTA was added this nuclease activity was inhibited (Figure 5.6). Also, the plasmid preparation showed polymorphism in its mobility that is typical for a circular DNA. Csd1 showed no preference for these and all of them seem to be digested with no traces of DNA left behind (Figure 5.6). This suggested that Csd1, indeed, exhibits endodeoxyribonuclease activity that is stimulated in the presence of a divalent metal. Encouraged by this, we asked whether it is adept at acting on single stranded DNA as well. Hence we employed the single stranded circular DNA from M13mp18 phage for the assay. When we incubated single stranded circular DNA with Csd1 we observed a band of same intensity as that of the control DNA, suggesting no degradation. In presence of metal the band seems to be shifted upward and was found stuck in the well, but when EDTA was added the band appeared at the same position as

that of the control DNA (Figure 5.6). This suggests the possibility of Csd1 binding to single stranded circular DNA only in presence of metal. Thus, Csd1 seems to be unable to process single stranded circular DNA even in presence of a divalent metal ion.

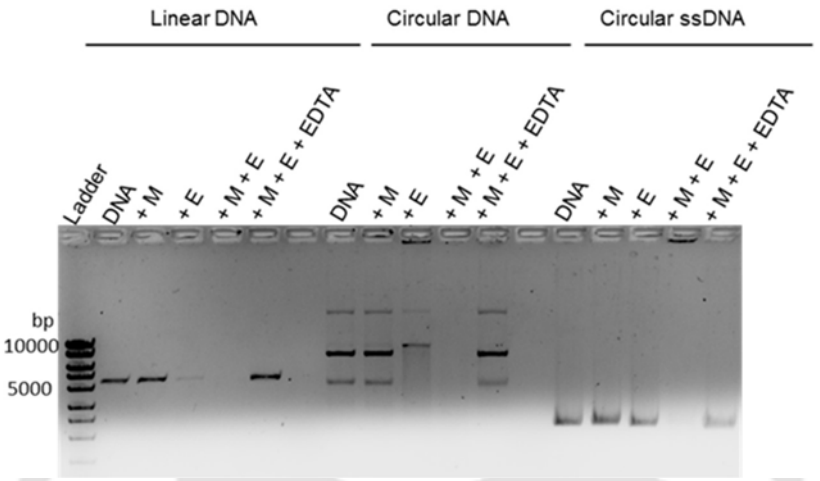


Figure 5.6 ***Activity of Csd1 against different forms of DNA.*** The DNase activity of Csd1 was assayed in the presence of double stranded linear, double stranded circular and single stranded circular DNA. In all panels, E represents Csd1 and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown. An upward shift of single stranded circular DNA in the presence of Csd1 and metal is noted which suggests binding as against the cleavage.

Thus, we found Csd1 to possess DNase activity that is enhanced in presence of divalent metal ion. However, Csd1 seems to be biased towards processing the double stranded DNA substrates, and shows only the binding to single stranded substrates in presence of metal. This is in contrast to Cas5d which recognizes and cleaves all forms of DNA in presence of metal (Punetha et al., 2014).

Initially, we had found that Csd1 was able to recognize and process the structured form of CRISPR RNA. Prompted by this, we asked whether the DNA recognition too is structure specific. For that, we utilized the sense, the antisense and the duplex forms of the CRISPR repeat DNA to clarify whether the sequence is recognized and cleaved in a manner similar to RNA. Both the sense and antisense showed the presence of stem and loop when

subjected to the fold prediction using MFOLD (Zuker, 2003) (Figure 5.7A). When we incubated Csd1 with sense, antisense and the duplex forms of CRISPR DNA we observed the drop in the intensity of all in presence of metal, suggesting that Csd1 cleaved all forms preferentially in the presence of the divalent metal (Figure 5.7B). Unlike the CRISPR repeat RNA where a single cleavage is shown to occur at positions G23, G24 and G28, there seemed to be no such preferential cleavage of the DNA substrates (Figure 5.7B). Intriguingly, though there was no cleavage against the circular single stranded DNA (Figure 5.6), the complementarity within the cognate CRISPR repeat DNA that would result in the formation of double stranded stem region, presumably could have activated the DNase activity or it is selectively inactive against the single stranded circular DNA. This corroborates that Csd1 displays sequence-independent nuclease activity against double stranded DNA substrates that is stimulated in the presence of the divalent metal ion.

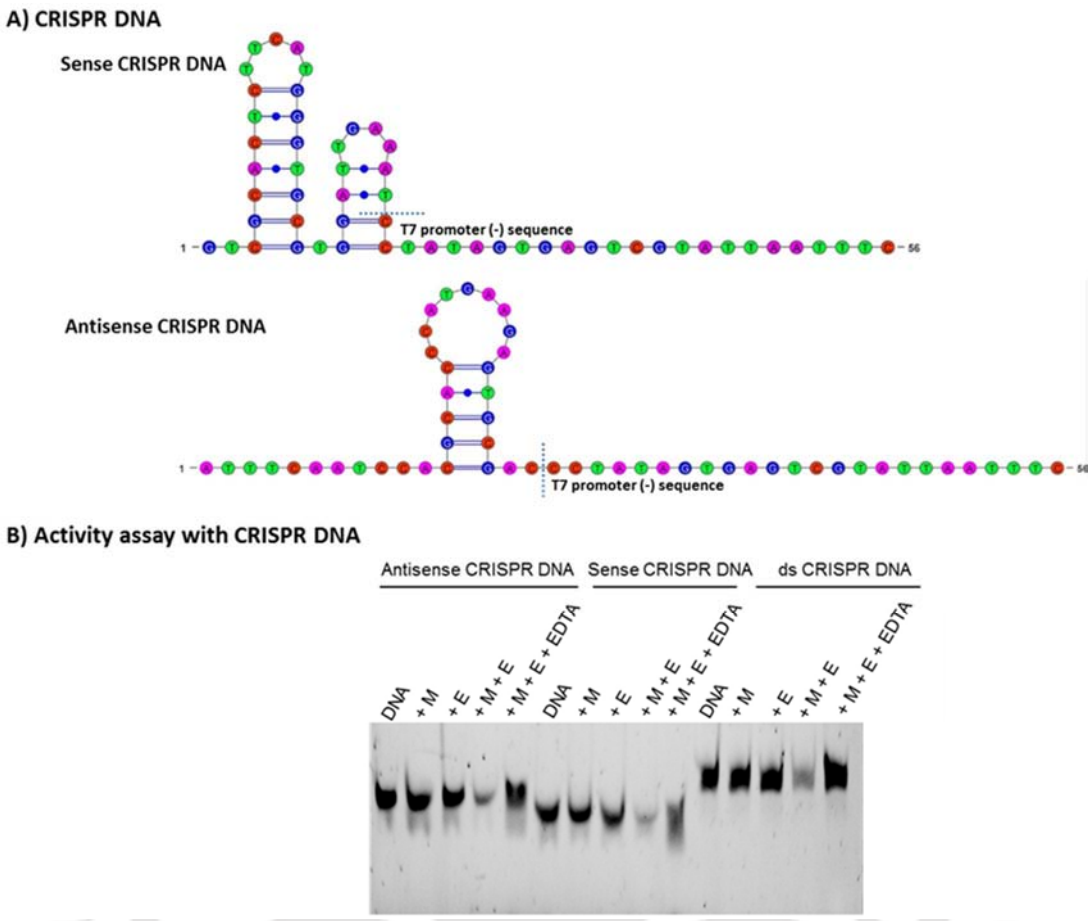


Figure 5.7 ***Csd1 activity against CRISPR DNA.*** (A) The sense and antisense CRISPR DNA with the additional T7 promoter sequence (-) is shown. The base Adenine is shown in purple, Thymine in green, Guanine in blue and Cytosine in red. The folds were predicted using MFOLD (Zuker, 2003) and the figures were prepared using VARNA (Darty et al., 2009). (B) The DNase activity of Csd1 was assayed in the presence of the single stranded antisense and sense strands as well as the duplex of sense and antisense CRISPR repeat DNA. In all panels, E represents Csd1 and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown.

In case of RNA substrates, we had observed extended processing of repeat RNA by Csd1 into specific products over time and this provoked us to assess the DNA processing over long time period. To test this, we performed time dependent assay. We incubated Csd1 with DNA in presence and absence of metal for different time intervals and monitored the pattern. In the absence of metal, the intensity of the band decreased sharply after 40 minutes of incubation and after which only slight reduction was observed with longer incubations. But the drop in the band intensity was more in presence of metal and within 60 minutes there was

no band observed, suggesting the complete degradation of DNA. There was no distinct product or intermediate accumulation even at longer incubation (Figure 5.8). Thus, the Csd1 seems to cleave the DNA to single nucleotide levels.

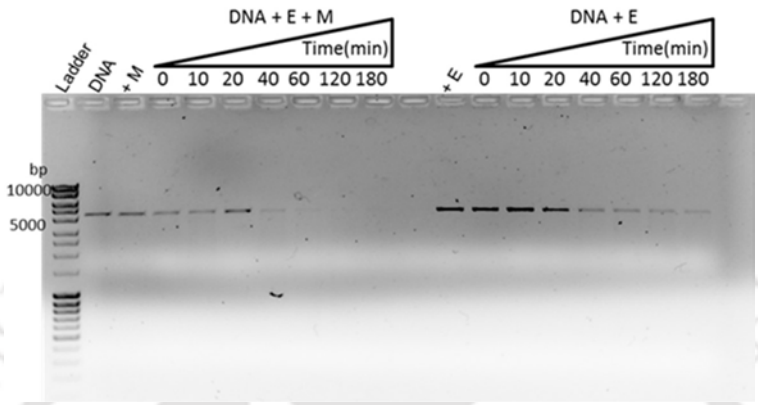


Figure 5.8 ***Time dependent DNase activity of Csd1.*** In all panels, E represents Csd1 and M denotes Mg^{2+}. The lanes containing DNA and ladder are indicated. Nuclease activity in presence of 10 mM Mg^{2+} with increasing time interval is shown. Csd1 cleaved almost entire DNA in 60 minutes in presence of metal.

5.3.2. Characterization of Csd2

Csd2 seems to belong to Cas7 family, the members of which are mostly RNA binding proteins. This RNA binding can have two implications, either it can have enzymatic role in processing the RNA or the structural role by providing the scaffold. The scaffold proteins are used in forming the backbone of multi-protein assemblies. Thus, Csd2 can be a nuclease or like Cas7 in type I-E, it may form the backbone of Cascade. Therefore, in order to understand the function of Csd2, we cloned it in suitable vector and purified it to perform the functional studies. The affinity purified fractions of Csd2 (Figure 5.9A) were pooled and taken further for gel filtration. Interestingly, the gel filtration profile of Csd2 showed two peaks, both of which corresponded to Csd2 when analyzed on SDS-PAGE (Figure 5.9B and 5.9C). Based on the elution volume, the second peak corresponds to ~31 kDa which seems to contain the

monomeric form of Csd2 while the first peak corresponds to ~90 kDa, which seems to be an oligomeric form of Csd2 encompassing three units. This suggests that Csd2 can exist in trimeric form as well, in contrast to the preliminary reports of Cascade composition where it is shown to be present in six copies (Nam et al., 2012). There is a possibility that the hexamer can form either by association of two trimeric forms or six single monomeric forms.

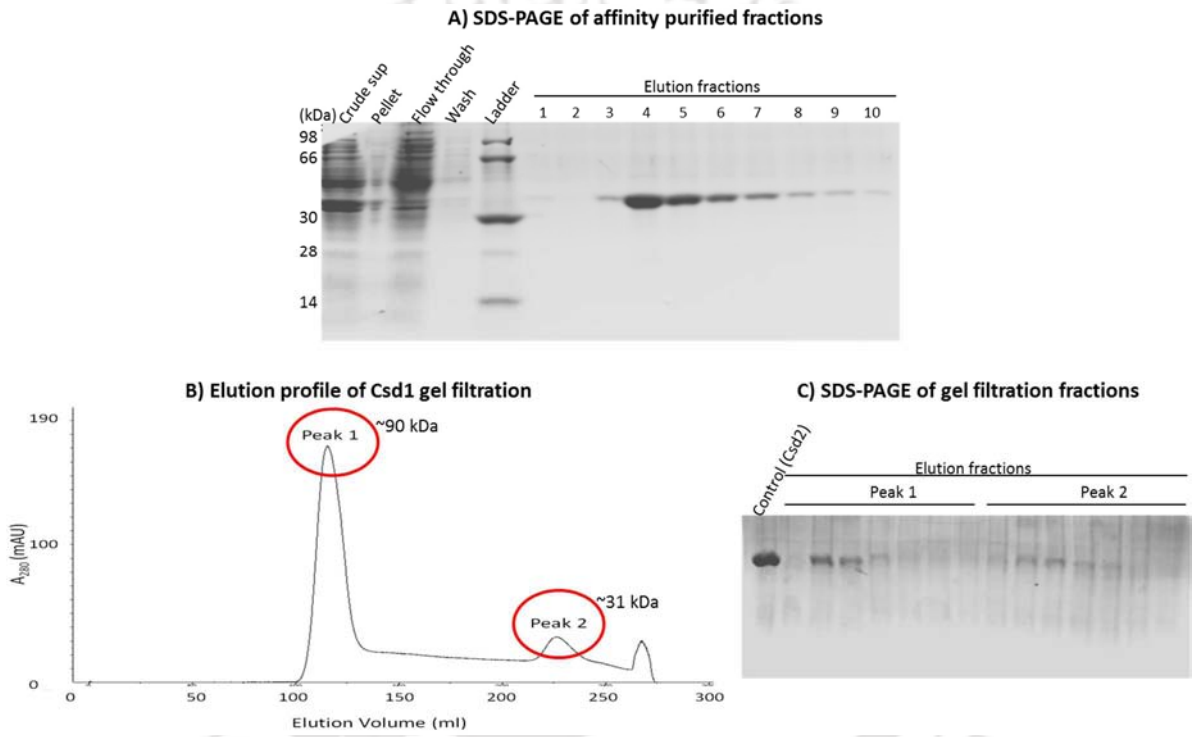


Figure 5.9 ***Csd2 purification.*** (A) The SDS-PAGE of the affinity purified fractions of Csd2 is shown. The obtained fractions were pooled and taken for further purification. The lanes containing the ladder and control samples are indicated. (B) The gel filtration profile of Csd2 obtained from calibrated HiLoad 26/60 Superdex 200 prep grade column (GE healthcare) is shown. The two peaks observed are marked with approximate molecular weight of the eluted fraction, based on the elution volume. (C) The SDS-PAGE containing fractions from both the peaks are shown. Fractions obtained from both the peaks contained Csd2. The lane containing Csd2 as control is indicated.

5.3.2.1. Investigating the RNase activity of Csd2

To test the nuclease activity of Csd2, we incubated it with 3' 6-FAM labelled repeat RNA at 37°C for 1 hr. We observed no cleavage of RNA when Csd2 was added, which

suggests that the repeat RNA is not processed. In contrast to this, Cas5d and Csd1 which were used as a control, produced a lower sized band migrating similar to T1 digest as anticipated, which indicates the processed product (Figure 5.10). Thus, unlike Cas5d and Csd1, Csd2 seems to possess no RNase activity.

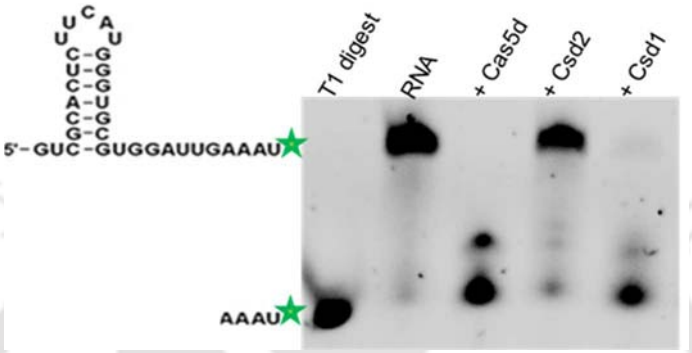


Figure 5.10 ***Activity assay to test Csd2 RNase activity.*** The purified Csd2 is incubated with 3'-end labeled repeat RNA in the absence of metal for 1 hr at 37°C. Cas5d and Csd2 RNase activity is shown for comparison. The lanes corresponding to the control RNA and T1 digest are indicated. The folded repeat RNA with its 3' 6-FAM label represented by star is shown.

5.3.2.2. Investigating the DNase activity of Csd2

Since we found the type I-C cascade components, Cas5d and Csd1 to possess metal dependent DNase activity, we were prompted to examine the behavior of the third protein component, Csd2 towards DNA substrates. To probe its nuclease activity, we utilized various forms of DNA substrate including double stranded DNA in both linear and circular form and single stranded circular DNA from M13mp18 phage. When we incubated Csd2 with various forms of DNA in absence or presence of metal at 37°C for 1 hr, we observed no change in the intensity or position of bands in any of the conditions. Thus, there seems to be no cleavage occurring in any of the DNA forms (Figure 5.11). Only a slight upward shift of single stranded circular DNA in the presence of Csd2 was observed, that might suggest binding as against the cleavage.

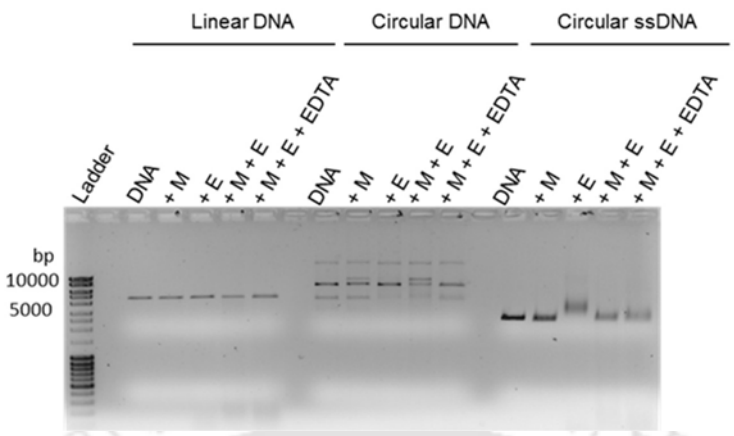


Figure 5.11 ***Activity assay to test Csd2 DNase activity.*** The DNase activity of Csd2 was assayed in the presence of double stranded linear DNA, double stranded circular DNA and M13mp18 phage single stranded circular DNA. In all panels, E represents Cas5d and M denotes Mg^{2+}. The presence of DNA and EDTA in the respective lanes is indicated. The lane where the ladder is loaded is shown. Slight upward shift of single stranded circular DNA in the presence of Csd2 might suggest binding as against the cleavage.

Thus, Csd2 seems to be inactive against DNA substrates too. To confirm this and also to observe any processing that might occur over time, we incubated the plasmid DNA with Csd2 for longer duration both in presence and absence of metal. But we could not observe any significant change in the intensity or position of the bands even with longer incubation (Figure 5.12), suggesting that Csd2 possess no DNase activity. Thus, in the absence of both RNase and DNase activity, Csd2 possibly plays only the structural role in Cascade.

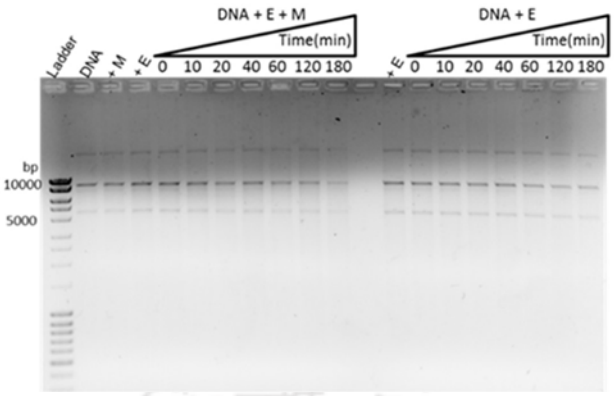


Figure 5.12 ***Time dependent assay to probe the nuclease activity of Csd2.*** In all panels, E represents Csd2 and M denotes Mg^{2+}. The lanes containing the control DNA and ladder are shown. Nuclease activity in presence of 10 mM Mg^{2+} with increasing time interval is shown. Csd2 did not cleave DNA even after 3 hrs.

5.3.3. Characterization of type I-C Cascade

The type I-C Cascade in *B. halodurans* consists of three components *viz.,* Cas5d, Csd1 and Csd2. Since we have already investigated the functionality of these individual proteins and found Cas5d and Csd1 to be active nucleases against both RNA and DNA substrates, we further questioned the functionality of these proteins as a part of the complex. Do they retain their nuclease activity as a complex or the activity gets modulated in the presence of other interacting partners? To address this, we set out to reconstitute the Cascade complex using the purified Cas proteins and then test the activity against RNA and DNA substrates. We explored both *in vitro* and *in vivo* reconstitution of the Cascade/type I-C complex.

5.3.3.1. Attempts of in vitro reconstitution of type I-C Cascade-like complex

We tried *in vitro* reconstitution of the complex in the absence of crRNA by incubating the individually purified proteins Cas5d, Csd1 and Csd2 (Figure 5.13A) for 30 minutes in ice

and then passed through membrane filter having 100 kDa cut off. The unbound or unassociated proteins passed out through the centrifugal filter as the molecular weight of Cas5d, Csd1 and Csd2 are 27, 72 and 31 kDa, respectively. The eluted complex was electrophoresed in denaturing conditions to observe the presence of Cas5d, Csd1 and Csd2. Interestingly, the bands corresponding to all the individual proteins were observed suggesting that the reconstituted complex preserves the integrity of the Cas proteins (Figure 5.13B). The complex thus formed is tested for the nuclease activity to observe the impact on the individual activities as a part of complex.

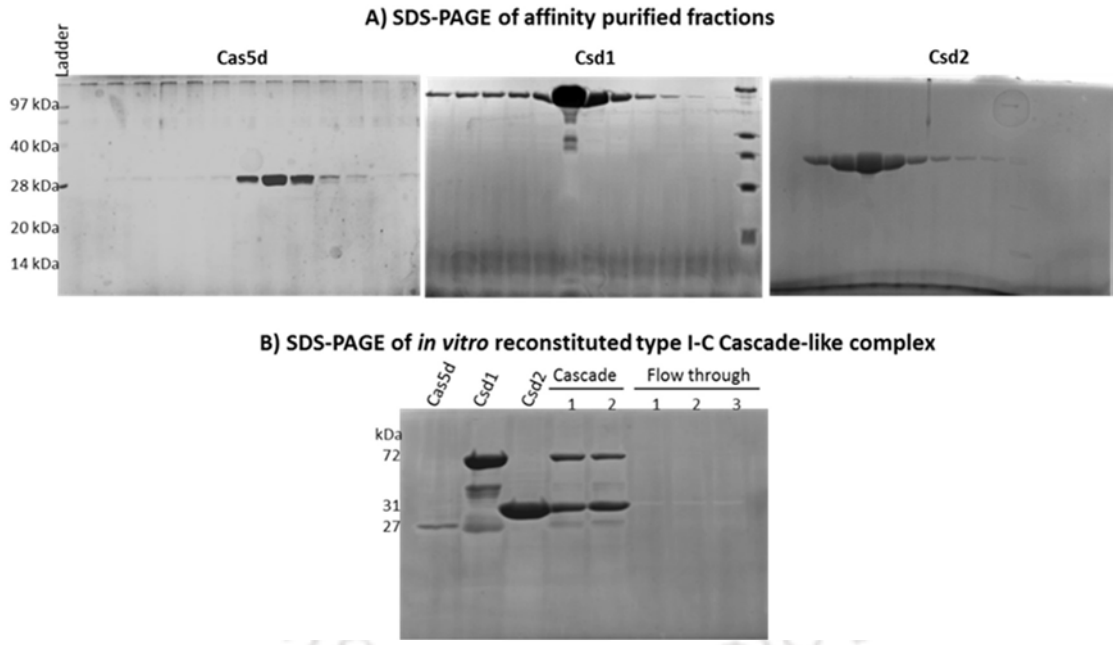


Figure 5.13 ***In vitro reconstitution of type I-C Cascade-like complex.*** (A) The SDS-PAGE of affinity purified Cas5d, Csd1, Csd2 having molecular weight of 27, 72 and 31 kDa, respectively, are shown. (B) Cascade-like complex formed from the *in vitro* reconstitution is shown. The lanes containing Cas5d, Csd1, Csd2 and the reconstituted complex are labelled. The retentate in the membrane filter retained the complex while the filtrate, labelled as flow through, didn't contain the complex.

5.3.3.2. Investigating the RNase activity of in vitro reconstituted complex

To test the nuclease activity, we used both 5' and 3' end-labelled CRISPR repeat RNA as substrate and incubated it with the Cas5d, Csd1 and Csd2 in different combinations along with the *in vitro* formed complex and monitored the effect. When we incubated 5' HEX and 3' 6-FAM labelled RNA with the various combination of individual proteins, we observed that the combination of Cas5d and Csd1 resulted in a prominent smaller size band, which migrated near T1 digest and seemed similar to the band from their individual activity. Also, another band just above this smaller band was seen, which was inferred to be an intermediate during the processing. But when Csd2 associated with Cas5d, the intermediate band seemed to be more prominent in intensity than the smaller sized band, suggesting that Csd2 might have slowed down the Cas5d activity. This effect of Csd2 was seen more dramatic when it associated with Csd1, in which the larger sized band was shown with higher intensity than the smaller sized band. This suggests that Csd2 slowed down Csd1 activity more effectively. When all the three proteins, Cas5d, Csd1 and Csd2 were present together, two smaller sized bands as compared to substrate band were present, referring to the intermediate and the final product. Thus, the activity seems to be a combined effect of the individual activity of the components (Figure 5.14). In conjunction with its inertness against RNA, whenever Csd2 associates with active nucleases Cas5d and Csd1, it appears to slow down their RNase activity. This led us to posit the possibility that Csd2 may down regulate the RNase activity of Cas5d and Csd1 when they assemble into the Cascade-like complex.

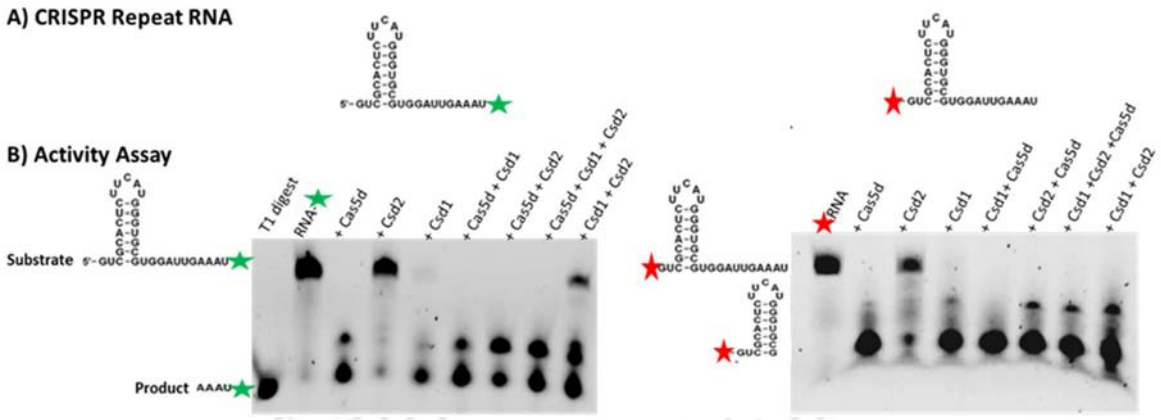


Figure 5.14 ***Combined effect of Csd1, Csd2 and Cas5d against RNA substrate.*** (A) The 3' 6-FAM and 5' HEX labelled repeat RNA used as substrate is shown. (B) The RNase activity assay performed with individual proteins in various combinations is shown. The lanes containing the control RNA and the proteins are marked. The position of the label in the substrate RNA is indicated by star, green showing the 3' 6-FAM and red showing 5' HEX labelling.

5.3.3.3. Investigating the DNase activity of in vitro reconstituted complex

Since we had found the metal dependent DNase activity in Cas5d and Csd1, we tried to examine the effect when they associate together along with the inert Csd2 to form a complex. It was interesting to explore whether the individual DNase activity is still retained after the complex formation. For, this we utilized the sense CRISPR DNA end-labelled with fluorescein tagged dUTP using deoxy terminal transferase (New England Biolabs) at the 3'-end. When we incubated the labelled CRISPR repeat DNA with Cas5d and Csd1 complex in presence of metal, we observed drastic decrease in the band intensity. Similar was the case, when Cas5d was incubated with Csd2, suggesting that DNase activity of Cas5d is not affected by these associations. The incubation of Csd1 and Csd2 together showed decrease in the band intensity, which was similar to the band intensity when Csd1 is present alone. The incubation of the labelled repeat DNA with the complex containing all three proteins showed drastic decrease in the band intensity (Figure 5.15). Thus, the activity seems to be a combination of activities of individual components, albeit predominated by Cas5d.

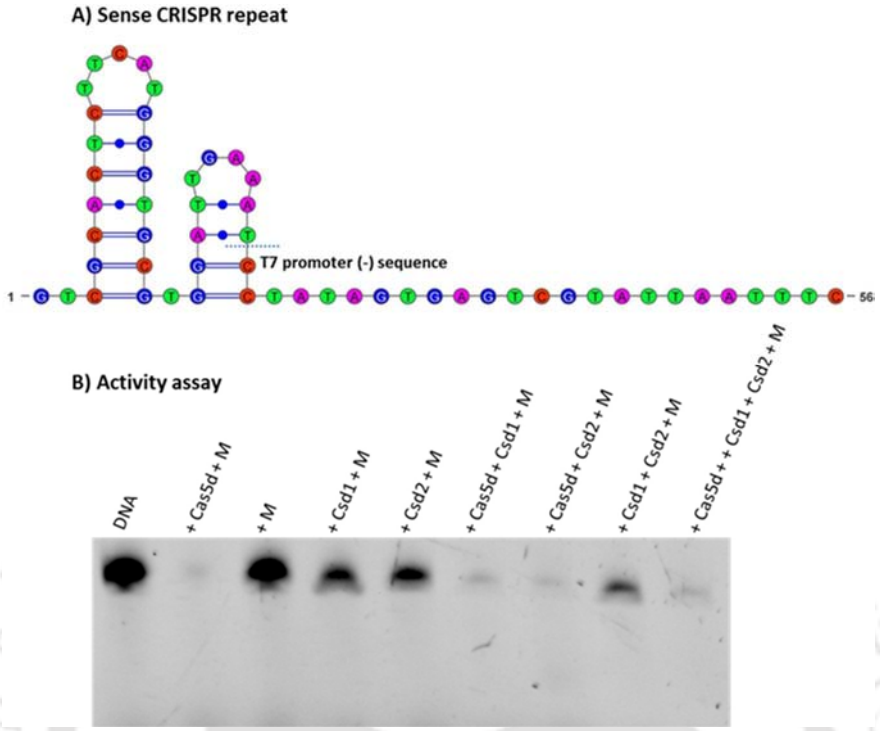


Figure 5.15 ***Combined effect of Csd1, Csd2 and Cas5d activity against DNA substrate.*** (A) The CRISPR repeat DNA with additional T7 promoter (-) sequence, which is used as substrate is shown. (B) The DNase activity assay is shown. In all lanes, M denotes Mg^{2+} and the lanes containing the respective enzymes are indicated. The lane containing the DNA and metal is shown. While Cas5d is proficient in cleaving the single stranded DNA, Csd1 shows weak cleavage while Csd2 seems inert.

Thus based on our assay, both the DNase and the RNase activity of the *in vitro* reconstituted complex seems to be a sum total of the individual activities of the components (Punetha et al., 2014). This raised the question on the actual persistence of these activities inside the cell and its utility. The promiscuous DNase activity can be a potential threat to the genome, if not regulated. Since this complex lacks crRNA, we also questioned whether the *in vitro* reconstituted complex is actually formed in the correct stoichiometry or structure as present in nature or this *in vitro* complex of associated proteins is different from the complex formed *in vivo*. Therefore, to address these questions, we tested the nuclease activity with the *in vivo* reconstituted type I-C complex.

5.3.3.4. Investigating the RNase activity of in vivo reconstituted Cascade

We utilized the *in vivo* reconstituted type I-C Cascade complex (see 5.2 Materials and method) and end-labelled repeat RNA substrates to perform the nuclease activity assays (Figure 5.16A). We hoped to assess the effect on the individual activities of the constituent proteins when they associate to form a Cascade along with crRNA and also to reveal any differences in the activity from the *in vitro* reconstituted complex lacking the crRNA. For this, we incubated 3' 6-FAM labelled CRISPR repeat RNA as substrate with Cascade for 20 minutes and observed a smaller size band as compared to the position of substrate band which suggests the processing of repeat by Cascade (Figure 5.16B). This band was similar to the band observed when Cas5d processed the repeat RNA. Based on our product mapping experiments, this would correspond to the cleavage at G21 in the 32 nt repeat which results in a visible fragment of 11 nt, since the label is at 3'-end (Figure 2.8 in Chapter 2). But as the concentration of the cascade is increased, further smaller sized bands appeared along with the decrease in the intensity of the earlier observed band, suggesting the sequential processing of the repeat RNA, *i.e.*, the bigger sized product getting further processed into smaller products (Figure 5.16B). The pattern of repeat processing by Cascade resembles that of Cas5d, while the RNase activity of Csd1 seems to be suppressed as a part of the Cascade (Figure 5.16B). Thus, unlike the Cas6e of Cascade/type I-E, Cas5d seems to process the repeat RNA both as an individual protein and as a part of the Cascade/type I-C.

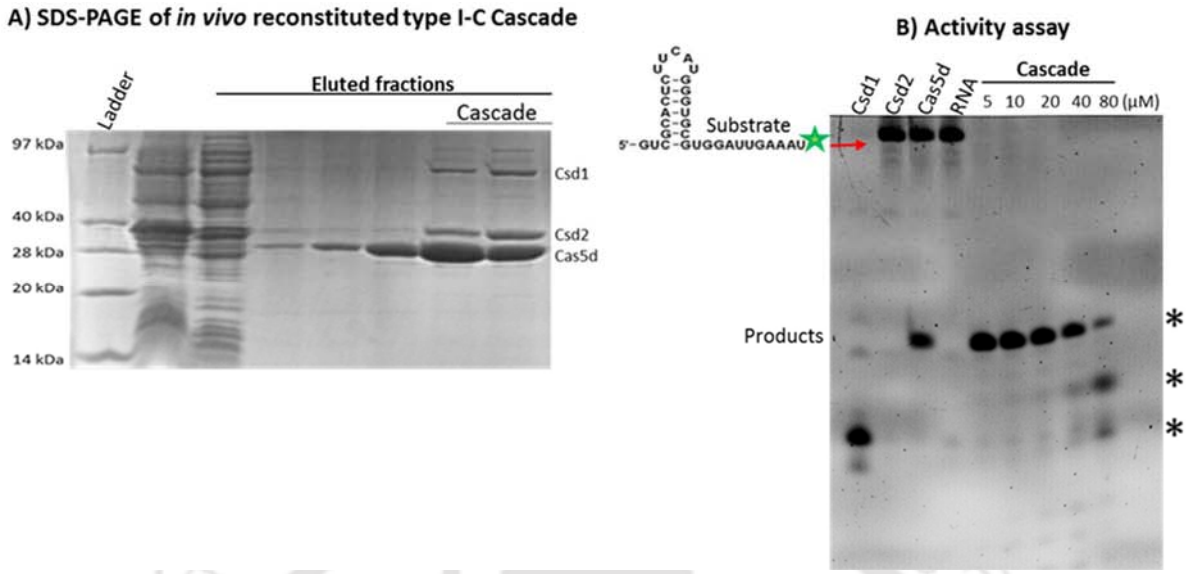


Figure 5.16 ***Activity assay to test RNase activity of type I-C Cascade.*** (A) The SDS-PAGE of the *in vivo* reconstituted Cascade containing Csd1, Csd2 and Cas5d having molecular weight of 72, 31, 27 kDa, respectively, is shown. (B) Activity assay to test RNase activity of type I-C Cascade is shown. The lanes containing the 3' 6-FAM labelled RNA substrate, the respective proteins and the Cascade are indicated. The increasing concentration of Cascade (μM) is shown. The reaction was incubated for 20 minutes at 37°C. * denotes the cleavage products.

5.3.3.5. Investigating the DNase activity of in vivo reconstituted Cascade

We further tried to probe the DNase activity of the Cascade. For, this we incubated the plasmid DNA with Cascade and also the individual protein components in presence and absence of divalent metal ion and monitored the effect. The plasmid preparation showed polymorphism in its mobility that is typical for a circular DNA. When Cas5d and Csd1 were incubated with DNA, we observed the conversion of supercoiled form into relaxed form. Similar pattern was observed, when Csd2 and Cascade were incubated with the DNA in presence of metal. In the cases, when DNA was incubated with Csd1 or Cascade, we observed an additional band stuck in the well, the intensity of which lowered in the presence of EDTA (Figure 5.17). This suggests the possibility of Csd1 or Cascade binding to DNA. The lanes containing DNA with Cas5d and Csd1 in presence of metal, showed degradation of

DNA, as anticipated, which was observed as decrease in the band intensity with a trail below, that was not detected in presence of EDTA (Figure 5.17). Thus, the individual DNase activity of Cas5d and Csd1 seems to be lost as a part of the Cascade. This was in contrast to the activity of the *in vitro* reconstituted complex, which showed combinatorial activity of the individual constituents (Figure 5.15). The plausible reason can be the individual protein did not associate properly in correct stoichiometry and orientation *in vitro*, which resulted in a complex different from the actual Cascade. The loss of DNase activity of Cascade implies that probably during the Cascade formation the DNA binding site or the metal binding site might get occluded from the surface, so that either the metal or the DNA substrate can no longer access it for binding. Consequently, it cannot switch its active site to process the DNA substrates. Thus, it seems that Cascade/type I-C formation segregates the promiscuous DNase activity of Cas proteins and enables it to retain the RNase activity to process the pre-CRISPR RNA transcript.

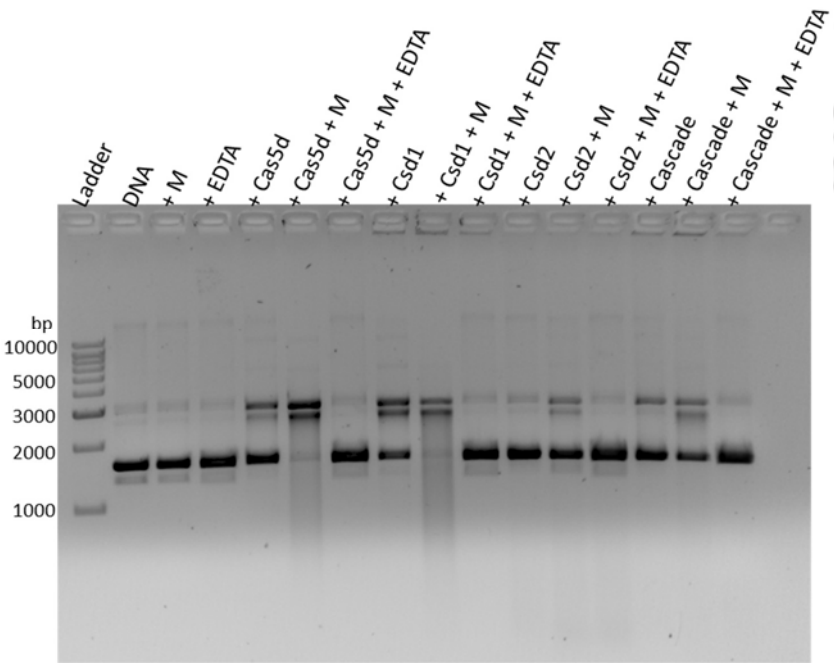


Figure 5.17 ***Activity assay to test DNase activity of type I-C Cascade.*** In all lanes, M denotes Mg^{2+} and the lanes containing the respective enzymes are indicated. The lanes containing the DNA, metal and EDTA are shown. While Cas5d and Csd1 are proficient in cleaving the plasmid DNA, Csd2 showed inertness. But the individual DNase activity was lost when Cascade was formed.

5.4. Summary

The Cascade/type I-C in *B. halodurans* comprises of only three types of protein *viz.,* Cas5d, Csd1 and Csd2 in specific stoichiometry. We identified Csd1 to be a fusion protein of its functional homolog Cse1 and Cse2 in type I-E. Owing to this covalent linkage, the copy number of the Cse2 that is fused to Cse1 in Csd1 is expected to differ and therefore the functionalities of Cascade/type I-C may deviate from its counterpart in type I-E. Interestingly, Csd1 was found to be nuclease with dual functionality, *i.e.*, it possessed both RNase and DNase activity, while Csd2 was inactive. Though RNA recognition seems to be specific, Csd1 apparently lack such specificity towards the DNA substrates. Moreover, Csd1 shows bias towards double stranded DNA. The Cascade exhibited RNase activity and processed the repeat RNA, raising the possibility of parallel processing of pre-crRNA along with the individual endoRNase, to elicit a rapid response to the infection. However, the DNase activity of the Cas5d and Csd1 seemed absent, when they are a part of the Cascade. This is perhaps due to the occlusion of DNA or metal binding sites, which might have occurred during the complex formation. Thus, type I-C system seems to be unique in possessing two dual functionality nucleases, which might be an outcome of evolutionary adaptation to suffice system requirement.

The CRISPR-Cas machinery, which provides an effective defense against foreign genetic elements in bacteria and archaea, is the only adaptive immune system known so far in the prokaryotes. Although the system is very dynamic owing to its adaptable and heritable nature, the benefit for the host in evolutionary terms seems to be transitional. This is because the rapidly incoming foreign DNA fragments, which are integrated as spacers in CRISPR loci can also be quickly lost, by either recombination or transpositional events. This dynamics of the system is evident by the high evolutionary rate of CRISPR array and the associated Cas proteins. The CRISPR-Cas system holds functional analogy to eukaryotic RNAi systems in utilizing small RNAs (crRNA and siRNA/miRNA in the respective systems) to direct the effector protein complexes to silence or destroy the invading nucleic acid. But the generation of small interfering RNAs seems to be distinct in both the cases. The RNAi system employs a metal dependent endoRNase Dicer, which utilizes two metal ion catalysis to process the dsRNA into the guide RNA (Jinek and Doudna, 2009), while all the known Cas endoRNases can metal independently generate the guide RNA. Thus, these small interfering RNAs are the key players of both defense systems, the generation of which holds the utmost importance for the functioning of the immune system.

The work presented in this thesis sheds light into the unique features of type I-C CRISPR-Cas system, which increases our understanding of maturation of crRNA in the novel prokaryotic immune system. The investigation of this system reveals that, in the absence of the Cas6 endoRNase, the role of crRNA processor can also be adopted by a subtype specific protein, which otherwise can be inactive or have a structural role. The distinctive feature of type I-C system is the presence two endonucleases *viz.,* Cas5d and Csd1 that recognize and process the pre-crRNA. This is in contrast to other known CRISPR-Cas subtypes, which usually harbor a single crRNA processor (Figure 6.1). The employment of two nucleases to

generate the guide RNA, might be an evolutionary adaptation required to make this system more efficient in triggering a rapid immune response. The enlightening feature revealed by the study on type I-C system is that, the basis of specificity inside the cell seems to be the coupling of transcription and processing, also called as co-transcriptional processing, which results in generation of uniform crRNA. The occurrence of the extended processing of CRISPR repeat over time is observed *in vitro,* which results in trimming of the repeat sequence, that forms the 5' handle. Thus, to avoid the resulting heterogeneity in crRNA, the system resorts to co-transcriptional processing. Another interesting feature revealed from this study is that, the substrate preference of the Cas endoRNases can be tuned by a metal cofactor, *i.e.,* the endoRNases of type I-C (Cas5d and Csd1) exhibit a metal dependent plasticity in their activity. Cas5d and Csd1 endoRNases possess a metal independent RNase activity and an additional metal dependent DNase activity (Figure 1.6). Interestingly, the processing of both the substrates show gross differences. The processing of CRISPR repeat RNA seems to proceed with specificity, resulting in the generation of specific products, while the DNase activity is apparently non-specific and the degradation occurs till the single nucleotide level. The two nucleases differ in the DNA substrates selection, while Csd1 has a bias towards double stranded DNA, Cas5d shows no such preference and cleaves all forms of DNA in presence of metal. This raises questions on the possible role(s) of this promiscuous DNase activity in type I-C. One can consider the promiscuous DNase activity exhibited by Cas5d and Csd1 as an evolutionary adaptation, suggesting their possible involvement in other stages of CRISPR-Cas immunity, wherein direct interaction with invading DNA is encountered. Recent studies have also shown that promiscuous restriction indeed confers selective advantage by increasing the survival fitness of bacteria to cope up against invading phages (Vasu et al., 2012). This characteristic feature has the potential to enhance the outreach of the defense strategy by countering the phages that escapes the bacterial defense

with reduced restriction sites or modification of phage genome. Therefore, the promiscuous DNase activity of Cascade-like complex, in association with Cas3, may elicit a rapid action response for target degradation during CRISPR interference. Yet another aspect of non-specific DNA targeting could be its involvement during the adaptation stage, where acquisition of new spacer from the invading genome requires fragmentation of the nucleic acids and the subsequent incorporation of this short fragment into the CRISPR locus. The phages can mutate the PAM or the seed regions of protospacers, to evade the CRISPR defense and become escape phages. To counter this, the Cascade complex associates with Cas3, Cas1 and Cas2 and primes the acquisition of more spacers from the mutated regions of the escape phage genome, for mounting the defense response during subsequent infection of the escape phage (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). This is undoubtedly beneficial to the host as it helps to adapt and become resistant to these escape phages. Therefore, the promiscuous DNase activity of Cas5d and Csd1 in the Cascade-like complex could come in handy either during the adaptation or the interference stage of the CRISPR immunity pathway. While the precise role of this nuclease activity during these processes needs further investigation, this observation allows one to forecast that the lineage specific functional variations operate in CRISPR-Cas systems across diverse microbial species, which may confer selective advantage for niche specific adaptation in protecting against genome predators.

Further, the nucleases Cas5d and Csd1 associate with the another subtype specific protein Csd2 to form the multiprotein assembly of type I-C antiviral defense complex. Among the type I-C Cascade components, Csd1 is the larger subunit. The nuclease activity of Csd1 in *B. halodurans* is in line with the nuclease activity exhibited by its ortholog Nar71 (MTH1090) in *Methanothermobacter thermoautotrophicus*. Though the region harboring the nuclease activity in Csd1 needs to be investigated, it appears that the Csd1 has undergone

adaptation in conformity with the requirement of type I-C immune response. Since we found Csd1 to be a fusion protein of its functional homolog Cse1 and Cse2 of type I-E. Owing to this covalent linkage the copy number of the Cse2 that is fused to Cse1 in Csd1 is expected to differ and thus the functionalities of Cascade-like complex in type I-C may deviate from its counterpart in type I-E. Cas5d, another nuclease of type I-C system, also seems to have undergone considerable adaption in comparison to the Cas5 of other known type I systems, which are inactive and play only a structural role as a part of Cascade. The nuclease activity of Cas5d can be attributed to the presence of active residues that are absent in Cas5 of other type I system. Further, the Cas5d possesses the remarkable ability of processing the repeats belonging to various CRISPR subtypes, suggesting a cross-species activity. This might be useful to organisms harbouring more than one CRISPR types, where it can process different CRISPR repeats associated with specific type and generate the guide RNA. Moreover, this seems to support the propagation by means of HGT, as the uptake of specific module can occur in case of requirement. Our experimentation showed, Csd2 to be inert against both RNA and DNA substrates and thus seems to play only the structural role by forming the backbone subunit of the Cascade complex (Figure 6.1). The Cascade of type I-C therefore, comprises of the two nuclease, Cas5d and Csd1 and an inactive protein, Csd2. The combinatorial activities of the individual components shown by the *in vitro* reconstituted Cascade, seems to be altered *in vivo*. In *in vivo*, the DNase activity of the Cascade seems to be abrogated, while it still retained the pre-crRNA processing ability (Figure 6.1). Moreover, the crRNA processing by Cascade seems to be a resultant of Cas5d RNase activity alone, suggesting the suppression of the RNase activity of Csd1 as a part of the Cascade. The abrogation of DNase activity of the individual components in Cascade can be the result of the occlusion of DNA or metal binding site during the complex formation or a modulation exerted by the interacting partners. Yet another speculation can be the formation of Cascade

immediately after the translation of Cas proteins, which thereby tweaks the promiscuous DNase activity of the Cas5d and Csd1, that may otherwise be deleterious to the cell. This non-specific activity can be tuned for driving rapid degradation of the invading DNA when needed. Thus, our work facilitated the characterization of the previously uncharacterized Cas proteins of type I-C system in *B. halodurans*. The multifarious role exhibited by these proteins highlights the remarkable adaptability of Cas protein, that helps in reducing the genetic burden of carrying many proteins for this purpose and thereby enhances the economy and versatility of the immune system. Further, the potential of dual-functionality nucleases can be harnessed for further applications.

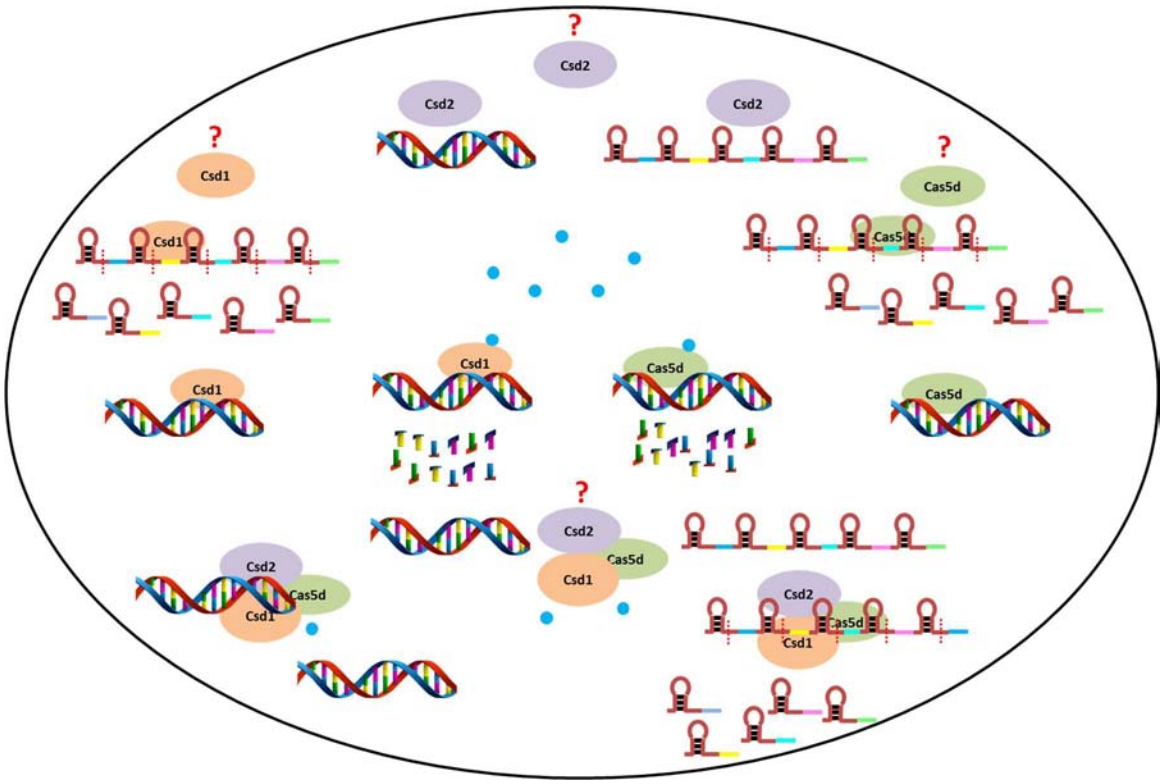


Figure 6.1. ***The characterization of the subtype specific proteins of type I-C CRISPR-Cas system.*** The three subtype specific proteins *viz.*, Csd1, Csd2 and Cas5d are shown in orange, purple and green respectively. The CRISPR array is shown consisting of stem-loop structured repeats in brown and diverse spacers in varied colours. The presence of metal in the vicinity is shown in blue circle. Csd1 and Cas5d processed the repeat RNA in metal independent, while DNA was processed in presence of metal cofactor. The Cascade complex only exhibits the RNase activity but remains inert to DNA, even in presence of metal.

The work opens the door in the area of other stages of CRISPR-Cas immunity, wherein the involvement of their DNase activity can be investigated. Also, the tunability of these nucleases can be explored to drive desirable functions like CRISPR-based controllable RNA processing or silencing. This can help in creation of standard genetic parts like promoters and repressors that can behave consistently across diverse genetic contexts. More broadly, an efficient genetic modification system can be made for new utility applications and therapies. If these Cas endoRNases can be tamed to overcome the bottleneck of the non-specificity, then this Cas machinery can also be utilized to target or replace genes, which can be instrumental in altering the germline of humans, animals and other organisms and also in modifying the genes of food crops. Also, the fundamental understanding of the system can help to combat the deadly diseases and circumvent the problem of multiple drug resistance, by specifically disrupting pathogen's CRISPR-Cas system, thereby making it susceptible to its invader (bacteriophages) and administrating the bacteriophage therapy. The bacteriophages, being host specific can selectively kill the pathogenic bacteria. Thus, the bacteriophage therapy seems to have an unexplored potential to cure deadly diseases. Overall the exploration of CRISPR-Cas system holds a promising future. The enormous potential of the CRISPR-Cas system is already evident by the tremendous application in the recent years, including genome editing and gene regulation with high precision.

Bibliography

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S., and Shinkai, A. (2010). Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. J Mol Biol *395*, 270-281.

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

Altenbuchner, J. (2016). Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 system. Appl Environ Microbiol *82*, 5421-5427.

Altfeld, M., Fadda, L., Frleta, D., and Bhardwaj, N. (2011). DCs and NK cells: critical effectors in the immune response to HIV-1. Nat Rev Immunol *11*, 176-186.

Anders, C., Bargsten, K., and Jinek, M. (2016). Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. Mol Cell *61*, 895-902.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569-573.

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

Ando, T., Xu, Q., Torres, M., Kusugami, K., Israel, D.A., and Blaser, M.J. (2000). Restriction-modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. Mol Microbiol *37*, 1052-1065.

Arslan, Z., Hermanns, V., Wurm, R., Wagner, R., and Pul, U. (2014). Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic Acids Res *42*, 7884-7893.

Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H., et al. (2013). Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Res *41*, 6347-6359.

Artsimovitch, I., Svetlov, V., Anthony, L., Burgess, R.R., and Landick, R. (2000). RNA polymerases from *Bacillus subtilis* and *Escherichia coli* differ in recognition of regulatory signals in vitro. J Bacteriol *182*, 6027-6035.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature *474*, 604-608.

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

Bachi, B., Reiser, J., and Pirrotta, V. (1979). Methylation and cleavage sequences of the EcoP1 restriction-modification enzyme. J Mol Biol *128*, 143-163.

Bair, C.L., and Black, L.W. (2007). A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. J Mol Biol *366*, 768-778.

Bandyopadhyay, P.K., Studier, F.W., Hamilton, D.L., and Yuan, R. (1985). Inhibition of the type I restriction-modification enzymes EcoB and EcoK by the gene 0.3 protein of bacteriophage T7. J Mol Biol *182*, 567-578.

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

Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol Cell *54*, 234-244.

Bassett, A.R., and Liu, J.L. (2014). CRISPR/Cas9 and genome editing in *Drosophila*. Journal of genetics and genomics *41*, 7-19.

Bassett, A.R., Tibbit, C., Ponting, C.P., and Liu, J.L. (2013). Highly efficient targeted mutagenesis of *Drosophila* with the CRISPR/Cas9 system. Cell Rep *4*, 220-228.

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

Benda, C., Ebert, J., Scheltema, R.A., Schiller, H.B., Baumgartner, M., Bonneau, F., Mann, M., and Conti, E. (2014). Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. Mol Cell *56*, 43-54.

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

Bernard, P., and Couturier, M. (1992). Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. J Mol Biol *226*, 735-745.

Bhaya, D., Davison, M., and Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet *45*, 273-297.

Blower, T.R., Pei, X.Y., Short, F.L., Fineran, P.C., Humphreys, D.P., Luisi, B.F., and Salmond, G.P. (2011a). A processed noncoding RNA regulates an altruistic bacterial antiviral system. Nat Struct Mol Biol *18*, 185-190.

Blower, T.R., Salmond, G.P., and Luisi, B.F. (2011b). Balancing at survival's edge: the structure and adaptive benefits of prokaryotic toxin-antitoxin partners. Curr Opin Struct Biol *21*, 109-118.

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

Boyer, H.W. (1971). DNA restriction and modification mechanisms in bacteria. Annu Rev Microbiol *25*, 153-176.

Brehm, S.L., and Cech, T.R. (1983). Fate of an intervening sequence ribonucleic acid: excision and cyclization of the *Tetrahymena* ribosomal ribonucleic acid intervening sequence *in vivo*. Biochemistry *22*, 2390-2397.

Brendel, J., Stoll, B., Lange, S.J., Sharma, K., Lenz, C., Stachler, A.E., Maier, L.K., Richter, H., Nickel, L., Schmitz, R.A., 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*. J Biol Chem *289*, 7164-7177.

Brosius, J., and Gould, S.J. (1992). On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci U S A *89*, 10706-10710.

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

Brown, S.P., Le Chat, L., De Paepe, M., and Taddei, F. (2006). Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. Curr Biol *16*, 2048-2052.

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

Bukowski, M., Rojowska, A., and Wladyka, B. (2011). Prokaryotic toxin-antitoxin systems--the role in bacterial physiology and application in molecular biology. Acta Biochim Pol *58*, 1-9.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science *273*, 1058-1073.

Calarco, J.A., and Friedland, A.E. (2015). Creating Genome Modifications in *C. elegans* Using the CRISPR/Cas9 System. Methods Mol Biol *1327*, 59-74.

Calvin, K., and Li, H. (2008). RNA-splicing endonuclease structure and function. Cell Mol Life Sci *65*, 1176-1185.

Carre-Mlouka, A., Gaumer, S., Gay, P., Petitjean, A.M., Coulondre, C., Dru, P., Bras, F., Dezelee, S., and Contamine, D. (2007). Control of sigma virus multiplication by the ref(2)P gene of *Drosophila melanogaster*: an in vivo study of the PB1 domain of Ref(2)P. Genetics *176*, 409-419.

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

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

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642-655.

Charpentier, E., and Marraffini, L.A. (2014). Harnessing CRISPR-Cas9 immunity for genetic engineering. Curr Opin Microbiol *19*, 114-119.

Chauhan, S., and Woodson, S.A. (2008). Tertiary interactions determine the accuracy of RNA folding. Journal of the American Chemical Society *130*, 1296-1303.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479-1491.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.L., and Brussow, H. (2004). Phage-host interaction: an ecological perspective. J Bacteriol *186*, 3677-3686.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol *31*, 230-232.

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

Christensen, S.K., Pedersen, K., Hansen, F.G., and Gerdes, K. (2003). Toxin-antitoxin loci as stress-response-elements: ChpAK/MazF and ChpBK cleave translated RNAs and are counteracted by tmRNA. J Mol Biol *332*, 809-819.

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res *36*, W197-201.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Crans, D.C., Smee, J.J., Gaidamauskas, E., and Yang, L. (2004). The chemistry and biochemistry of vanadium and the biological activities exerted by vanadium compounds. Chem Rev *104*, 849-902.

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics *25*, 1974-1975.

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

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

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

Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol *64*, 475-493.

Diez-Villasenor, C., Almendros, C., Garcia-Martinez, J., and Mojica, F.J. (2010). Diversity of CRISPR loci in *Escherichia coli*. Microbiology *156*, 1351-1361.

Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J., and Mojica, F.J. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. RNA Biol *10*, 792-802.

Donahue, J.P., Israel, D.A., Peek, R.M., Blaser, M.J., and Miller, G.G. (2000). Overcoming the restriction barrier to plasmid transformation of *Helicobacter pylori*. Mol Microbiol *37*, 1066-1074.

Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

Dulbecco, R. (1952). Mutual exclusion between related phages. J Bacteriol *63*, 209-217.

Dybvig, K., Sitaraman, R., and French, C.T. (1998). A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. Proc Natl Acad Sci U S A *95*, 13923-13928.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H., Tjian, R., and Doudna, J.A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature *538*, 270-273.

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

Ebina, H., Misawa, N., Kanemura, Y., and Koyanagi, Y. (2013). Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. Scientific reports *3*, 2510.

Engelberg-Kulka, H., Amitai, S., Kolodkin-Gal, I., and Hazan, R. (2006). Bacterial programmed cell death and multicellular behavior in bacteria. PLoS Genet *2*, e135.

Fineran, P.C., Blower, T.R., Foulds, I.J., Humphreys, D.P., Lilley, K.S., and Salmond, G.P. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. Proc Natl Acad Sci U S A *106*, 894-899.

Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature *532*, 517-521.

Forde, A., and Fitzgerald, G.F. (1999). Bacteriophage defence systems in lactic acid bacteria. Antonie Van Leeuwenhoek *76*, 89-113.

Fox, K.L., Srikhanta, Y.N., and Jennings, M.P. (2007). Phase variable type III restriction-modification systems of host-adapted bacterial pathogens. Mol Microbiol *65*, 1375-1379.

Fozo, E.M., Hemm, M.R., and Storz, G. (2008a). Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol Rev *72*, 579-589.

Fozo, E.M., Kawano, M., Fontaine, F., Kaya, Y., Mendieta, K.S., Jones, K.L., Ocampo, A., Rudd, K.E., and Storz, G. (2008b). Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. Mol Microbiol *70*, 1076-1093.

French, R.C., Lesley, S.M., Graham, A.F., and van, R.C. (1951). Studies on the relationship between virus and host cell. III. The breakdown of P32 labelled T2r+ bacteriophage adsorbed to *E. coli* previously infected by other coliphages of the T group. Can J Med Sci *29*, 144-148.

Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiacovo, M.P., Church, G.M., and Calarco, J.A. (2013). Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. Nat Methods *10*, 741-743.

Frols, S., Ajon, M., Wagner, M., Teichmann, D., Zolghadr, B., Folea, M., Boekema, E.J., Driessen, A.J., Schleper, C., and Albers, S.V. (2008). UV-inducible cellular aggregation of the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by pili formation. Mol Microbiol *70*, 938-952.

Galperin, M.Y., and Koonin, E.V. (2012). Divergence and convergence in enzyme evolution. J Biol Chem *287*, 21-28.

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

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

Garvey, P., Hill, C., and Fitzgerald, G.F. (1996). The Lactococcal Plasmid pNP40 Encodes a Third Bacteriophage Resistance Mechanism, One Which Affects Phage DNA Penetration. Appl Environ Microbiol *62*, 676-679.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A *109*, E2579-2586.

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

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

Gogarten, J.P., Doolittle, W.F., and Lawrence, J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol Biol Evol *19*, 2226-2238.

Gogarten, J.P., and Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol *3*, 679-687.

Gong, B., Shin, M., Sun, J., Jung, C.H., Bolt, E.L., van der Oost, J., and Kim, J.S. (2014). Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proc Natl Acad Sci U S A *111*, 16359-16364.

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

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

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol Microbiol *10*, 1057-1065.

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

Guy, C.P., Majernik, A.I., Chong, J.P., and Bolt, E.L. (2004). A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. Nucleic Acids Res *32*, 6176-6186.

Haaber, J., Moineau, S., and Hammer, K. (2009). Activation and transfer of the chromosomal phage resistance mechanism AbiV in *Lactococcus lactis*. Appl Environ Microbiol *75*, 3358-3361.

Haerter, J.O., and Sneppen, K. (2012). Spatial Structure and Lamarckian Adaptation Explain Extreme Genetic Diversity at CRISPR Locus. Mbio *3*, *e00126-12.*

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

Hale, C.R., Cocozaki, A., Li, H., Terns, R.M., and Terns, M.P. (2014). Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. Genes Dev *28*, 2432-2443.

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

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

Hamilton, H.L., and Dillard, J.P. (2006). Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Mol Microbiol *59*, 376-385.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244-251.

Hatoum-Aslan, A., Maniv, I., and Marraffini, L.A. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. Proc Natl Acad Sci U S A *108*, 21218-21222.

Hatoum-Aslan, A., Maniv, I., Samai, P., and Marraffini, L.A. (2014). Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. J Bacteriol *196*, 310-317.

Hatoum-Aslan, A., Samai, P., Maniv, I., Jiang, W., and Marraffini, L.A. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J Biol Chem *288*, 27888-27897.

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

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

Hayes, R.P., and Ke, A. (2015). One More Piece Down to Solve the III-A CRISPR Puzzle. J Mol Biol *427*, 228-230.

Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B., Rajashankar, K., Bailey, S., Wiedenheft, B., and Ke, A. (2016). Structural basis for promiscuous PAM recognition in type I-E Cascade from *E. coli*. Nature *530*, 499-503.

Hazan, R., and Engelberg-Kulka, H. (2004). *Escherichia coli* mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. Mol Genet Genomics *272*, 227-234.

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

Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature *519*, 199-202.

Hermans, P.W., van Soolingen, D., Bik, E.M., de Haas, P.E., Dale, J.W., and van Embden, J.D. (1991). Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect Immun *59*, 2695-2705.

Hirano, H., Gootenberg, J.S., Horii, T., Abudayyeh, O.O., Kimura, M., Hsu, P.D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F., *et al.* (2016a). Structure and Engineering of *Francisella novicida* Cas9. Cell *164*, 950-961.

Hirano, S., Nishimasu, H., Ishitani, R., and Nureki, O. (2016b). Structural Basis for the Altered PAM Specificities of Engineered CRISPR-Cas9. Mol Cell *61*, 886-894.

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S.J., Naidich, S., Garcia, M., Kahn, E., Bergmire-Sweat, D., and Musser, J.M. (1999). Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. Emerg Infect Dis *5*, 254-263.

Hooton, S.P., and Connerton, I.F. (2014). *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Frontiers in microbiology *5*, 744.

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

Horvath, P., Coute-Monvoisin, A.C., Romero, D.A., Boyaval, P., Fremaux, C., and Barrangou, R. (2009). Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol *131*, 62-70.

Hoskisson, P.A., and Smith, M.C. (2007). Hypervariation and phase variation in the bacteriophage 'resistome'. Curr Opin Microbiol *10*, 396-400.

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

Hrle, A., Su, A.A., Ebert, J., Benda, C., Randau, L., and Conti, E. (2013). Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. RNA Biol *10*, 1670-1678.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262-1278.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M.D., Jr., Zhou, S., Rajashankar, K., Kurinov, I.*, et al.* (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol *21*, 771-777.

Hyman, P., and Abedon, S.T. (2010). Bacteriophage host range and bacterial resistance. Adv Appl Microbiol *70*, 217-248.

Iida, S., Meyer, J., Bachi, B., Stalhammar-Carlemalm, M., Schrickel, S., Bickle, T.A., and Arber, W. (1983). DNA restriction--modification genes of phage P1 and plasmid p15B. Structure and *in vitro* transcription. J Mol Biol *165*, 1-18.

Ishiguro, E.E., Kay, W.W., Ainsworth, T., Chamberlain, J.B., Austen, R.A., Buckley, J.T., and Trust, T.J. (1981). Loss of virulence during culture of *Aeromonas salmonicida* at high temperature. J Bacteriol *148*, 333-340.

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

Jackson, R.N., Golden, S.M., van Erp, P.B., Carter, J., Westra, E.R., Brouns, S.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. Science *345*, 1473-1479.

Jackson, S.A., Koduvayur, S., and Woodson, S.A. (2006). Self-splicing of a group I intron reveals partitioning of native and misfolded RNA populations in yeast. RNA *12*, 2149-2159.

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

Jansen, R., van Embden, J.D., Gaastra, W., and Schouls, L.M. (2002b). Identification of a novel family of sequence repeats among prokaryotes. Omics *6*, 23-33.

Jaroszewski, L., Li, Z., Cai, X.H., Weber, C., and Godzik, A. (2011). FFAS server: novel features and applications. Nucleic Acids Res *39*, W38-44.

Jeltsch, A., and Pingoud, A. (1996). Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J Mol Evol *42*, 91-96.

Jiang, F., and Doudna, J.A. (2015). The structural biology of CRISPR-Cas systems. Curr Opin Struct Biol *30*, 100-111.

Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867-871.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477-1481.

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

Jinek, M., and Doudna, J.A. (2009). A three-dimensional view of the molecular machinery of RNA interference. Nature *457*, 405-412.

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

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science *343*, 1247997.

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

Jung, T.Y., An, Y., Park, K.H., Lee, M.H., Oh, B.H., and Woo, E. (2015). Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. Structure *23*, 782-790.

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

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H., Marr, M.T., 2nd, and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. Genes Dev *25*, 2502-2512.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490-495.

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res *29*, 3742-3756.

Koduvayur, S.P., and Woodson, S.A. (2004). Intracellular folding of the *Tetrahymena* group I intron depends on exon sequence and promoter choice. RNA *10*, 1526-1532.

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

Koo, Y., Ka, D., Kim, E.J., Suh, N., and Bae, E. (2013). Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. J Mol Biol *425*, 3799-3810.

Koonin, E.V., and Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol *10*, 679-686.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). Multiple links between transcription and splicing. RNA *10*, 1489-1498.

Kramer, F.R., and Mills, D.R. (1981). Secondary structure formation during RNA synthesis. Nucleic Acids Res *9*, 5109-5124.

Kroger, M., Hobom, G., Schutte, H., and Mayer, H. (1984). Eight new restriction endonucleases from *Herpetosiphon giganteus*--divergent evolution in a family of enzymes. Nucleic Acids Res *12*, 3127-3141.

Kruger, D.H., and Bickle, T.A. (1983). Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. Microbiol Rev *47*, 345-360.

Kumar, V., and Jain, M. (2015). The CRISPR-Cas system for plant genome editing: advances and opportunities. Journal of experimental botany *66*, 47-57.

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

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat Rev Microbiol *8*, 317-327.

Lai, D., Proctor, J.R., and Meyer, I.M. (2013). On the importance of cotranscriptional RNA structure formation. RNA *19*, 1461-1473.

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

Lepikhov, K., Tchernov, A., Zheleznaja, L., Matvienko, N., Walter, J., and Trautner, T.A. (2001). Characterization of the type IV restriction modification system BspLU11III from *Bacillus* sp. LU11. Nucleic Acids Res *29*, 4691-4698.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505-510.

Li, M., Wang, R., Zhao, D., and Xiang, H. (2014). Adaptation of the *Haloarcula hispanica* CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res *42*, 2483-2492.

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

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

Lintner, N.G., Frankel, K.A., Tsutakawa, S.E., Alsbury, D.L., Copie, V., Young, M.J., Tainer, J.A., and Lawrence, C.M. (2011a). The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. J Mol Biol *405*, 939-955.

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

Llave, C. (2010). Virus-derived small interfering RNAs at the core of plant-virus interactions. Trends Plant Sci *15*, 701-707.

Lu, M.J., and Henning, U. (1994). Superinfection exclusion by T-even-type coliphages. Trends Microbiol *2*, 137-139.

Lutz, B., Faber, M., Verma, A., Klumpp, S., and Schug, A. (2014). Differences between cotranscriptional and free riboswitch folding. Nucleic Acids Res *42*, 2687-2696.

Maguire, M.E., and Cowan, J.A. (2002). Magnesium chemistry and biochemistry. Biometals *15*, 203-210.

Mahen, E.M., Harger, J.W., Calderon, E.M., and Fedor, M.J. (2005). Kinetics and thermodynamics make different contributions to RNA folding *in vitro* and in yeast. Mol Cell *19*, 27-37.

Mahen, E.M., Watson, P.Y., Cottrell, J.W., and Fedor, M.J. (2010). mRNA secondary structures fold sequentially but exchange rapidly *in vivo*. PLoS biology *8*, e1000307.

Mahony, J., McGrath, S., Fitzgerald, G.F., and van Sinderen, D. (2008). Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. Appl Environ Microbiol *74*, 6206-6215.

Makarova, K.S., Anantharaman, V., Aravind, L., and Koonin, E.V. (2012). Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biol Direct *7*, 40.

Makarova, K.S., Anantharaman, V., Grishin, N.V., Koonin, E.V., and Aravind, L. (2014). CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Frontiers in genetics *5*, 102.

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

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

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

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

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol *13*, 722-736.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2009). Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol Direct *4*, 19.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). The basic building blocks and evolution of CRISPR-CAS systems. Biochem Soc Trans *41*, 1392-1400.

Mali, P., Esvelt, K.M., and Church, G.M. (2013). Cas9 as a versatile tool for engineering biology. Nat Methods *10*, 957-963.

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

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

Marinus, M.G., and Casadesus, J. (2009). Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol Rev *33*, 488-503.

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

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

Masepohl, B., Gorlitz, K., and Bohme, H. (1996). Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. Biochim Biophys Acta *1307*, 26-30.

Maxwell, K.L. (2016). Phages Fight Back: Inactivation of the CRISPR-Cas Bacterial Immune System by Anti-CRISPR Proteins. PLoS Pathog *12*, e1005282.

McGrath, S., Fitzgerald, G.F., and van Sinderen, D. (2002). Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol *43*, 509-520.

Meisel, A., Bickle, T.A., Kruger, D.H., and Schroeder, C. (1992). Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. Nature *355*, 467-469.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343-349.

Merkhofer, E.C., Hu, P., and Johnson, T.L. (2014). Introduction to cotranscriptional RNA splicing. Methods Mol Biol *1126*, 83-96.

Meyer, I.M., and Miklos, I. (2004). Co-transcriptional folding is encoded within RNA genes. BMC Mol Biol *5*, 10.

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. Science *335*, 428-432.

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

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

Mojica, F.J., Diez-Villasenor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol *36*, 244-246.

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

Molineux, I.J. (1991). Host-parasite interactions: recent developments in the genetics of abortive phage infections. New Biol *3*, 230-236.

Mulepati, S., and Bailey, S. (2011). Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J Biol Chem *286*, 31896-31903.

Mulepati, S., Heroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479-1484.

Mullings, R., Bennett, S.P., and Brown, N.L. (1988). Investigation of sequence homology in a group of type-II restriction/modification isoschizomers. Gene *74*, 245-251.

Murray, N.E. (2000). Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev *64*, 412-434.

Naito, T., Kusano, K., and Kobayashi, I. (1995). Selfish behavior of restriction-modification systems. Science *267*, 897-899.

Nakata, A., Amemura, M., and Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. J Bacteriol *171*, 3553-3556.

Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P., and Ke, A. (2012). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure *20*, 1574-1584.

Nam, K.H., Kurinov, I., and Ke, A. (2011). Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. J Biol Chem *286*, 30759-30768.

Nariya, H., and Inouye, M. (2008). MazF, an mRNA interferase, mediates programmed cell death during multicellular *Myxococcus* development. Cell *132*, 55-66.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A.*, et al.* (1999). Evidence for lateral

gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature *399*, 323-329.

Niewoehner, O., Jinek, M., and Doudna, J.A. (2014). Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. Nucleic Acids Res *42*, 1341-1353.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935-949.

Nordstrom, K., and Forsgren, A. (1974). Effect of protein A on adsorption of bacteriophages to *Staphylococcus aureus*. J Virol *14*, 198-202.

Nordstrom, K., Forsgren, A., and Cox, P. (1974). Prevention of bacteriophage adsorption to *Staphylococcus aureus* by immunoglobulin G. J Virol *14*, 203-206.

Numata, T., Inanaga, H., Sato, C., and Osawa, T. (2015). Crystal Structure of the Csm3-Csm4 Subcomplex in the Type III-A CRISPR-Cas Interference Complex. J Mol Biol *427*, 259-273.

Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol Cell *62*, 824-833.

Nunez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015a). Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature *527*, 535-538.

Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol *21*, 528-534.

Nunez, J.K., Lee, A.S., Engelman, A., and Doudna, J.A. (2015b). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature *519*, 193-198.

Olieric, V., Weinert, T., Finke, A.D., Anders, C., Li, D., Olieric, N., Borca, C.N., Steinmetz, M.O., Caffrey, M., Jinek, M., *et al.* (2016). Data-collection strategy for challenging native SAD phasing. Acta crystallographica Section D, Structural biology *72*, 421-429.

Orlowski, J., and Bujnicki, J.M. (2008). Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. Nucleic Acids Res *36*, 3552-3569.

Osawa, T., Inanaga, H., Sato, C., and Numata, T. (2015). Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. Mol Cell *58*, 418-430.

Pawluk, A., Bondy-Denomy, J., Cheung, V.H.W., Maxwell, K.L., and Davidson, A.R. (2014). A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of *Pseudomonas aeruginosa*. Mbio *5*, e00896-14.

Pecota, D.C., and Wood, T.K. (1996). Exclusion of T4 phage by the hok/sok killer locus from plasmid R1. J Bacteriol *178*, 2044-2050.

Pedersen, K., Christensen, S.K., and Gerdes, K. (2002). Rapid induction and reversal of a bacteriostatic condition by controlled expression of toxins and antitoxins. Mol Microbiol *45*, 501-510.

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K.H., Bernardis, S., Ke, A., and DeLisa, M.P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. Mol Microbiol *79*, 584-599.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605-1612.

Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. Cell Mol Life Sci *62*, 685-707.

Pohl, C., Kiel, J.A., Driessen, A.J., Bovenberg, R.A., and Nygard, Y. (2016). CRISPR/Cas9 Based Genome Editing of *Penicillium chrysogenum*. ACS synthetic biology *5*, 754-764.

Portillo, M.C., and Gonzalez, J.M. (2009). CRISPR elements in the Thermococcales: evidence for associated horizontal gene transfer in *Pyrococcus furiosus*. Journal of applied genetics *50*, 421-430.

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

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

Proudfoot, N. (2000). Connecting transcription to messenger RNA processing. Trends Biochem Sci *25*, 290-293.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

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

Punetha, A., Sivathanu, R., and Anand, B. (2014). Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system. Nucleic Acids Res *42*, 3846-3856.

Putnam, C.D., and Tainer, J.A. (2005). Protein mimicry of DNA and pathway regulation. DNA repair *4*, 1410-1420.

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

Ramia, N.F., Spilman, M., Tang, L., Shao, Y., Elmore, J., Hale, C., Cocozaki, A., Bhattacharya, N., Terns, R.M., Terns, M.P., *et al.* (2014). Essential structural and functional roles of the Cmr4 subunit in RNA cleavage by the Cmr CRISPR-Cas complex. Cell Rep *9*, 1610-1617.

Reeks, J., Graham, S., Anderson, L., Liu, H., White, M.F., and Naismith, J.H. (2013). Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. RNA Biol *10*, 762-769.

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H., and Fineran, P.C. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res *42*, 8516-8526.

Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S., Dryden, D.T., Dybvig, K., et al. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res *31*, 1805-1812.

Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res *43*, D298-299.

Robson, J., McKenzie, J.L., Cursons, R., Cook, G.M., and Arcus, V.L. (2009). The vapBC operon from *Mycobacterium smegmatis* is an autoregulated toxin-antitoxin module that controls growth via inhibition of translation. J Mol Biol *390*, 353-367.

Rocha, E.P., Danchin, A., and Viari, A. (2001). Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. Genome research *11*, 946-958.

Rollins, M.F., Schuman, J.T., Paulus, K., Bukhari, H.S., and Wiedenheft, B. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from *Pseudomonas aeruginosa*. Nucleic Acids Res *43*, 2216-2222.

Roossinck, M.J. (2011). The good viruses: viral mutualistic symbioses. Nat Rev Microbiol *9*, 99-108.

Rosenshine, I., Tchelet, R., and Mevarech, M. (1989). The mechanism of DNA transfer in the mating system of an archaebacterium. Science *245*, 1387-1389.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C.V., Spagnolo, L., and White, M.F. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell *52*, 124-134.

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

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164-1174.

Samson, J.E., Magadan, A.H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol *11*, 675-687.

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

Saunders, N.J., Peden, J.F., Hood, D.W., and Moxon, E.R. (1998). Simple sequence repeats in the *Helicobacter pylori* genome. Mol Microbiol *27*, 1091-1098.

Schnettler, E., de Vries, W., Hemmes, H., Haasnoot, J., Kormelink, R., Goldbach, R., and Berkhout, B. (2009). The NS3 protein of rice hoja blanca virus complements the RNAi suppressor function of HIV-1 Tat. EMBO Rep *10*, 258-263.

Scholl, D., and Merril, C. (2005). The genome of bacteriophage K1F, a T7-like phage that has acquired the ability to replicate on K1 strains of *Escherichia coli*. J Bacteriol *187*, 8499-8503.

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

Seib, K.L., Peak, I.R., and Jennings, M.P. (2002). Phase variable restriction-modification systems in *Moraxella catarrhalis*. FEMS Immunol Med Microbiol *32*, 159-165.

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

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

Shah, S.A., and Garrett, R.A. (2011). CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. Research in microbiology *162*, 27-38.

Shah, S.A., Hansen, N.R., and Garrett, R.A. (2009). Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. Biochem Soc Trans *37*, 23-28.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L., et al. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. Nat Biotechnol *31*, 686-688.

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

Shao, Y., Richter, H., Sun, S., Sharma, K., Urlaub, H., Randau, L., and Li, H. (2016). A Non-Stem-Loop CRISPR RNA Is Processed by Dual Binding Cas6. Structure *24*, 547-554.

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

Siomi, H., and Siomi, M.C. (2009). On the road to reading the RNA-interference code. Nature *457*, 396-404.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology *12*, 246-258.

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

Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H., and Stagg, S. (2013). Structure of an RNA silencing complex of the CRISPR-Cas immune system. Mol Cell *52*, 146-152.

Srikhanta, Y.N., Fox, K.L., and Jennings, M.P. (2010). The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. Nat Rev Microbiol *8*, 196-206.

Staals, R.H., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K., et al. (2013). Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. Mol Cell *52*, 135-145.

Staals, R.H., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., et al. (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. Mol Cell *56*, 518-530.

Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. Bioessays *33*, 43-51.

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

Studier, F.W., and Bandyopadhyay, P.K. (1988). Model for how type I restriction enzymes select cleavage sites in DNA. Proc Natl Acad Sci U S A *85*, 4677-4681.

Sumby, P., and Smith, M.C. (2002). Genetics of the phage growth limitation (Pgl) system of *Streptomyces coelicolor* A3(2). Mol Microbiol *44*, 489-500.

Sun, J., Jeon, J.H., Shin, M., Shin, H.C., Oh, B.H., and Kim, J.S. (2014). Crystal structure and CRISPR RNA-binding site of the Cmr1 subunit of the Cmr interference complex. Acta crystallographica Section D, Biological crystallography *70*, 535-543.

Sutherland, I.W., Hughes, K.A., Skillman, L.C., and Tait, K. (2004). The interaction of phage and biofilms. FEMS Microbiol Lett *232*, 1-6.

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

Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, C., Nwokeoji, A.O., Dickman, M.J., Horvath, P., and Siksnys, V. (2014). Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. Mol Cell *56*, 506-517.

Tan, A., Hill, D.M., Harrison, O.B., Srikhanta, Y.N., Jennings, M.P., Maiden, M.C., and Seib, K.L. (2016). Distribution of the type III DNA methyltransferases modA, modB and modD among *Neisseria meningitidis* genotypes: implications for gene regulation and virulence. Scientific reports *6*, 21015.

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

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

Taylor, D.W., Zhu, Y., Staals, R.H., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science *348*, 581-585.

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

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

Treiber, D.K., and Williamson, J.R. (2001). Beyond kinetic traps in RNA folding. Curr Opin Struct Biol *11*, 309-314.

Trevino, A.E., and Zhang, F. (2014). Genome editing using Cas9 nickases. Methods in enzymology *546*, 161-174.

Tsui, T.K.M., and Li, H. (2015). Structure Principles of CRISPR-Cas Surveillance and Effector Complexes. Annu Rev Biophys *44*, 229-255.

Ussery, D.W., Binnewies, T.T., Gouveia-Oliveira, R., Jarmer, H., and Hallin, P.F. (2004). Genome update: DNA repeats in bacterial genomes. Microbiology *150*, 3519-3521.

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

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol *12*, 479-492.

van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B.A., and Schouls, L.M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J Bacteriol *182*, 2393-2401.

van Erp, P.B., Jackson, R.N., Carter, J., Golden, S.M., Bailey, S., and Wiedenheft, B. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in *Escherichia coli*. Nucleic Acids Res *43*, 8381-8391.

Van Melderen, L. (2010). Toxin-antitoxin systems: why so many, what for? Curr Opin Microbiol *13*, 781-785.

Van Melderen, L., and Saavedra De Bast, M. (2009). Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet *5*, e1000437.

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

Van Valen, L. (1973). A new evolutionary law. Evolutionary Theory *1*, 1-30.

Vasu, K., Nagamalleswari, E., and Nagaraja, V. (2012). Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. Proc Natl Acad Sci U S A *109*, E1287-1293.

Venclovas, C., Timinskas, A., and Siksnys, V. (1994). Five-stranded beta-sheet sandwiched with two alpha-helices: a structural link between restriction endonucleases EcoRI and EcoRV. Proteins *20*, 279-282.

Vestergaard, G., Garrett, R.A., and Shah, S.A. (2014). CRISPR adaptive immune systems of Archaea. RNA Biol *11*, 156-167.

Viswanathan, P., Murphy, K., Julien, B., Garza, A.G., and Kroos, L. (2007). Regulation of dev, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. J Bacteriol *189*, 3738-3750.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K., et al. (2015). Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res *43*, 10848-10860.

Vovis, G.F., Horiuchi, K., and Zinder, N.D. (1974). Kinetics of methylation of DNA by a restriction endonuclease from *Escherichia coli* B. Proc Natl Acad Sci U S A *71*, 3810-3813.

Waite-Rees, P.A., Keating, C.J., Moran, L.S., Slatko, B.E., Hornstra, L.J., and Benner, J.S. (1991). Characterization and expression of the *Escherichia coli* Mrr restriction system. J Bacteriol *173*, 5207-5219.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840-853.

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

Wang, X., Yao, D., Xu, J.G., Li, A.R., Xu, J., Fu, P., Zhou, Y., and Zhu, Y. (2016). Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. Nat Struct Mol Biol *23*, 868-70.

Watanabe, K., Ishibashi, K., Nakashima, Y., and Sakurai, T. (1984). A phage-resistant mutant of *Lactobacillus casei* which permits phage adsorption but not genome injection. J Gen Virol *65 (Pt 5)*, 981-986.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Wei, Y., Chesne, M.T., Terns, R.M., and Terns, M.P. (2015a). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. Nucleic Acids Res *43*, 1749-58.

Wei, Y., Terns, R.M., and Terns, M.P. (2015b). Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev *29*, 356-361.

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

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

Westra, E.R., Swarts, D.C., Staals, R.H., Jore, M.M., Brouns, S.J., and van der Oost, J. (2012b). The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet *46*, 311-339.

White, M.F. (2009). Structure, function and evolution of the XPD family of iron-sulfur-containing 5'-->3' DNA helicases. Biochem Soc Trans *37*, 547-551.

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

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

Wiedenheft, B., van Duijn, E., Bultema, J.B., Waghmare, S.P., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J., Boekema, E.J., Dickman, M.J., et al. (2011b). RNA-guided

complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A *108*, 10092-10097.

Williams, E., Lowe, T.M., Savas, J., and DiRuggiero, J. (2007). Microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus* exposed to gamma irradiation. Extremophiles *11*, 19-29.

Wilson, G.G., and Murray, N.E. (1991). Restriction and modification systems. Annu Rev Genet *25*, 585-627.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576-4579.

Woodson, S.A. (2002). Folding mechanisms of group I ribozymes: role of stability and contact order. Biochem Soc Trans *30*, 1166-1169.

Wyers, F., Petitjean, A.M., Dru, P., Gay, P., and Contamine, D. (1995). Localization of domains within the Drosophila Ref(2)P protein involved in the intracellular control of sigma rhabdovirus multiplication. J Virol *69*, 4463-4470.

Yang, W., Lee, J.Y., and Nowotny, M. (2006). Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. Mol Cell *22*, 5-13.

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

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759-771.

Zhang, H., Zheng, X., and Zhang, Z. (2010). The role of vacuolar processing enzymes in plant immunity. Plant Signal Behav *5*, 1565-1567.

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

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., and Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature *515*, 147-150.

Zheng, Y., Roberts, R.J., and Kasif, S. (2004). Identification of genes with fast-evolving regions in microbial genomes. Nucleic Acids Res *32*, 6347-6357.

Zhu, X., and Ye, K. (2015). Cmr4 is the slicer in the RNA-targeting Cmr CRISPR complex. Nucleic Acids Res *43*, 1257-1267.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res *31*, 3406-3415.

List of publications

1. Punetha A, Sivathanu R, Anand B: **Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system**. Nucleic Acids Res 2014, 42(6):3846-3856.
2. Punetha A, Nimkar S, Chhetry S, Anand B: **On the importance of co-transcriptional processing during the maturation of crRNA in CRISPR-Cas type I-C system** (manuscript under preparation).

List of poster presentations in conferences

1) INDO-US Conference and Workshop on Recent Advances in Structural Biology & Drug Discovery at Indian Institute of Technology Roorkee, 9-11 Oct 2014.

2) 42nd National Seminar on Crystallography and International Workshop on Application of X-ray Diffraction for Drug Discovery at JNU, New Delhi, 21-23 Nov 2013.

3) International Conference on Biomolecular Forms and Functions at IISC, Bangalore 8-11 Jan 2013.

4) Structural and Biophysical method for biological macromolecules in solution organized by EMBO at CCMB, Hyderabad, 29-6 Dec 2012.