Molecular Mechanism Governing the CRISPR Adaptation in CRISPR-Cas Type I-E System

A Thesis Submitted in Partial Fulfilment of the Requirements for the Award of the Degree of Doctor of Philosophy

by

Yoganand.K.N.R

Registration Number: 11610606

Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati

Dedication

I dedicate this thesis to my grandparents, parents, wife, daughters and teachers.

Indian Institute of Technology Guwahati
Department of Biosciences and Bioengineering

Statement

I do hereby declare that the matter embodied in this thesis entitled "**Molecular Mechanism Governing the CRISPR Adaptation in CRISPR-Cas Type I-E System"** is the result of work carried out in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the supervision of **Dr. B. Anand** and **Dr. R. Swaminathan**.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

Yoganand.K.N.R.

Roll no: 11610606

Indian Institute of Technology Guwahati
Department of Biosciences and Bioengineering

Certificate

It is certified that the work described in this thesis entitled **"Molecular Mechanism Governing the CRISPR Adaptation in CRISPR-Cas Type I-E System"** by Mr. Yoganand.K.N.R. for the award of degree of Doctor of Philosophy is an authentic record of the results obtained from the research work carried out under our supervision in the Department of Biosciences and Bioengineering, IITG. The work embodied in this thesis has not been submitted elsewhere for a degree.

Dr. B. Anand
Thesis Supervisor
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Assam 781039, India

Prof. R. Swaminathan
Co-supervisor
Department of Biosciences and
Bioengineering Indian Institute of Technology
Guwahati Assam 781039, India

Acknowledgement

Completing this thesis marks a new milestone in my life. At the end of this long, arduous and enlightening research voyage, I would like to whole-heartedly thank the people who stood by me and encouraged my dream of becoming a scientist.

First and foremost, I would like to thank my thesis advisor, Dr. B. Anand, without whose support and guidance, this accomplishment would not have been possible. Dr. Anand encouraged me to explore my scientific curiosity, and his clear inputs and precise directions were vital to the successful execution of my experiments. In my difficult times, motivation given by him acted as a catalyst to improve myself both scientifically and personally.

Next, I would like to express my immense gratitude towards Prof. R. Swaminathan, my thesis co-supervisor; my doctoral committee, Dr. Nitin Chaudhary, Prof. Aiyagari Ramesh of the Department of Biosciences and Bioengineering and Prof. Parameswar K. Iyer of the Department of Chemistry for their valuable suggestions and encouragement. Their insightful advice and timely counsel were crucial to the fruitful progression of my thesis. I would also like to thank Dr. Senthil Kumar Sivaprakasam, Dr. Rajkumar P. Thummer and Dr. Priyadarshi Satpati of the Department of Biosciences and Bioengineering for their valuable suggestions during annual evaluation presentations.

I want to thank Heads of the Department during my PhD tenure, Prof. Latha Rangan, Prof. Arun Goyal, Prof. V. Venkata Dasu and Prof. Kannan Pakshirajan as well as the other faculty members for their support and advice.

I owe my thanks to the Department of Biosciences and Bioengineering and Central Instrumentation Facility, IIT Guwahati for providing the necessary research facilities to accomplish my PhD thesis objectives. I want to extend my thanks to all the Department of Biosciences and Bioengineering staff members for providing me with the logistical support essential to perform my research.

I acknowledge the geniality of all the investigators who shared their plasmids and bacterial strains either directly or through the addgene repository.

I thank all the current and past MAB lab members for making a cheerful environment in the lab. I want to acknowledge Sivathanu, Siddharth and Manasa for their helping hand in

performing experiments. I will always remember long, troubleshooting discussions that I had with Himanshu. I thank Manasa, Siddharth and Sunanada for their help in proofreading the thesis. I will forever cherish and miss the discussions, debates and banters had with all my labmates. I also thank all my friends at IIT Guwahati for the beautiful memories that we created together.

The vibrant atmosphere and facilities offered by IIT Guwahati had allowed me to enhance my knowledge and skills. The humility of IITG fraternity, who hails from every corner of the country, gave me great motivation and personal learning. A stroll in the IIT Guwahati campus's enthralling scenic beauty is a stress buster and this had taught me the importance of living with nature. I thank the fraternity of IIT Guwahati for the best days of my life at the campus.

I extend my gratitude to the funding agencies that provided financial support to conduct my research. I want to acknowledge the research grant support for MAB lab from the Department of Biotechnology, Govt. of India and Department of Science and Technology, Govt. of India. Next, I acknowledge MHRD, Govt. of India for the fellowship during my PhD tenure.

I thank my friends and relatives, whose constant love and support helped me push through every roadblock I encountered in my PhD career. Finally, I think back to my parents and life partner's unconditional love, personal sacrifices, and patience, that made me reach this stage.

Thank you!!!

Yoganand

Table of Contents

List of Figures

List of Tables

Abbreviations

°C	Degree Celsius
3'-OH	3'-Hydroxyl
aa	Amino acid
Abi	Abortive infection
AT sequence	Adenine Thymine sequence
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
bp	Base pair
BSA	Bovine serum albumin
CARF	CRISPR-associated Rossmann fold
Cas genes	CRISPR-associated genes
Cascade	CRISPR-associated complex for antiviral defense
CBS	Cas binding site
Chi site	Crossover hotspot instigator site
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
dcas9	Nuclease-deactivated variant of Cas9
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
DTT	Dithiothreitol
eAgos	Eukaryotic argonautes
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift assay
ExoIII	Exonuclease III
FAM	Fluorescein amidite
FRET	Fluorescence resonance energy transfer

gRNA	Guide RNA
HEPN domain	Higher eukaryotes and prokaryotes nucleotide-binding domain
hrs	Hours
IAS	Integrase anchoring site
IBS	IHF binding site
IHF	Integration host factor
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kb	Kilobases
KD	Dissociation constant
kDa	Kilodalton
LB medium	Luria-Bertani medium
LS	Large subunit
M	Molar
MGE	Mobile genetic element
mins	Minutes
ml	Millilitre
mM	Millimolar
MTase	Methyltransferase
nM	Nanomolar
nt	Nucleotide
NUC lobe	Nuclease lobe
OD600	Optical density at 600 nm
PAGE	Polyacrylamide gel electrophoresis
pAgos	Prokaryotic argonautes
PAM	Protospacer adjacent motif
PCR	Polymerase chain reaction
PDB	Protein data bank
PFS	Protospacer flanking sequence
Pgl system	Phage growth limitation system
PICI	Phage-inducible chromosomal island
PMSF	Phenylmethylsulfonyl fluoride
pre-crRNA	Pre-CRISPR RNA
RAMP	Repeat associated mysterious proteins
REase	Restriction endonuclease

REC lobe	Recognition lobe
R-M system	Restriction modification system
RNA	Ribonucleic acid
RNAi	RNA interference
RNase	Ribonuclease
rpm	Revolutions per minute
SAM	S-Adenosyl methionine
SDS	Sodium Dodecyl Sulphate
sgRNA	Single guide RNA
SS	Small subunit
ssDNA	Single stranded DNA
ssRNA	Single stranded RNA
T5exo	T5 exonuclease
TA system	Toxin-antitoxin system
TAE	Tris-Acetate-EDTA buffer
TBE	Tris-Borate-EDTA buffer
tracrRNA	Transactivating CRISPR RNA
WT	Wildtype
β-ME	β-mercaptoethanol
µg	Microgram
µl	Microlitre
µM	Micromolar

Chapter I

Introduction

1. Chapter I

1.1. Introduction

Among the plethora of living beings that co-exist in nature, prokaryotes compose a simple, unique and primitive class of organisms called archaea and bacteria. Unlike their eukaryotic counterparts, these unicellular creatures are devoid of extensive cellular compartmentalisation. Despite the presence of a cell wall and phospholipid membrane that enclose the cytoplasm, prokaryotes lack a separate nuclear envelope. Here, the genetic material is organised into a highly condensed structure termed 'nucleoid.' Along with replication, transcription and translation, all the other anabolic and catabolic biochemical processes occur in the cytoplasm of prokaryotes.

In contrast to higher-order living forms, prokaryotes are omnipresent on the earth. These creatures comfortably thrive in diverse physical environments that range from snow-capped mountains to volcanic lava, freshwater ponds to marine hydrothermal vents and surface soil to the human gut. To counter these extreme physical challenges in the habitat, prokaryotes demonstrate numerous biochemical pathways such as photosynthesis, heavy metal sequestering and metabolism of toxic chemical compounds. Here, the occurrence of extreme genetic diversity among the different genus acts as a bedrock in imparting the variability in lifestyles to these microbial communities.

Predominantly prokaryotes reproduce asexually via binary fission. Here, the copy of the genome is duplicated and shared to the progeny. In addition to this, they display various horizontal gene transfer mechanisms such as natural transformation, conjugation and phage directed transduction. Many bacteria and archaea possess natural competency and uptake fragments of nucleic acids from the surrounding environment. Recombination and integration of such acquired DNA into the genome or maintaining them as extrachromosomal self-replicative plasmids results in displaying new traits by the hosts to counter various selection pressures. Besides this, enhanced transfer of DNA assisted by conjugation or viral particles (phages) generate a swift variability in the genetic makeup. The process of conjugation between two individual prokaryotes is mediated by extracellularly projected hair-like appendages termed 'pilus'. Whereas, in phage mediated DNA transfer (or transduction) newly generated viral particles encase small pieces of host genetic material along with its genome.

Transmission of such payloads into freshly infected hosts results in the generation of genetic diversity. Here, comprehensive pilus compatibility during conjugation and broad host range of the phages during transduction promote the prolific interspecies transfer of genetic material and result in rapid evolutionary rate. These acquired genetic regions exist as genomic islands and routinely encode proteins that confer accessory yet critical functions such as stress tolerance, antibiotic resistance, defence against virulent phages and pathogenicity. Some such genomic islands termed 'Transposable elements' (or 'jumping genes') generates mutant strains by self-excision and integration at different sites on the genome. DNA transposition in concomitance with conjugation and transduction can engender extreme heterogeneity in genetic composition. Thus, this potentially results in the emergence of resilient prokaryotic variants such as extremely multi-drug resistant pathogenic bacteria.

In addition to hostile physical conditions in the habitat, prokaryotes are often challenged with biological selection factors such as Mobile Genetic Elements (MGEs *viz*., phages and plasmids). Despite conferring genetic diversification, these MGEs conditionally make fatal depletion of host's cellular resources for their proliferation and propagation. Phages are the most abundant forms of life and in comparison to bacteria, they are outnumbered by a tenfold margin (Brussow and Hendrix, 2002). These statistics indicate the extremity of parasitic encounters that prokaryotes face. Owing to such natural selection pressures, prokaryotes developed a myriad of biomolecular defensive pathways that counter phage pathogenesis at various levels.

1.2. The evolutionary arms race between prokaryotic hosts and pathogenic phages

"Red Queen hypothesis" proposes that the continual coevolution of prey and predator maintains the relative fitness and brings in population equilibrium (Van Valen, 1973). The relationship of prokaryotic hosts and pathogenic phages stands out as a befitting example to this hypothesis. Though the host defence mechanisms appear to overpower the infections, in reality, phages swiftly evolve and dispel these challenges in an elegant manner. Here, prokaryotes and phages persistently coevolve and develop various mechanisms (Labrie et al., 2010; Rostol and Marraffini, 2019) and anti-mechanisms (Ofir and Sorek, 2018; Samson et

al., 2013b) to avert the peril of extinction and embrace the genetic diversification. This remarkable trade-off termed "evolutionary arms race" gave rise to numerous defence mechanisms that act at different stages of this prokaryote-phage encounter.

1.3. The life cycle of a bacteriophage and the multi-layered defences of the prokaryotic hosts

Though phage particles demonstrate varied host specificities, the primary stages involved in their life cycle are markedly identical (Figure 1.1). During the initial stage, i.e., adsorption, phages recognise host-specific, surface-exposed cellular components as receptors. Here, phage tail fibres identify the receptors and adhere to the host surface. Upon successful anchoring, the genetic material stored in the capsid of the phage is injected into the bacterial cytoplasm via the tail tube and base plate. This viral genetic material can either be stably integrated into the host genome as a prophage (a process termed "lysogeny") or it can excessively replicate and translate to produce viral proteins by hijacking host cellular resources. Assembly of these viral proteins and packaging of duplicated phage DNA yields numerous phage particles. Upon this, phage-encoded holins and endolysins (viral lysozyme) ruptures the bacterial cells and release the virions (Young, 2014). In contrary to this, during the lysogenic phase, prophage remains latent and is steadily inherited by the progeny. Whereas, a sporadic exposure to extreme stress conditions can induce the prophages to self-excise and undergo a lytic pathway (Figure 1.1).

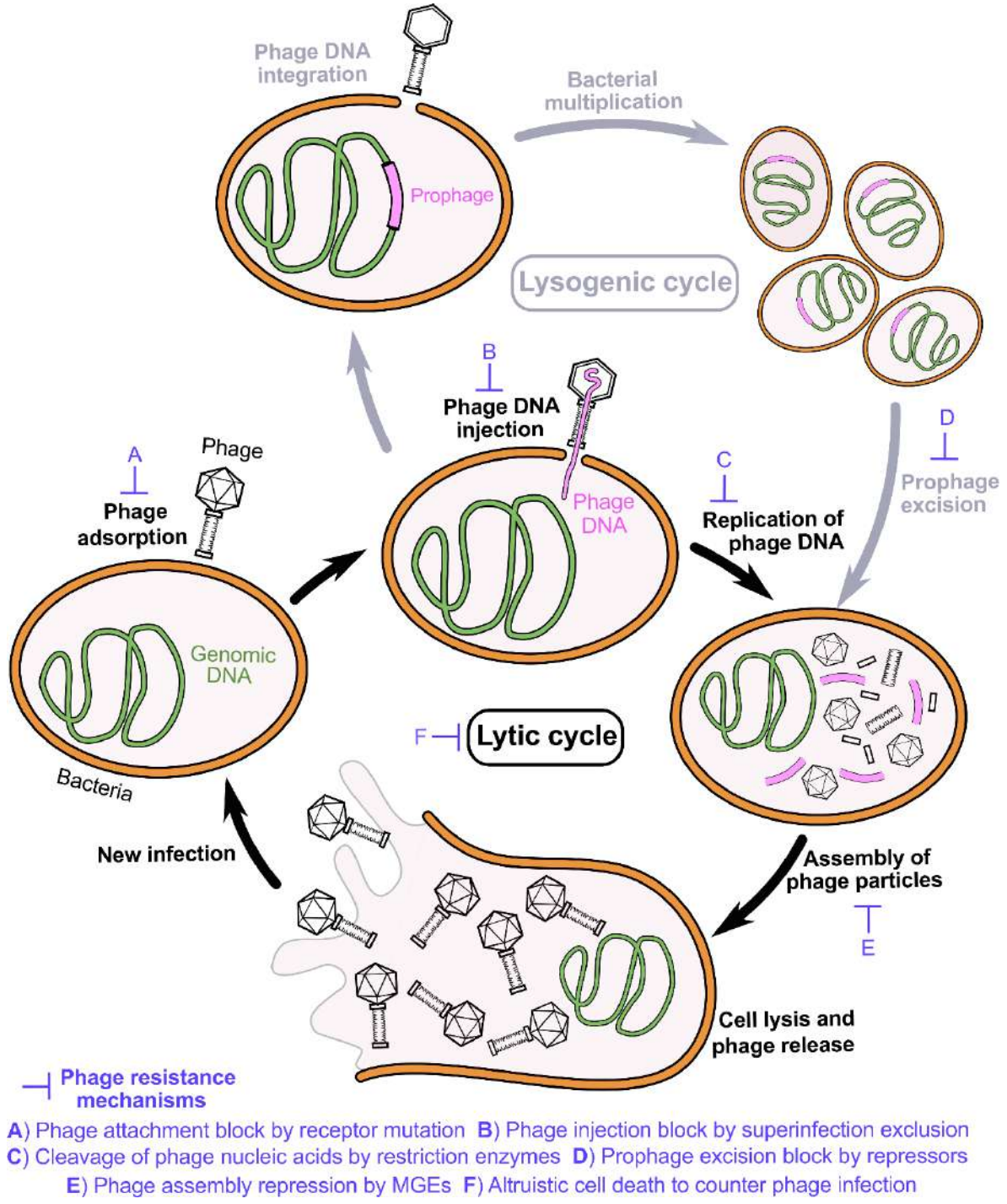


Figure 1.1: Phage resistance mechanisms target different stages of the phage life cycle

Picture describing various steps involved in lysogenic and lytic cycles of phage. Once the phage adsorbs to a prokaryotic cell's surface, it injects the genetic material into the cytoplasm. Upon this, the injected material can replicate and transcribe profusely to generate and release new phage particles by host lysis or enter a lysogenic cycle by integrating phage nucleic acid into the host genome as a prophage. In response to phage infection, prokaryotic

hosts also display countermeasures and evade the phage propagation at various stages (A to E in blue).

To fend off the infections, prokaryotes have developed various types of defence pathways that enact at each level of the phage life cycle (Figure 1.1) (Dy et al., 2014b).

1.3.1. Prevention of phage attachment

Phages choose various surface molecules (such as components of the cell wall, surface proteins, extracellular glycoconjugates, pili and flagella) as receptors by which they adsorb to the host surface (Bertozzi Silva et al., 2016; Steven et al., 1988). Bacteria limits this phage adsorption by masking the receptors with the help of intricate coating such as capsule and biofilms (Branda et al., 2005; Hammad, 1998; Hanlon et al., 2001; Scholl et al., 2005; Vidakovic et al., 2018). In addition to this, hosts modify receptors via mutation (Braun-Breton and Hofnung, 1981; Morona et al., 1984) or post-translational modifications (Harvey et al., 2018) to block phage infection. Few gram-negative bacteria attain adsorption block by outer membrane vesicle (OMV) formation. OMVs pinches off from the cell, during which they seclude the receptor molecules along with adsorbed phages from the host surface (Kulp and Kuehn, 2010; Manning and Kuehn, 2011; Reyes-Robles et al., 2018).

To counter the adsorption block, phages have evolved to produce hydrolytic enzymes to clear-off the barriers (Baker et al., 2002; Castillo and Bartell, 1974; Hanlon et al., 2001; Hynes et al., 1995; Scholl et al., 2001) or choose different host surface molecules as a receptor (Meyer et al., 2012; Werts et al., 1994). In this regard, *Bordetella* phage BPP-1 had developed a new pathway termed as “diversity-generating retroelements” to mutate its tail protein and generate variants up to 10^{12}. Such tail protein mutants choose various host surface molecules as receptors and ward off the adsorption block (Doulatov et al., 2004; Naorem et al., 2017).

1.3.2. Phage adsorption block by MGEs

In many cases, prophages bestow the host with stress tolerance, antibiotic resistance, genetic diversification and biofilm development (Ramisetty and Sudhakari, 2019; Wang et al., 2010; Wang and Wood, 2016). In addition to these benefits, the incumbent prophages also limit future viral invasions via a multitude of mechanisms called superinfection exclusions (Sie). Based on the type of phages that are resisted, Sie is classified into homotypic Sie (blocks similar phages) and heterotypic Sie (blocks different phages).

Sie systems can act at various stages of the phage life cycle. During consequent phage infection, Sie systems encode elements that either physically occludes the receptor (Pedruzzi et al., 1998) or block the receptor assembly (such as pili) (Chung et al., 2014). Thus, by preventing the infection of new phages, prophages establish a kind of symbiotic relationship with the host.

In addition to prophages, various plasmids also confer phage resistance to the host. This is achieved by producing proteins that specifically block phage receptors (Riede and Eschbach, 1986) or by encoding whole pathways that make an extracellular matrix to encase surface receptors (Forde et al., 1999; Forde and Fitzgerald, 2003).

1.3.3. Blocking the injection of phage genome

Once the phages adsorb successfully at the host surface, they transfer the genetic material into the cytoplasm via the cell membrane. Bacterial hosts demonstrate the second level immunity at this stage and block the injection of the viral genome. Generally, pre-infected prophages direct this mechanism by Sie. Here, host cellular components facilitating the phage nucleic acid injection are physically occluded by proteins encoded from incumbent prophages (Ko and Hatfull, 2018; Lu et al., 1993; Rossmann et al., 2004; Sun et al., 2006). In a few hosts, plasmids also confer resistance against phage DNA injection (Garvey et al., 1996).

1.3.4. Inhibition of phage multiplication by selective degradation of viral nucleic acids

Upon successful injection into the host cell, the phage genome can integrate into the host and cause latency via lysogenic cycle or can vigourously replicate and encode new phages by hijacking host cellular resources. To counter such aggravated viral response, bacterial hosts have developed various mechanisms to recognise invading MGE and nucleolytically degrade them. Restriction-modification (R-M) system, phage growth limitation (Pgl) system, prokaryotic argonautes (pAgos) and CRISPR-Cas systems (CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats; Cas: CRISPR associated proteins) are few such examples.

1.3.4.1. R-M systems

R-M systems are the most common intracellular defence systems in bacteria and archaea (Roberts et al., 2015). This system confers immunity through DNA methyltransferase (MTase) and restriction endonuclease (REase). During replication of host genome, MTases recognise particular short sequence (4-8 bp in length) and specifically methylate the adenine or cytosine bases in it (Tock and Dryden, 2005). REase also recognises the same sequence as of MTase and cleaves the DNA. But methylation of endogenous DNA by MTase prevents the nuclease activity of REase. Such molecular tagging by MTase signals the REase to differentiate between endogenous methylated DNA from exogenous non-methylated MGE. During infection, REase cuts the invading phage DNA that is unmethylated and restricts its propagation. Usually, S-adenosylmethionine (SAM), Mg^{2+} and ATP serve as cofactors for the enzymatic activity (Dryden et al., 1993; Dryden et al., 2001; Vovis et al., 1974).

Based on the number of functional subunits, type of target sequence and the mechanism of action, R-M systems are classified into four types (Loenen et al., 2014; Roberts et al., 2003). Among these, type I and type III MTase-REase complexes translocate by pulling of DNA strands upon unmethylated target recognition. In type I R-M system, REase reels in the DNA leading to the collision with another such complex or firm molecular blockade that stalls the translocation and results in DNA cleavage at the impact point (Janscak et al., 1999;

Studier and Bandyopadhyay, 1988). The point of cleavage in type I R-M system is variable and can be thousands of bases away from the recognition site. Unlike type I R-M systems, the target cleavage in type III R-M system can occur only upon head-to-head collision of two oppositely translocating type III R-M complexes (Meisel et al., 1992; Meisel et al., 1995; Reich et al., 2004). During this impact, type III REase subunit cleave at a fixed distance of 25-27 bp from the recognition motif (Bachi et al., 1979; Hadi et al., 1979).

Type II REase and MTase do not form a complex and exert their activities individually. The target sites are usually palindromic and 4-8 bp in length. Here, monomeric MTase methylates the recognition motif in self genomic DNA (Polisky et al., 1975). Type II REase activity does not involve DNA translocation (Sistla and Rao, 2004; Smith and Wilcox, 1970). It targets unmethylated DNA and typically cleaves inside the recognition motif or within a fixed distance of 1-5 bp from the recognition motif (Pingoud and Jeltsch, 2001).

Phages evolve and resist R-M systems by tagging their DNA nucleotides with various chemical modifications (Warren, 1980). For example, T4 phage repels type I-III R-M systems by incorporating hydroxymethylcytosine and glucosyl-hydroxymethylcytosine into the genome (Kornberg et al., 1961). In response to these challenges, prokaryotes have acquired many type IV REases that specifically target the DNA that contains modified nucleotides at the recognition motif. Typically, type IV R-M systems do not have MTase and target the regions that harbour nucleotide modifications such as methylation, hydroxymethylation, phosphorothioation and glucosyl-hydroxymethylation (Loenen and Raleigh, 2014). Among type IV enzymes, *Escherichia coli* McrBC remains to be a well-characterised example. The functional multimeric complex of McrBC comprises two units of McrC sandwiched between two sets of McrB hexamers (Nirwan et al., 2019). Upon identifying modified nucleotide, McrBC remains bound to the recognition motif and translocates by pulling the DNA. The nucleolytic cleavage occurs upon the collision of one more such translocating complex or any other molecular blockade within 3000 bp from the initial recognition site (Panne et al., 1999; Sutherland et al., 1992).

1.3.4.2. R-M like systems

The phage growth limiting (Pgl) system is an example of R-M like system (Chinenova et al., 1982). This system encodes four proteins (PglW, PglX, PglY and PglZ) and the activity of PglX methyltransferase is phase variable. The active form of PglX methyltransferase is produced in Pgl^+ phase. Sporadic frame-shift mutations in *pgl* locus during replication results in inactivated methylase and other Pgl proteins; this phase is termed as Pgl^- (Sumby and Smith, 2003). Upon initial infection, the phage can successfully propagate in Pgl^+ strains. During the release, PglX methylates the phage DNA. These modified phages can reinfect and reproduce in Pgl^- hosts but not in Pgl^+. Methylated phage DNA is recognised as an exogenous target by Pgl^+ and restricted by a predicted nuclease action (Hoskisson et al., 2015).

Intriguingly, genes encoding various phage defence pathways secluded together on the genome as "defence islands" (Makarova et al., 2011c). Therefore, recent genome mining studies relied on the "guilt-by-association" approach and discovered numerous novel immune mechanisms encoded from the prokaryotic defence islands (Doron et al., 2018; Goldfarb et al., 2015; Ofir et al., 2018). Two such identified anti-phage pathways are bacteriophage exclusion (BREX) system (Goldfarb et al., 2015; Gordeeva et al., 2019) and defence island system associated with restriction-modification (DISARM) (Ofir et al., 2018). BREX system is composed of six genes (*brxA*, *brxB*, *brxC*, *brxL*, *pglX* and *pglZ*), out of these two are derived from the Pgl system. Similar to R-M systems, BREX modifies and marks the self-DNA by methylating the recognition motif. Unlike the Pgl system, BREX silences the phage infection upon the first instance of DNA injection itself (Goldfarb et al., 2015; Gordeeva et al., 2019). Surprisingly, BREX lacks nuclease activity; therefore, the mechanism by which BREX constitutes phage resistance remains elusive. DISARM also displays R-M like anti-phage response. This system encodes five genes and methylates the recognition motif of self-DNA. Intriguingly, the deletion of predicted DISARM nuclease gene was shown to confer undeterred phage resistance (Ofir et al., 2018).

1.3.4.3. pAgos

Eukaryotic argonautes (eAgos) are known for their role in RNA-based immunity against viral infections and gene silencing. They generally utilise small ssRNA fragments derived from double stranded RNA as guides. Nucleoprotein complex of eAgo-RNA guide detects and cleaves the target RNA based on complementarity (Muller et al., 2020). Studies had revealed the presence of eAgo homologues in genomic defence islands of the prokaryotes (termed as pAgos) (Makarova et al., 2009). In contrast to eAgos, which confer only RNA-guided RNA interference, pAgos are functionally diverse and exhibit either DNA-guided (Swarts et al., 2015) or RNA-guided (Lisitskaya et al., 2018; Olovnikov et al., 2013) nucleic acid interference. A successful interfering nuclease action necessitates the presence of 5'-phosphate group in the guides and sequence match between guide and the target region (Swarts et al., 2015). In rare exceptions, guides with 5'-hydroxyl groups were also preferred (Kaya et al., 2016). Preferential *in vivo* capture of plasmid derived DNA guides by pAgos suggests their targeting towards MGEs (Swarts et al., 2014). The mechanism of guide acquisition, target recognition and target cleavage are not fully understood. CRISPR-Cas system, another RNA-guided immune pathway targeting the nucleic acids of MGEs, is detailed in Section 1.5.

1.3.5. Blocking of lytic phase induction

Once the phages escape nucleic acids interference systems, viral genome either integrate into the host and remain latent in lysogeny or profusely multiplicate to generate new phage particles and lyse the host cells. In such a scenario, various complex biomolecular switches regulate the decision-making process. For example, in coliphage λ, transcription factor Cro activates lysogeny and also inhibits the expression of another transcription factor CI. In turn, CI inhibits Cro expression and activates the lytic phase. In this way, the CI-Cro bi-stable genetic switch regulates the λ phage life cycle (Fang et al., 2018; Johnson et al., 1981). While maintaining lysogeny, λ prophage produces Cro in abundance to block lytic response and protects the host against cell death. In addition to this, higher intracellular Cro

levels confer Sie against homologous lytic phages by inhibiting CI expression (Johnson et al., 1981).

1.3.6. Hijacking the assembly of bacteriophages

Successful initiation of the lytic program rapidly depletes host cellular resources and leads to the production of viral proteins and nucleic acids in bulk. These molecules assemble to generate virions that are released by killing the host. To counter this challenge, hosts have acquired MGEs such as phage-inducible chromosomal islands (PICIs). This system remains integrated into the host genome and activates only during the lytic phase of certain temperate phages (also known as "helper phages") (Penades and Christie, 2015). Upon activation, PICIs severely retard the viral reproduction by hijacking the phage assembly components and packing its own genetic material. *Staphylococcus aureus* pathogenicity islands (SaPIs) are one such well-studied PICIs that parasitise the lytic phage 80α (Lindsay et al., 1998). The Excision-replication-packaging (ERP) cycle of SaPI is activated during the lytic phase of phage 80α (Tormo-Mas et al., 2010). The PICI viral particles released after ERP cycle infects new cells and remain in lysogeny. Though this mechanism leads to host death and PICI propagation, it severely impedes the propagation of lytic phages to the surrounding population.

1.3.7. Altruistic approaches to counter phage attack

Prokaryotic hosts encompass various stress-responsive systems that cause programmed cell death during the phage infection cycle. Such altruistic approach by the host ensure the survivability of surrounding cells and confer an evolutionary advantage against phage predators. Based on the composition and mechanism of action, these can be categorised as Abortive infection (Abi) and toxin-antitoxin (TA) systems.

Abi systems encode toxic proteins that expressed or activated during the phage infection cycle. These toxins result in fatal membrane damage or perturbation of various key biological pathways such as replication (Emond et al., 1997; Wang et al., 2011), transcription

(Durmaz and Klaenhammer, 2007; Parreira et al., 1996) and translation (Bingham et al., 2000; Morad et al., 1993). Abi toxins either directly interact with essential host factors and inhibit or degrade them. Some Abi toxins shut the activity of host proteins by signal transduction pathways (such as phosphorylation) (Depardieu et al., 2016).

Abi and TA systems have a marginal differentiation and even have overlapped activities in few instances (Dy et al., 2014a; Samson et al., 2013a). TA systems also act as a kill switch and eradicate phage infected cells, in addition to this, TA system encompasses an antitoxin that silences the activity of toxic proteins in unstressed cells that are free of phage invasions (Harms et al., 2018). The toxins in TA system are highly stable and possess lethal activities like blocking cell division (Pimentel et al., 2014), membrane damage (Makroczyova et al., 2014), nucleic acid degradation (Winther et al., 2016), inhibition of essential pathways such as ATP synthesis (Cheng et al., 2014), replication (Aakre et al., 2013) and translation (Schifano et al., 2016). The complementary antitoxins are unstable and the stoichiometry of toxin and antitoxin play a key role in regulating the TA system activity. Antitoxins can either be proteins or RNA. Based on the nature of antitoxin and the mechanism of toxin inhibition, TA systems are categorised into six types (Harms et al., 2018). In addition to phage resistance, TA systems also regulate the maintenance of MGEs such as plasmids (Harms et al., 2018; Ogura and Hiraga, 1983).

1.4. Innate immunity vs adaptive immunity

Based on the specificity and versatility of the infection retaliation mechanisms, the bacterial defence pathways can be categorised as either innate or adaptive. Innate immune systems are in-born and display a generic response towards the pathogenic encounters (all the anti-phage pathways discussed in Section 1.3 belong to innate immune systems). On the other hand, adaptive immune systems are target specific, record and retain the molecular memory of new infections and exert a strong retaliation upon recurring infections. Till date, only the CRISPR-Cas system is known to wield an adaptive immune response in prokaryotes.

1.5. CRISPR-Cas system

Adaptive immunity is believed to be a characteristic of vertebrates. Here, during the life course, the humoral immune response recognises the immunogens and records such encounters as a molecular memory. Such memory guides the adaptive immune response to identify and retaliate against recurring infections with enhanced specificity and efficacy. Owing to the simple unicellular architecture of the prokaryotes, it was strongly believed that they do not harbour sophisticated machinery of adaptive immune system. In such a scenario, the discovery of CRISPR-Cas to act as an adaptive immune pathway had awestruck the scientific community. From the time of discovery to till date, CRISPR-Cas garnered a keen interest and had been widely characterised. In recent years, CRISPR-Cas has emerged as an indispensable genome engineering tool for several labs. True to this spirit, the nobel prize in chemistry 2020 was awarded for CRISPR-based genome editing. Unlike the humoral immunity, where the antibodies bring in countermeasures against immunogens, the CRISPR-Cas wields its response through a non-coding small RNA guided mechanism to destroy the foreign nucleic acids.

1.5.1. Components of CRISPR-Cas system

CRISPR-Cas systems are widespread, ~85% of archaeal and ~40% of bacterial sequenced genomes encompass them (Makarova et al., 2020b). The striking feature of this system is a CRISPR array; it is a region on the genomic DNA, which constitutes a ~20-40 bp of partially-palindromic repetitive sequences (termed as 'repeats') that are intervened by similarly sized but variable sequence stretches (termed as 'spacers') (Figure 1.2). Thus, this gives rise to the name Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs). The mysterious pattern of CRISPR array was noted back in 1987 while analysing the sequence of *iap* gene (isozyme of alkaline phosphatase) in *E. coli* (Ishino et al., 1987). Later investigations revealed that the spacers match phage and other MGE sequences (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005), thus hinting the possibility of the CRISPR system to be a defence system in prokaryotes. Bioinformatic analysis in 2006 had suggested that the CRISPR-Cas systems could function analogously to eukaryotic RNA interference

(RNAi) systems to defend against MGEs (Makarova et al., 2006). Finally, in 2007, experimental demonstration of new spacer acquisition from the phage infections and protection against the recurring phage encounters in *Streptococcus thermophilus* has uncovered the CRISPR-Cas to be an adaptive immune system (Barrangou et al., 2007). To distinguish spacers in CRISPR array from the equivalent sequence in MGE, the latter ones are named as "protospacers" (Deveau et al., 2008). An AT sequence rich leader region of about a few hundred bases in length is usually situated adjacent to repeat-spacer array. The leader sequence is highly conserved at a species level (Jansen et al., 2002) and encompasses a promoter region to regulate the transcription of CRISPR array (Brouns et al., 2008; Lillestol et al., 2006; Lillestol et al., 2009). In addition to various DNA elements (*viz.*, repeats, spacers and leader), a cluster of CRISPR associated (*cas*) genes exist in close conjunction with most of the CRISPR arrays. Presence of helicase and nuclease domains in Cas proteins has lent credence to presume CRISPR-Cas as a nucleic acid-dependent pathway (Jansen et al., 2002). The research advancements in CRISPR-Cas biology had revealed that the Cas proteins act as a workhorse to propel host counterattack. Owing to the expansion in the number prokaryotic genomes being sequenced, numerous variants of *cas* genes were identified (Makarova et al., 2020b). Despite such variability in the Cas proteins among various organisms, the overall mechanism by which the CRISPR-Cas protects the host remains similar.

1.5.2. CRISPR-Cas pathway: mechanism of action

CRISPR-Cas confers adaptive immune response to the host through a three-stage process (adaptation, maturation and interference) (Figure 1.2). During the adaptation stage, the infection of phages or other MGEs is identified and recorded. In this stage, a set of Cas proteins form an adaptation complex (Nuñez et al., 2014; Xiao et al., 2017b) and capture a small stretch of nucleic acid fragment from the MGEs (called "spacer precursors" or "prespacers"). These fragments are integrated at the leader-proximal repeat of the CRISPR array as "spacers" (Barrangou et al., 2007; Yosef et al., 2012). A concomitant duplication of the first repeat accompanies the prespacer integration event (Datsenko et al., 2012; Goren et al., 2012; Yosef et al., 2012), thus resulting in the maintenance of repeat-spacer architecture.

Next, the promoter sequence present in the leader region signals the transcription and leads to the expression of the whole spacer-repeat array as a single long pre CRISPR RNA (pre-crRNA) (Figure 1.2) (Brouns et al., 2008). In the maturation stage, a different set of Cas proteins recognise the repeat regions on the pre-crRNA and process the inert transcript to generate short regulatory mature CRISPR RNA (crRNA). Such crRNA generated by the site-specific endo-RNase activity of maturation proteins contains a trimmed repeat region and the sequence derived from a single spacer unit (Hochstrasser and Doudna, 2015; Punetha et al., 2018).

After the processing step, the crRNA-maturation complex associates with other Cas proteins to form a ribonucleoprotein surveillance complex. Upon the recurring infection, the crRNA acts as a guide RNA (gRNA) and directs the surveillance complex to identify target nucleic acids via sequence complementarity (Figure 1.2). The spacer region in crRNA is derived from the protospacer sequence of the viral nucleic acids during adaptation. Therefore, the complementary interaction of the protospacer with the crRNA symbolises the recurring MGE invasion. Such target recognition event induces structural changes in the surveillance complex and initiates the interference stage. During this step, the surveillance complex recruits interference nuclease at the target region. Upon this, interfering Cas proteins digest the foreign nucleic acids by nuclease action and silence the infection (Brouns et al., 2008; Garneau et al., 2010; Marraffini and Sontheimer, 2008).

The RNAi mechanism by CRISPR-Cas (Hille et al., 2018), pAgos (Lisitskaya et al., 2018) and eAgos (Wilson and Doudna, 2013) are all guided by the complementarity of small RNA towards the target. Unlike the other mechanisms, CRISPR-Cas memorises the infection in the form of acquired spacers and even transfers this genetic memory to the progeny. Owing to these properties, CRISPR-Cas was declared to be a legitimate adaptive immune system in prokaryotes.

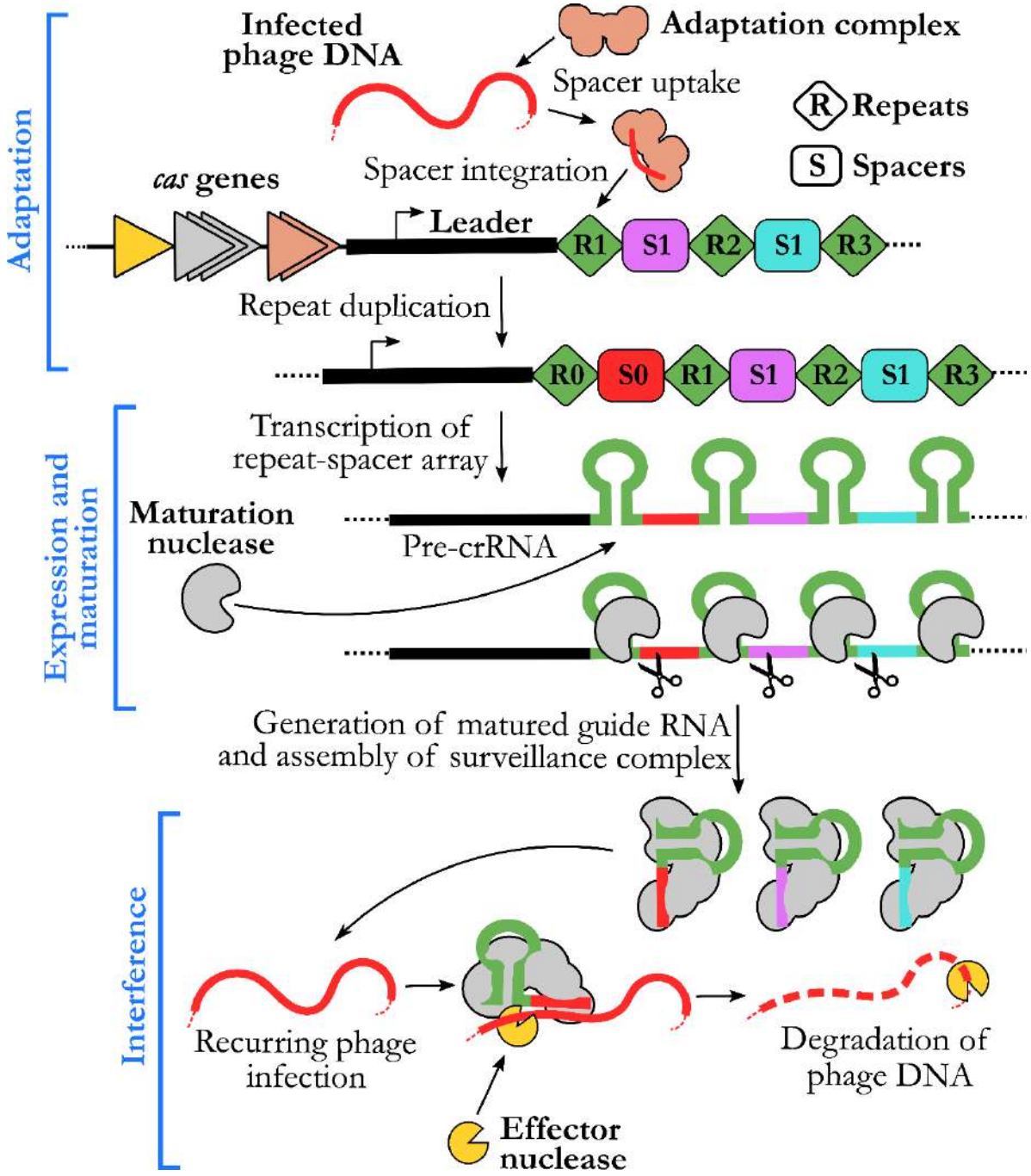


Figure 1.2: Mechanistic features of CRISPR-Cas system

Picture describing three stages of CRISPR-Cas immune response against MGEs (based on type I CRISPR-Cas system). CRISPR locus contains an array of repeats (R1-R3), intervened by spacer sequences (S1-S2). A leader element lies in upstream of the first repeat (R1). A set of *cas* genes is located close to CRISPR locus (Brown, Grey and Yellow triangles). During the adaptation stage, nucleic acid fragment from the infected MGE (red curved line) is derived and incorporated into the CRISPR locus as a new spacer (S0). Adaptation complex

formed by a set of Cas proteins (brown blob) catalyses this integration reaction at the leader proximal end. A repeat duplication event during spacer integration generates a new repeat (R0) in upstream to S0. The promoter located in the leader region signals the transcription of the repeat-spacer array to form pre-crRNA. During the maturation stage, an RNase encoded by *cas* operon recognises the repeat regions and processes the pre-crRNA. This generates a matured RNA (crRNA or guide RNA) with a single repeat-spacer unit. A set of Cas proteins associate with crRNA to form a surveillance complex (Grey blob). Upon recurring phage infection, this complex identifies the foreign nucleic acids by sequence complementarity to crRNA and recruit an interfering nuclease. During the final interference step, the nuclease degrades invaded nucleic acids.

1.5.3. Composition and classification of the CRISPR-Cas system

The above section describes the three stages of CRISPR-Cas in a simplified way. Owing to the remarkable diversity in the *cas* loci composition and Cas protein sequences among various prokaryotic species, the CRISPR-Cas display a high variation in its molecular mechanism. With ever-expanding sequence information in genomic and metagenomic databases, various types of *cas* genes are identified with rapid pace. Additionally, divergent organisation of *cas* operons are evinced due to pervasive exchange of *cas* genes among various hosts and MGEs (Almendros et al., 2014; Garrett et al., 2011; Puigbo et al., 2017). None of the *cas* genes is shared by all the CRISPR-Cas variants. Hence, to classify the CRISPR-Cas system, computational strategies such as protein sequence comparisons, protein phylogenetic analysis, sequence similarity-based clustering and genomic organisation of *cas* genes were used (Makarova et al., 2020b). In addition, experimentally validated data regarding various Cas proteins and structural analysis was also considered as parameters for the classification.

Based on the involvement of Cas proteins in different activities, they are grouped into four distinct functional modules: adaptation (spacer uptake), expression processing (pre-crRNA processing), interference (surveillance complex formation, target binding and target cleavage) and ancillary or signal transduction (cellular regulators involved in accessory CRISPR-Cas functions) modules (Figure 1.3). During the interference step, crRNA-Cas effector complex guides the target recognition and cleavage. Based on the architecture of the effector complex, CRISPR-Cas systems are broadly divided into two classes. In class 1, this complex comprises of multiple Cas proteins, whereas, class 2 variants have a single multidomain Cas effector protein.

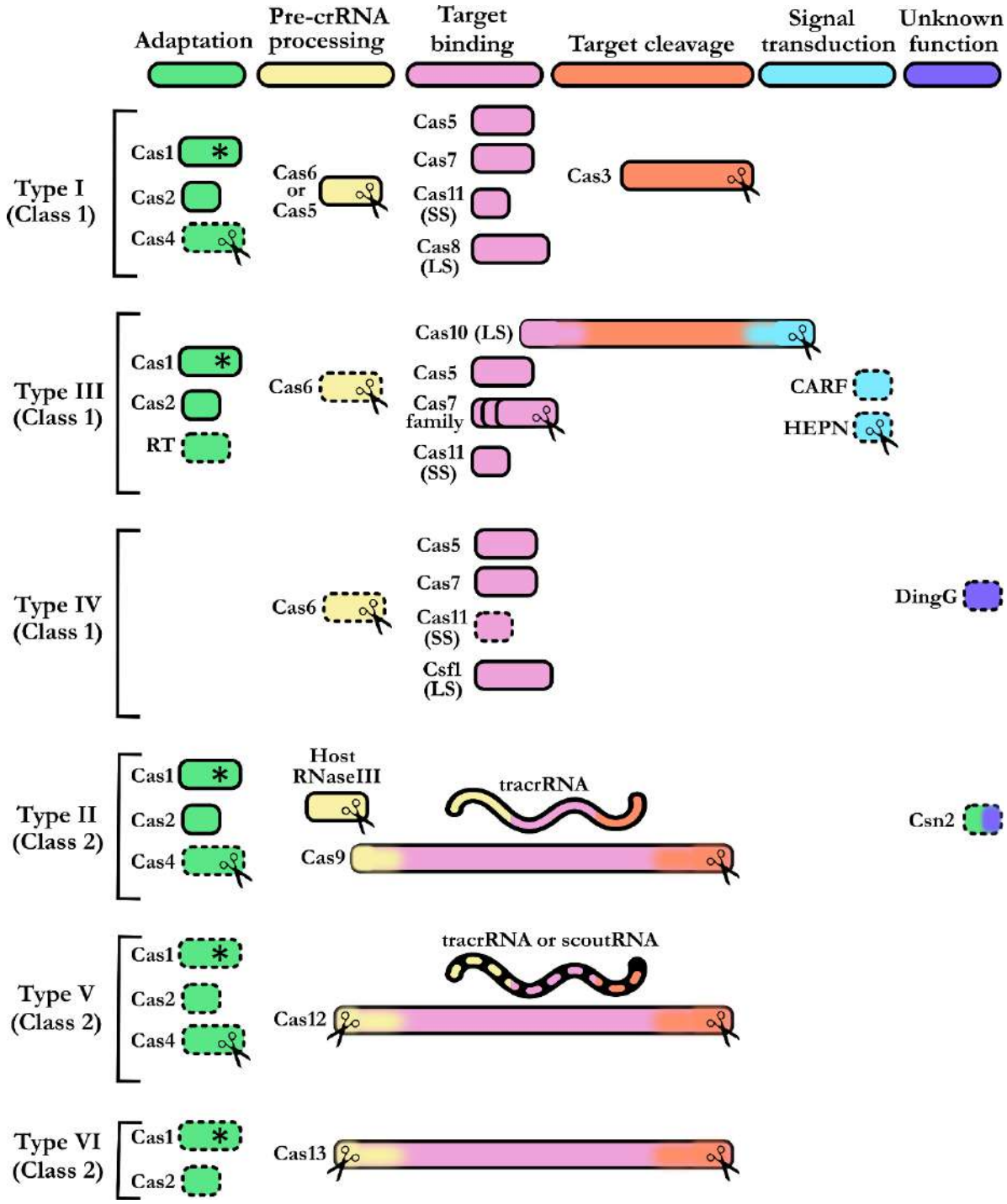


Figure 1.3: Modular organisation and classification of the CRISPR-Cas system

The classification shown here is based on (Makarova et al., 2020b). The *cas* genes present in each CRISPR-Cas type are categorised according to the function displayed by them [adaptation (Green), pre-crRNA processing (Yellow), target sequence binding (Purple), target cleavage (Orange), signal transduction (Cyan) and unknown function (Blue)]. Based on the number of proteins involved in target binding and cleavage CRISPR-Cas systems are

divided into two classes. In class 1, a complex formed by multiple Cas proteins exerts target binding and cleavage, whereas, in class 2 a single Cas protein with multiple domains display this activity. Based on the nuclease involved in interference activity, each class is further divided into three types each [Class I: Type I (Cas3), Type III (Cas7, Cas10 and HEPN nucleases) and Type IV (unknown); Class II: Type II (Cas9), Type V (Cas12) and Type VI (Cas13)]. The key protein catalysing spacer integration and nuclease activity (integration, maturation and interference) are indicated with an asterisk and scissor symbol, respectively. Proteins that act as large subunit and small subunit of target binding complex are depicted as LS and SS, respectively. Multiple colour shading on a few Cas proteins indicates their role in multiple functions. TracrRNA (curved line) element is required in maturation, target binding and interference in all type II and few type V candidates. The dotted boundaries on a few *cas* genes and tracrRNA in type V indicate that these are not conserved among all the individuals of the particular CRISPR-Cas type.

1.5.3.1. Class 1 CRISPR-Cas systems

Based on the type of signature proteins that exerts the interference activity, each class is further divided into three types. In class 1, interference proteins Cas3 and Cas10 are regarded as the signature proteins for type I and III systems, respectively (Figure 1.3). Type IV, the other type in class 1 is uncharacterised and seems to lack any interference proteins. Here a surveillance complex component named Csf2 (Özcan et al., 2019; Pinilla-Redondo et al., 2020) is chosen as a signature protein.

The loci architecture of *cas* operon and variability in the sequence of Cas8 (a large protein component of effector complex) were considered as parameters to categorise the type I system to various subtypes (I-A to I-G) (Makarova et al., 2020b). Generally, in these subtypes, Cas1, Cas2 and Cas4 constitute the adaptation module (Figure 1.3). Among these, Cas1 encompass main catalytic centre for the spacer integration, whereas, Cas2 acts as a scaffold to the structural framework of the active Cas integrase complex. An accessory nuclease Cas4 encompasses a RecB domain and assists in the prespacer trimming. Cas4 is absent in subtypes I-E and I-F and Cas1 is fused to the Cas4 in subtype I-G (I-G was formerly known as I-U). In many of the organisms, solo-Cas4 is encoded outside the *cas* locus or from the MGEs (Hudaiberdiev et al., 2017). During maturation, Cas6 displays an endoribonucleolytic activity and is responsible for the processing of pre-crRNA transcript. Various other proteins like Cas5, Cas7, Cas8 and Cas11 are involved in binding of crRNA to form effector complex. Cas8 and Cas11 are the large (LS) and small subunits (SS) of the

effector complex (Jackson et al., 2014; Jore et al., 2011; Mulepati et al., 2014). Cas5, Cas6 and Cas7 belong to Repeat Associated Mysterious Proteins (RAMPs) family. They encompass a unique protein fold that constitutes the RNA Recognition Motif (RRM or ferredoxin fold) as a core for the ribonucleotide interactions (Makarova et al., 2011a; Wang and Li, 2012). In some of the subtypes, LS and SS are fused as a single protein (Makarova et al., 2020b). Noticeably, type I-C lacks Cas6 and its activity is substituted by Cas5 (Nam et al., 2012; Punetha et al., 2014). In type I systems, Cas3 is signalled to digest the targets by surveillance complex during the interference step. Generally, Cas3 possess an N-terminal HD nuclease domain fused to a helicase domain (Brouns et al., 2008; Sinkunas et al., 2011). Remarkably, in subtype I-F, Cas2 is fused to the N-terminus of Cas3. A newly discovered variant of subtype I-F (I-F3) encode a set of transposases that appeared to be derived from Tn7-like transposons (Peters et al., 2017). This variant lacks interference proteins but the effector complex associates with transposases and carry out site-directed transposition of CRISPR-Cas loci at the sequence matching to crRNA (Halpin-Healy et al., 2020; Klompe et al., 2019).

Type III systems contain Cas10 as an LS of the effector complex (Figure 1.3) (Osawa et al., 2015; Staals et al., 2014). Cas10 encompasses HD nuclease domain and a palm domain (similar to that of nucleotide cyclase and polymerases). Based on the variations in Cas10 and loci architecture, this system is characterised to six subtypes (III-A to III-F). Here, Cas1, Cas2 and Cas6 are present in fewer individuals. It is believed that type III CRISPR-Cas systems could share the missing proteins from the other CRISPR-Cas loci in the host. Notably, adaptation modules in some of the type III individuals harbour reverse transcriptase (as a separate RT or RT-Cas1 fusion or Cas6-RT-Cas1 fusion) to acquire spacers from the RNA phages (Mohr et al., 2018; Silas et al., 2017; Silas et al., 2016). In type III effector complex, Csm2 (subtypes III-A and III-D) and Cmr5 (subtypes III-B and III-C) acts as SS. Various types of Cas5 (type III-A: Csm4; type III-B: Cmr3) and Cas7 (type III-A: Csm3 and Csm5; type III-B: Cmr1, Cmr4 and Cmr6) family proteins form the backbone of effector complex (Guo et al., 2019; Liu et al., 2019; Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015). Despite the small set of *cas* core genes, recent comprehensive computational analysis of CRISPR-Cas loci neighbourhood has identified many uncommon membrane transport proteins and signal-transduction proteins (Faure et al., 2019a; Makarova et al., 2014; Shah et al., 2019; Shmakov et al., 2018). These proteins are believed to be involved in accessory functions of CRISPR-Cas system and are grouped into an ancillary functional module. Majority of these proteins harbour notable features such as CRISPR-associated rossmann fold

(CARF) domain and higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain (Anantharaman et al., 2013; Makarova et al., 2020b). One such protein, Csx1 (type III-B) was found to show similar function as that of Csm6, a core Cas protein of type III-A. The signal transduction between Cas10 and the CARF domain proteins upon target recognition was shown to trigger the non-specific RNase activity by HEPN domain of the latter and restrict the spread of infection (Garcia-Doval et al., 2020; Jia et al., 2019; Makarova et al., 2020a; Molina et al., 2019).

Type IV systems are comparatively less studied among the class I CRISPR-Cas variants. This system encompasses highly diverged *cas* genes that share partial similarities with type I and III effector complex proteins. They are majorly encoded from plasmids and lack adaptation module and interference nucleases (Faure et al., 2019b). Effector complex proteins such as a smaller version of Cas8 (Csf1), diverged variants of Cas7 (Csf2) and Cas5 (Csf3) constitutes this system (Özcan et al., 2019) (Figure 1.3). Previous studies used Csf1 as a signature protein to identify the occurrence of a type IV system (Makarova et al., 2015). Owing to the high level of sequence divergence in the Csf1, a recent computational study has suggested Csf2 to be a better candidate as a signature protein (Pinilla-Redondo et al., 2020). Relying on the genome loci architectures and evolutionary relationships, this study also classified type IV system into five subtypes (IV-A to IV-E). In addition to effector proteins mentioned above, Cas6-like protein (Csf5), DinG helicase (Csf4) and signal transduction protein CysH were also found in few of the type IV individuals (Faure et al., 2019a; Pinilla-Redondo et al., 2020). To compensate the absence of many critical Cas proteins, type IV systems are predicted to display a functional cross-talk with other co-existing CRISPR-Cas systems (mostly type I systems) and share the Cas protein modules (Pinilla-Redondo et al., 2020).

1.5.3.2. Class 2 CRISPR-Cas systems

Class 2 systems constitute a minimal number of Cas proteins (Figure 1.3). Generally, they contain a multidomain effector protein that drives maturation, crRNA binding and interference. In recent times, scientific community harnessed the potential of Cas9, a class 2 effector protein and developed multiple strategies to edit genomes in various organisms. In

the quest of efficient genome editing tool, researchers garnered a lot of interest in class 2 systems. This resulted in a rapid surge in discovery and characterisation of numerous class 2 variants. Recent computational study has classified these systems into 3 types and 17 subtypes (Makarova et al., 2020b). Based on the type of effector nuclease, i.e., Cas9, Cas12 and Cas13, this class is categorised into type II, V and VI, respectively (Figure 1.3).

Cas9, the signature protein of type II system has a bilobed architecture. Here, dsDNA targets were identified by α-helical recognition lobe (REC lobe) and these targets were cleaved by two domains (RuvC and HNH) of the nuclease lobe (NUC lobe) (Jinek et al., 2014). Based on the composition and architecture of CRISPR-Cas loci, type II system is partitioned to 3 subtypes (II-A to II-C) (Makarova et al., 2020b). Though the complex formed by Cas1 and Cas2 is sufficient for the integration of prespacer (Wright and Doudna, 2016; Xiao et al., 2017b), Cas9 and Csn2 (subtype II-A) were also found to be essential for CRISPR adaptation *in vivo* (Heler et al., 2015; Nussenzweig et al., 2019; Wei et al., 2015b). Subtypes II-B and II-C2 lacks Csn2, but harbour Cas4 nuclease (a usual member of the type I adaptation module). Surprisingly, type II-C1 lacks both Csn2 and Cas4. Such architectural differences in type II system highlights the existence of complex variations in the mechanism of spacer uptake. Type II systems display a remarkable difference in the maturation stage. The processing of pre-crRNA requires the base pairing of transactivating CRISPR RNA (tracrRNA) followed by nucleolytic cleavage of cellular factor RNase III (Deltcheva et al., 2011). In addition to this, the binding of Cas9 is also essential for the maturation. Upon pre-crRNA processing, the dual RNA-Cas9 complex acts as effector and brings in the recognition and cleavage of dsDNA viral targets (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012).

Type V systems are recognised by the signature protein Cas12 (Figure 1.3). Like Cas9 nuclease, Cas12 also contain REC and NUC lobes. But in Cas12, NUC comprises of only RuvC like domain (Swarts and Jinek, 2019; Yamano et al., 2016). Based on the architecture of CRISPR-Cas loci and functional diversity of effector protein, type V system is further divided into 10 subtypes (V-A to V-I and V-U) (Makarova et al., 2020b). Unlike others, type V system seems to demonstrate extreme divergence in all three stages of CRISPR-Cas response. Adaptation module contains Cas1, Cas2 and Cas4 (V-A) or Cas1 alone (V-C) or none (V-U). Therefore, type V variants can demonstrate spacer adaptation on their own or could share adaptation Cas components with other co-existing CRISPR-Cas types in the host. During the maturation, Cas12 directly processes the pre-crRNA (V-A) (Fonfara et al., 2016)

or can rely on tracrRNA interactions (V-B). Cas12-crRNA effector complex usually targets dsDNA (Zetsche et al., 2015), in addition; it was observed that the effector nucleases of different subtypes show collateral and targeted activities against ssDNA and/or ssRNA (Chen et al., 2018; Harrington et al., 2018; Yan et al., 2019). Contrasting to other subtypes Cas12g (V-G) displays RNA directed ssRNA targeting (Yan et al., 2019). Subtype V-U5 (or V-K) encodes nuclease null Cas12k along with Tn7- like transposon elements (Shmakov et al., 2017). Akin to variant I-F3, type V-K effector complex lack nuclease action, but it transposes the CRISPR-Cas locus into a target site defined by the sequence of crRNA (Strecker et al., 2019).

Type VI system is an RNA-guided RNA targeting system. The signature protein Cas13 displays a bilobed architecture like other class 2 effectors (Knott et al., 2017; Liu et al., 2017b). Based on the sequence differences among Cas13 and the variation in ancillary module distribution, type VI is characterised into 4 subtypes (VI-A to VI-D) (Makarova et al., 2020b; Smargon et al., 2017). This system seldom harbours adaptation module, but it could also potentially associate with other CRISPR-Cas types in the host to share the spacer acquisition machinery (Hoikkala et al., 2020; Toro et al., 2019). Cas13 alone processes the pre-crRNA (East-Seletsky et al., 2016; Zhang et al., 2019a). Upon this, Cas13-crRNA effector complex identifies the invading viral RNA and exerts the nuclease activity. The NUC lobe of Cas13 comprises of two HEPN domains to exercise the nuclease mediated targeting of foreign RNA (Abudayyeh et al., 2016; East-Seletsky et al., 2017; Shmakov et al., 2015). In addition to this, the structural changes induced by target recognition triggers the non-specific and collateral RNase activity to suppress the invasion of foreign genetic elements (Abudayyeh et al., 2016; Knott et al., 2017; Liu et al., 2017b).

1.5.4. CRISPR adaptation records infection memory in prokaryotes

CRISPR adaptation or spacer acquisition stage builds the infection memory against foreign nucleic acids. A robust CRISPR adaptation confers an effective CRISPR-Cas defence to the host. Presence of this stage certifies CRISPR-Cas to be a legitimate adaptive immune system in prokaryotes. The acquired spacers in the CRISPR array act as a heritable infection memory data bank. Naïve and primed are the two types of CRISPR adaptation mechanisms

displayed by most of the prokaryotic hosts. Naïve adaptation refers to the uptake of spacers from the MGE that is not previously encountered. To fend off the CRISPR-Cas interference, phages mutate their protospacer regions and disrupt the target recognition by effector complex. In response to this, many prokaryotes swiftly and selectively acquire spacers from the mutated invader upon recognition of mismatched priming. This process termed primed adaptation effectively aid the host to counter mutated phages (Datsenko et al., 2012; Jackson et al., 2019; Semenova et al., 2016). The process of spacer uptake was experimentally demonstrated in various type of prokaryotic hosts. Initial studies demonstrated spacer uptake from infected phages (Barrangou et al., 2007) and plasmids (Garneau et al., 2010) by the CRISPR-Cas system of *S. thermophilus*. Till date, type I-E system of *E. coli* is the best-studied model for the CRISPR adaptation.

Cas1 and Cas2 are the fundamental molecular players for the CRISPR adaptation and they are the most conserved proteins of CRISPR-Cas system (Makarova et al., 2020b). In *E. coli*, the promoters of CRISPR-Cas operon are usually repressed by a nucleoid protein H-NS (Pougach et al., 2010; Pul et al., 2010; Westra et al., 2010). Here, either H-NS has to be deleted (Swarts et al., 2012) or an episomal copy of *cas1* and *cas2* should be induced (Diez-Villasenor et al., 2013; Yosef et al., 2012) to activate spacer acquisition. Of all the Cas proteins, Cas1-2 integrase (complex of Cas1 and Cas2) alone suffices for naïve adaptation in *E. coli* (Yosef et al., 2012). In other types, along with Cas1-2, different Cas proteins were also found to be indispensable for spacer uptake (Fagerlund et al., 2017; Heler et al., 2015; Li et al., 2014; Mohr et al., 2018; Vorontsova et al., 2015; Wei et al., 2015b). Despite the extreme variability in the spacer sequence, majority of CRISPR-Cas type I, II and V individuals display a conservation of 2-5 nt protospacer adjacent motif (PAM) at the bordering region of source site on MGE (Deveau et al., 2008; Mojica et al., 2009; Yosef et al., 2012; Zetsche et al., 2015). In addition to specifying the prespacers for integration, PAM also guides interference machinery to differentiate between self and foreign nucleic acids (refer section 1.5.6.1) (Gleditzsch et al., 2019; Mojica et al., 2009; Sashital et al., 2012). The routinely proposed models of CRISPR adaptation is based on the understandings from well characterised type I-E and II-A systems (Hille et al., 2018; Sasnauskas and Siksnys, 2020). The adaptation process can be envisaged to encompass following subset of events: selection and capture of prespacer fragments, processing of prespacer and integration of prespacer into the CRISPR locus. Owing to the diversity of *cas* operon and the type of factors involved in spacer uptake, a variety of deviations in molecular mechanism of adaptation is observed among CRISPR-Cas types.

1.5.4.1. Differentiation of self- versus non-self targets during prespacer selection

Detection and differentiation of foreign nucleic acids during adaptation step is the key for averting autoimmunity by the interference. PAM directed spacer acquisition results in biased uptake of certain sequences, nonetheless, such preference is inadequate to distinguish self-DNA from foreign elements. PAMs are short and host genomes have a same level of PAM distribution like MGE (Levy et al., 2015). Moreover, large quantities of self-targeting spacers were acquired in absence of interference and/or by overexpression of adaptation proteins. In interference active cells, such self-targeting spacers result in autoimmunity and cell death (Levy et al., 2015; Wei et al., 2015b; Yosef et al., 2012). The inability of Cas adaptation machinery to differentiate self- versus non-self targets calls for the presence of supplementary mechanism to avert autoimmunity.

In *E. coli*, prespacers are observed to be sourced from the regions that are prone to double-stranded breaks (DSBs) (such as replication origin and termination sites) (Levy et al., 2015). These regions form the nerve centres for continuous DNA polymerase activity and harbour stalled replication forks. The replication stress at these fragile points result in DSBs. RecBCD mediated DNA recombination process repair these lethal DNA lesions (Figure 1.4A) (Dillingham and Kowalczykowski, 2008). During this process, RecBCD multi-subunit motor unwinds and slices the DNA until it encounters a crossover hotspot instigator (Chi) site. Upon encountering the Chi site, nuclease activity of the RecBCD complex is hampered and restricted to a single strand. Chi site ssDNA region associates with RecA to form a nucleoprotein filament to facilitate homology directed recombinational repair (Del Val et al., 2019; Smith, 2012). The predominant RecBCD nuclease action between DSB and Chi site results in production of ample DNA fragments belonging to this region. The adaptation complex of *E. coli* captures the prespacers from such pool of DNA debris, thus making the replication stalling points and chi sites a hotspot for prespacers (Figure 1.4A) (Levy et al., 2015). *E. coli* genome harbours one chi site for ~4.6 kb, in contrast to this, MGEs rarely have a chi site (Blattner et al., 1997; Dillingham and Kowalczykowski, 2008; Smith, 2001). In addition, many types of MGEs are injected into the host in a linear form (exposed dsDNA ends akin to DSBs) and they are highly replicative (leads to more stalled replication forks). These properties of MGEs allow the RecBCD to degrade larger proportions of extraneous DNA and ensure preferential supply of non-self prespacers to the Cas adaptation machinery

(Figure 1.4B) (Levy et al., 2015). Apart from CRISPR-Cas, RecBCD pathway was shown to assist bacterial argonautes to acquire DNA guides generated from MGEs (Kuzmenko et al., 2020).

RecBCD is not ubiquitous. Gram-positive bacteria contain AddAB helicase-nuclease, a functional paralogue to RecBCD (Wigley, 2013). In type II-A system of *Streptococcus pyogenes*, AddAB facilitates the preferential acquisition of spacers from the regions between injected linear phage DNA end and the nearest Chi sequence. Such bias could potentially serve as a strategy to differentiate self- versus non-self (Modell et al., 2017). CRISPR-Cas operons are diverse and are harboured in a different range of hosts. Therefore, it is well expected to have a variety of mechanisms to differentiate self- versus non-self targets. Future efforts towards understanding this information could unveil a range of elusive pathways involved in the evolutionary arms race of prokaryotes and MGEs.

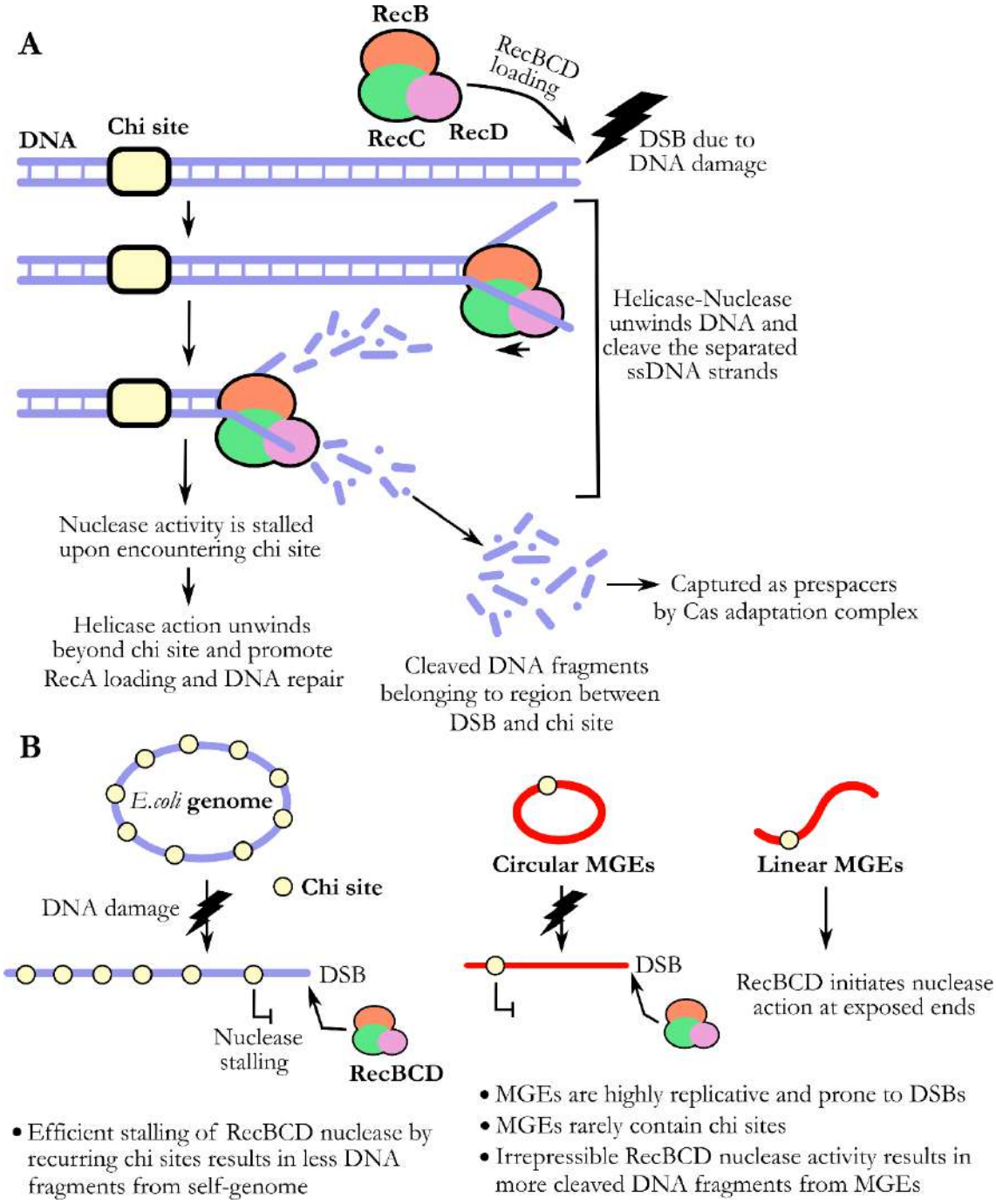


Figure 1.4: RecBCD mediated differentiation of self- and nonself- prespacer selection during CRISPR adaptation

(A) Model depicting the RecBCD directed recombinational repair of DSBs. Upon encountering DSB, RecBCD shreds the DNA region from DSB to Chi site (Yellow). These cleaved DNA fragments are fuelled as prespacer substrates for the Cas adaptation complex.

(B) MGEs are more prone to RecBCD mediated digestion in comparison to the host genome. On average, *E. coli* harbours one chi site per ~4.6 kb in the genome (Blue circle).

Therefore, though the DSBs exist, RecBCD nuclease activity is frequently stalled by recurring chi sites. This results in the generation of a minimal number of digested fragments from the host genome. Unlike this, MGEs (Red) harbour very few or no chi sites. Due to high replicative nature, MGEs are routinely prone to DSBs (few types of MGEs are inherently linear and have exposed DNA ends). Unrestricted RecBCD nuclease activity on MGEs, result in the generation of more digested fragments. Such biased RecBCD nuclease activity towards MGEs result in the generation of more prespacer substrates derived from foreign DNA.

1.5.4.2. Capture and processing of prespacers

Spacers in a CRISPR array are of same length and usually encompass a PAM sequence (type I, II and V) at the protospacer region on the invader. These parameters of spacers are critical for effective interference by CRISPR-Cas system and require stringent selection and processing mechanisms by Cas adaptation complex. A close look at the integrated spacers in *E. coli* shows that the adaptation complex always integrates 33 bp of spacer which abuts a 5'-AAG-3' PAM (where 'G' is destined to be the first residue of the spacer) (Mojica et al., 2009; Yosef et al., 2012). In *E. coli*, prespacers are mostly sourced from remanent DNA fragments from RecBCD activity (Figure 1.4) (Levy et al., 2015). Modulation in helicase, exonuclease and endonuclease activities of RecBCD results in the production of single-stranded DNA fragments ranging from tens to thousands of nucleotides in length (Dillingham and Kowalczykowski, 2008; Muskavitch and Linn, 1982b). The structural framework of Cas1-2 gauges and guides the prespacer selection from the raw material generated by recombinational repair. Cas1-2 integrase is a stable heterohexameric complex of Cas1 and Cas2 (Figure 1.5A) (Nuñez et al., 2014; Wang et al., 2015). This complex recruits the DNA substrate that contains 23 bp duplex with prolonged 3'-overhangs on either end. Also, the C-terminal tail of Cas1 seems to recognize the PAM by stable molecular interactions (Wang et al., 2015). Though Cas1-2 interaction with longer DNA fragments seems to map the prespacer boundaries in *E. coli*, the exact mechanism by which prespacers are trimmed is yet to be understood (Figure 1.5A).

Type I variants (except type I-E and I-F) encode Cas4 accessory nuclease (Makarova et al., 2020b). Recent studies in *Sulfolobus solfataricus* (type I-A), *Sulfolobus islandicus* (type I-A), *Bacillus halodurans* (type I-C), *Synechocystis sp*.6803 (type I-D), *Pyrococcus furiosus*

(type I-B) and *Geobacter sulfurreducens* (type I-G) (Almendros et al., 2019; Kieper et al., 2018; Lee et al., 2018; Liu et al., 2017d; Rollie et al., 2018; Shiimori et al., 2018; Zhang et al., 2019b) highlighted the indispensable role of Cas4 during spacer acquisition. In these organisms, the site-specific nuclease activity of Cas4 was shown to confer PAM selection and prespacer processing. (Figure 1.5B) (Makarova et al., 2018). The protospacers in *P. furiosus* encompass a PAM and downstream motif. Intriguingly, two different Cas4 (Cas4-1 and Cas4-2) are involved in motif selection and prespacer processing. Cas4-1 is encoded from *cas* operon and specifies PAM selection, whereas, Cas4-2 is remotely encoded in the genome and is responsible for selection of downstream motif (Shiimori et al., 2018). *G. sulfurreducens* has a Cas4 protein fused to the N-terminus of the Cas1 (Almendros et al., 2019). Fission of Cas4 and Cas1 proteins abolished the prespacer integration in type I-G. Mutation in RecB nuclease domain of Cas4/I-G hampered the rate of spacer uptake and PAM selection fidelity. Unlike other type I systems, where Cas4 mediated processing is required for sizing of prepacers, the Cas4/I-G nuclease activity was found to be dispensable for trimming the prespacers (Almendros et al., 2019). In one case, it was observed that an extended variant of Cas2 with an unorthodox C-terminus DnaQ exonuclease domain assists *Streptococcus thermophilus* DGCC7710 (type I-E) in prespacer trimming (Drabavicius et al., 2018). Cas2-DnaQ domain fusion is non-ubiquitous, and even model organisms for spacer acquisition studies like *E. coli* (type I-E) do not harbour this.

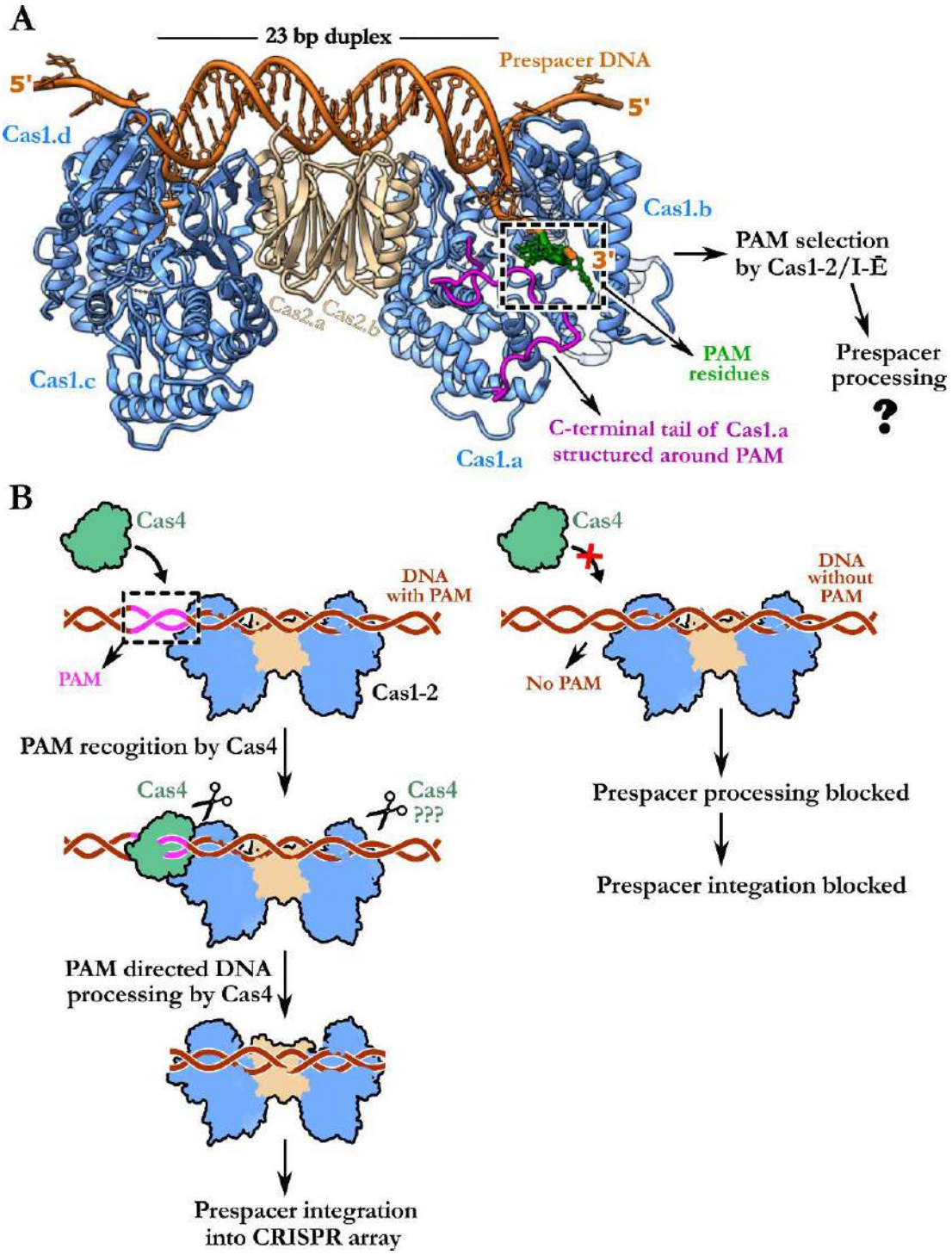


Figure 1.5: Capturing and processing of prespacers by type I system

(A) Structure of *E. coli* Cas1-2-prespacer integrase complex is depicted (PDB: 5DQZ). Four protomers of Cas1 (indicated as Cas1.a-d in Blue) and two protomers of Cas2 (indicated as Cas2.a-b in Tan) are shown. A 23 bp duplex of prespacer DNA (Brown) lies on the Cas1-2 platform. On either end of the prespacer duplex, DNA strands are unwound and the 3'-single strand is captured by the Cas1 protomers. PAM residues (Green) on 3'-strand

are boxed in a dotted line. These residues are recognised by C-terminal tail of Cas1.a. (Magenta). After capturing of prespacer and recognition of PAM by *E. coli* Cas1-2 (type I-E), the longer prespacers are trimmed by an unknown mechanism.

(B) The Cas4 (Green) in CRISPR-Cas subtypes I-A, I-B, I-C and I-D or the Cas4 domain of Cas4-Cas1 fusion in type I-G trims the DNA upon recognising the PAM region (in Magenta) of prespacer bound by Cas1-2. The second copy of Cas4 in type I-B was shown to trim the non-PAM end upon recognising a short motif, whereas, in other subtypes, it is not clear whether Cas4 processes this end. Here, the dual role of PAM recognition and prespacer processing by Cas4 results in the generation of integration competent prespacers.

Unlike type I systems, all the Cas proteins (Cas1, Cas2, Cas9 and Csn2) and tracrRNA of type II-A are essential for naïve adaptation (Heler et al., 2015; Wei et al., 2015b). Cas1-2/II-A display a similar architecture to Cas1-2/I-E, and is believed to define the length of prespacer (Nunez et al., 2015a; Xiao et al., 2017b). Here, Cas9 traces the substrates and specify the PAM containing prespacers for acquisition. Mutation in the PAM interacting domain of Cas9 resulted in acquisition of random prespacers without any PAM specificity. Introduction of nuclease null mutations in Cas9 resulted in unaltered spacer acquisition, suggesting that the Cas9 is not involved in prespacer processing (Heler et al., 2015; Wei et al., 2015b). A new proposal for the mechanism of prespacer capture in type II-A system was given based on recent structural data of type II-A adaptation complex (Ka et al., 2018; Wilkinson et al., 2019). $Cas1_8$–$Cas2_4$–$Csn2_8$ complex binds the free ends of nucleic acids generated by DNA repair machinery (such as AddAB). Csn2 octamer core interacts with DNA and slides the adaptation complex till it encounters a Cas9 bound to PAM. Upon this, the selected prespacers containing PAM are trimmed and loaded on to type II-A adaptation complex by an unknown mechanism.

1.5.4.3. Mechanism of prespacer integration

Though Cas1-2 complex seems to be essential, it is not sufficient for the spacer uptake *in vivo*. Various conserved motifs present in leader and repeat elements guide and ensure fidelity of prespacer integration (Arslan et al., 2014; Goren et al., 2016; Grainy et al., 2019; Kieper et al., 2019; Kim et al., 2019a; McGinn and Marraffini, 2016b; Wang et al., 2016; Wei

et al., 2015a; Yosef et al., 2012). Once the legitimate prespacers are captured and processed, Cas1-2 integrase complex catalyses the prespacer incorporation into the CRISPR array (Li et al., 2014; McGinn and Marraffini, 2016b; Nunez et al., 2015b; Rollie et al., 2015; Wei et al., 2015b; Yosef et al., 2012). The mechanism of prespacer integration is similar to that of transposases and retroviral integrases (Li and Craigie, 2005; Nunez et al., 2015b; Yang et al., 1996). The intrinsic sequence specificity of Cas1 in adaptation complex promotes spacer integration at the leader proximal repeat (Rollie et al., 2015). Here, the 3'-OH ends of the prespacer make nucleophilic attacks at the top strand of leader-1st repeat junction and the bottom strand of 1st repeat-1st spacer junction. These both half-site integration reactions results in ligation of prespacer's 3'-OH ends to the repeat's 5'-phosphate ends (full-site integration product) (Figure 1.6) (Nunez et al., 2015b). Two inverted sequence motifs in the CRISPR repeat anchor the Cas1-2-prespacer complex. These motifs act as a molecular ruler and guide the integrase machinery to make a nucleophilic attack at a fixed distance from them (Goren et al., 2016). Despite the presence of several repeat-spacer units in the CRISPR array, the site of integration of new prespacer has always been at the leader-repeat junction resulting in the integration of the prespacer and concomitant duplication of the first repeat (Diez-Villasenor et al., 2013; Goren et al., 2012; Swarts et al., 2012; Yosef et al., 2012). Polarised prespacer incorporation preserves the chronology of the integration events such that the newest spacer is closer to the leader proximal end and the oldest spacer at the distal end (Figure 1.6). Such bias during integration ensures a quick expression of leader proximal spacers and provide an efficient interference response against recent infections (McGinn and Marraffini, 2016b). Unlike *in vivo* spacer integration, *E. coli* Cas1-2 seems to lack homing site specificity towards the leader-repeat junction *in vitro*, thus leading to integration at all CRISPR repeats and many non-CRISPR sequences (Nunez et al., 2015b). This intriguing observation suggests the involvement of specificity determining factors that guide directional prespacer integration in *E. coli*.

Type I-E adaptation machinery mandates the requirement of longer leader region (~60 bp in *E. coli*) for a successful spacer integration (Yosef et al., 2012). Unlike this, in type II-A system, a repeat proximal short motif of the leader region (~5 bp leader anchoring site (LAS)) was sufficient to promote site-directed spacer integration (McGinn and Marraffini, 2016b; Wright and Doudna, 2016; Xiao et al., 2017b). Type II-A integration mechanism has been studied with purified Cas1-2 integrases from *S. pyogenes* and *Enterococcus faecalis* (Wright and Doudna, 2016; Xiao et al., 2017b). Like type I-E prespacer integration, type II-A is also

driven by 3'-OH attack of the prespacer at leader-repeat junction. Upon recognition of LAS-repeat site, type II-A adaptation machinery integrates the prespacer towards one end of the 1st repeat (half-site integration). The terminal 5 bp motifs on either end of the repeat are critical for recruitment and activity of integrase complex. The 2nd nucleophilic attack of half-site intermediate is contingent upon the length of prespacer and the structural deformities induced by Cas1-2 in repeat region. Failure of 2nd half-site reaction results in branched DNA intermediates due to the integration of prespacer on only one strand of dsDNA. In such a scenario, Cas1-2 disintegrate the prespacer from the half-site integration intermediate and repairs the nick (Wright and Doudna, 2016; Xiao et al., 2017b). This quality control step can ensure the full-site integration of prespacers only at the legitimate target site (i.e., leader adjoining repeat) and can also protect the cells from lethal effects of DNA breaks due to illicit half-site prespacer integrations (Wright and Doudna, 2016).

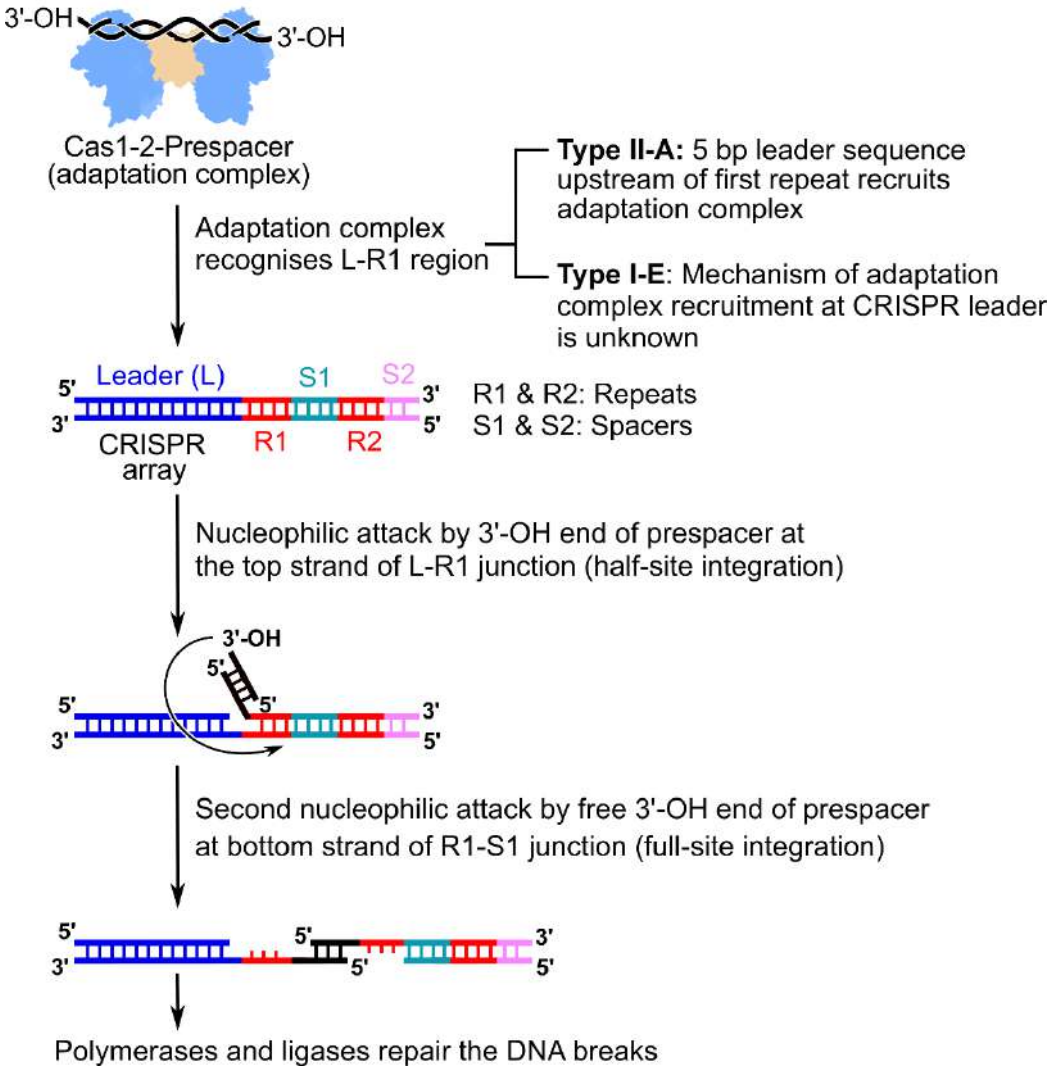


Figure 1.6: Mechanism of prespacer integration

Schematic representation of the steps involved in prespacer integration. Once the Cas1-2 integrase is loaded with the processed prespacer, it recognises the target site at the CRISPR array (i.e., Leader-Repeat1 (L-R1) junction). In type II-A system, 5 bp leader region upstream to first repeat is sufficient for the recognition. Whereas, in type I-E system, the factors directing the target site recognition are unidentified. Once integrase complex anchors to the target site, the 3'-OH end of the prespacer makes a nucleophilic attack at top strand of L-R1 junction and get itself integrated into one strand of CRISPR array (half-site integration). A second nucleophilic attack by the free 3'-OH end on the bottom strand at Repeat1-Spacer1(S1) junction results in full-site integration of the prespacer. A fully expanded CRISPR array is resulted upon sealing of DNA breaks by host polymerases and ligases.

Type III-B system of *Marinomonas mediterranea* (MMB-1) is a unique variant with the capability to acquire both DNA and RNA prespacers. Here, short RNA fragments were ligated into the target site via 3'-OH nucleophilic attack. The integrated RNA prespacer is reverse transcribed to cDNA by the Cas1-reverse transcriptase fusion protein (Silas et al., 2016). The CRISPR adaptation stage is the least understood among the three stages of CRISPR-Cas defence pathway. Present knowledge on the mechanism of spacer uptake is derived from the research conducted on type I and II systems. Future studies focussed on the understanding of CRISPR adaptation in type III to VI system could reveal if the mechanism of spacer capture, processing and integration events are similar to that of type I and II systems. Several CRISPR variants in subtypes III to VI lack few or all adaptation proteins (Makarova et al., 2020b). Therefore, these systems could display a novel mechanism of spacer integration with existing Cas proteins or can share the adaptation machinery of co-existing CRISPR-Cas system.

1.5.5. Expression and maturation of pre-crRNA generates active small RNA guides against MGEs

The infection record (in the form of spacers) of the CRISPR array is transmitted to the interference machinery in the form of small RNA. This process starts with transcription of whole CRISPR array followed by processing of long pre-crRNA transcripts at the repeat sequences. The mature crRNA contains a portion of repeat and protospacer (corresponding to a single repeat-spacer unit).

1.5.5.1. Transcription of the CRISPR array

The AT-rich leader encompasses the promoters to transcribe the whole repeat-spacer array (Brouns et al., 2008; Pougach et al., 2010; Pul et al., 2010). In some cases, transcription of CRISPR array is regulated by the interaction of repressor and activator proteins (Medina-Aparicio et al., 2011; Pougach et al., 2010; Pul et al., 2010; Westra et al., 2010). In other cases, CRISPR expression is constitutive and upregulated upon phage infection (Agari et al., 2010).

Interestingly, each repeat of *Neisseria meningitidis* (type II-C) carries its own promoter and can transcribe individual crRNA. This process obviates the requirement of a separate maturation step (Zhang et al., 2013).

1.5.5.2. Processing of pre-crRNA

Mechanism of crRNA maturation is highly similar across the variants of class 1 systems. In most of the type I and III systems, Cas6 endoribonuclease processes the pre-crRNA (Brouns et al., 2008; Carte et al., 2010; Gesner et al., 2011; Haurwitz et al., 2012; Shao et al., 2016; Sternberg et al., 2012). Cas6 recognises the repeats on pre-crRNA and cleaves within them. Owing to the palindromic nature, most types of repeats form a folded stem-loop structure (Figure 1.7A) (Kunin et al., 2007). Cas6 recognises repeat motifs and stem-loop structure and cuts the region downstream to the stem-loop (Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011). Among type I systems, I-C lacks Cas6 and the maturation of crRNA is performed by Cas5d (Garside et al., 2012; Nam et al., 2012; Punetha et al., 2014). The Cas6 or Cas5d cleavage following stem-loop region leaves the mature crRNA with repeat derived short handle at the 5'-end and stem-loop at the 3'-end (Figure 1.7A). After the cleavage, Cas6 or Cas5d remain bound to the structured repeat and recruit other Cas proteins to form the effector complex (Hochstrasser et al., 2016; Jore et al., 2011; Sashital et al., 2011). Type I-A and I-B contain non-palindromic and unstructured repeats (Kunin et al., 2007). In these systems, usually, a dimer of Cas6 remodels the repeat to form a stem-loop like structure, a favourable conformation for the cleavage (Figure 1.7B). Upon processing the unstructured repeat, Cas6 disassociates from matured crRNA in type I-A and I-B (Reeks et al., 2013; Richter et al., 2013; Sefcikova et al., 2017; Shao and Li, 2013; Shao et al., 2016).

The repeats in type III system are usually unstructured or contain tiny loops (Kunin et al., 2007). Therefore, in order to facilitate cleavage, Cas6 of type III is expected to restructure the repeat region like in I-A and I-B systems (Figure 1.7B). Once the cleavage occurs in repeat region, Cas6 disassociates and do not involve in the assembly of effector complex. A secondary processing step by unknown nucleases at 3'-end of Cas6 leads to trimming of 3'-residual repeat sequence and generates fully matured crRNA (Hatoum-Aslan et al., 2011;

Rouillon et al., 2013; Zhang et al., 2012). Notably, type III-C and III-D lack Cas6 (Makarova et al., 2020b). Like type I-C, Cas5 is predicted to process pre-crRNA in type III-C and III-D.

Csf5, a Cas6 variant of type IV system was shown to facilitate the maturation the crRNA. The structured repeat was cleaved towards 3'-end of the stem-loop. Here, Csf5 remains bound to the processed crRNA and becomes part of the effector complex (Özcan et al., 2019).

Generally, class 2 CRISPR-Cas systems lack Cas6 homologues. Maturation and interference in class 2 are bought about by the same multi-domain effector protein (Figure 1.3) (Makarova et al., 2020b). Though type II effector Cas9 is required for the maturation, it is not sufficient. Here, a *trans*-encoded small RNA termed "*trans*-activating crRNA (tracrRNA)" is required for the crRNA processing. The tracrRNA has a complementary region to the CRISPR repeat and the sequence motifs that are recognised by Cas9 (Figure 1.7C) (Deltcheva et al., 2011; Jinek et al., 2012). The Cas9-tracrRNA complex anchors to the complementary site on repeat sequence of pre-crRNA and forms duplex RNA regions. Host factor RNase III recognises the dsRNA regions and co-processes both CRISPR repeat and tracrRNA (Deltcheva et al., 2011). A secondary processing step by unknown nucleases trims the 5'-residual repeat and a part of spacer sequence in crRNA (Figure 1.7C). Once the processing is finished Cas9:tracrRNA:crRNA complex can detect the viral target and exert nuclease action against them (Deltcheva et al., 2011; Gasiunas et al., 2012; Jinek et al., 2012). Intriguingly, RNase III mediated processing of crRNA is dispensable for the formation of effector complex in type II-C variant. Here, transcription of individual crRNA by the promoters in repeat sequence generates functional crRNA (Zhang et al., 2013).

Cas12, an effector protein of type V system performs crRNA processing and target cleavage. Cas12a of type V-A recognises the structured part of the repeat (Dong et al., 2016; Fonfara et al., 2016) and cleaves few bases upstream of the stem-loop. These intermediate crRNA is further trimmed at both 5'- and 3'-ends by unknown nucleases to form fully matured crRNA with a 5'-repeat derived stem-loop (Figure 1.7D) (Fonfara et al., 2016). Recently numerous variants of type V were reported (Makarova et al., 2020b) and the maturation mechanism in these systems is yet to be characterised. Type V-B, V-E, V-F and V-G contain tracrRNA; therefore, pre-crRNA processing mechanism of these variants might be similar to that of type II system (Makarova et al., 2020b; Shmakov et al., 2015). Maturation in type V-C and V-D variants require a short non-coding RNA termed 'scout RNA' (Harrington et al.,

2020). Unlike tracrRNA, scout RNA have only 5 bp complementarity with repeat sequence of pre-crRNA and is predicted to contain minimal secondary structural folds. Cas12c/d cleaves the pre-crRNA at a fixed length from the dsRNA duplex region (formed by base pairing point of scout RNA and repeat motif of pre-crRNA). Like tracrRNA in type II systems, scout RNA is also an integral component of the effector complex in type V-C and V-D systems (Harrington et al., 2020).

Type VI systems do not harbour tracrRNA. Here, Cas13 effector is sufficient for processing pre-crRNA (Abudayyeh et al., 2016; East-Seletsky et al., 2017; Makarova et al., 2020b; Shmakov et al., 2015; Smargon et al., 2017). Surprisingly, pre-crRNA processing is not essential for interference in type VI-A, although interference efficiency is enhanced when mature crRNA is employed (East-Seletsky et al., 2017). Cas13a recognises the repeat by sequence- and structure-based interactions. Like Cas12a, Cas13a also cuts the repeat at few bases upstream of the stem-loop (Figure 1.7E) (East-Seletsky et al., 2016; Knott et al., 2017). In contrast to the usual architecture of CRISPR array, type VI-B displays variable-sized repeats (36 and 88 bp) in a single CRISPR locus. Despite this, the spacer size is highly conserved (30 bp). Cas13b cleaves at the 3'-end of repeat terminal base; this leads to the generation of mature crRNA with full spacer sequence followed by entire repeat sequence. Cas13b processes long and short repeat variants in a similar fashion, thereby generating mature crRNA of two different sizes (66 nt crRNA contains 30 nt spacer and 36 nt repeat; 118 nt crRNA includes 30 nt spacer and 88 nt repeat) (Smargon et al., 2017).

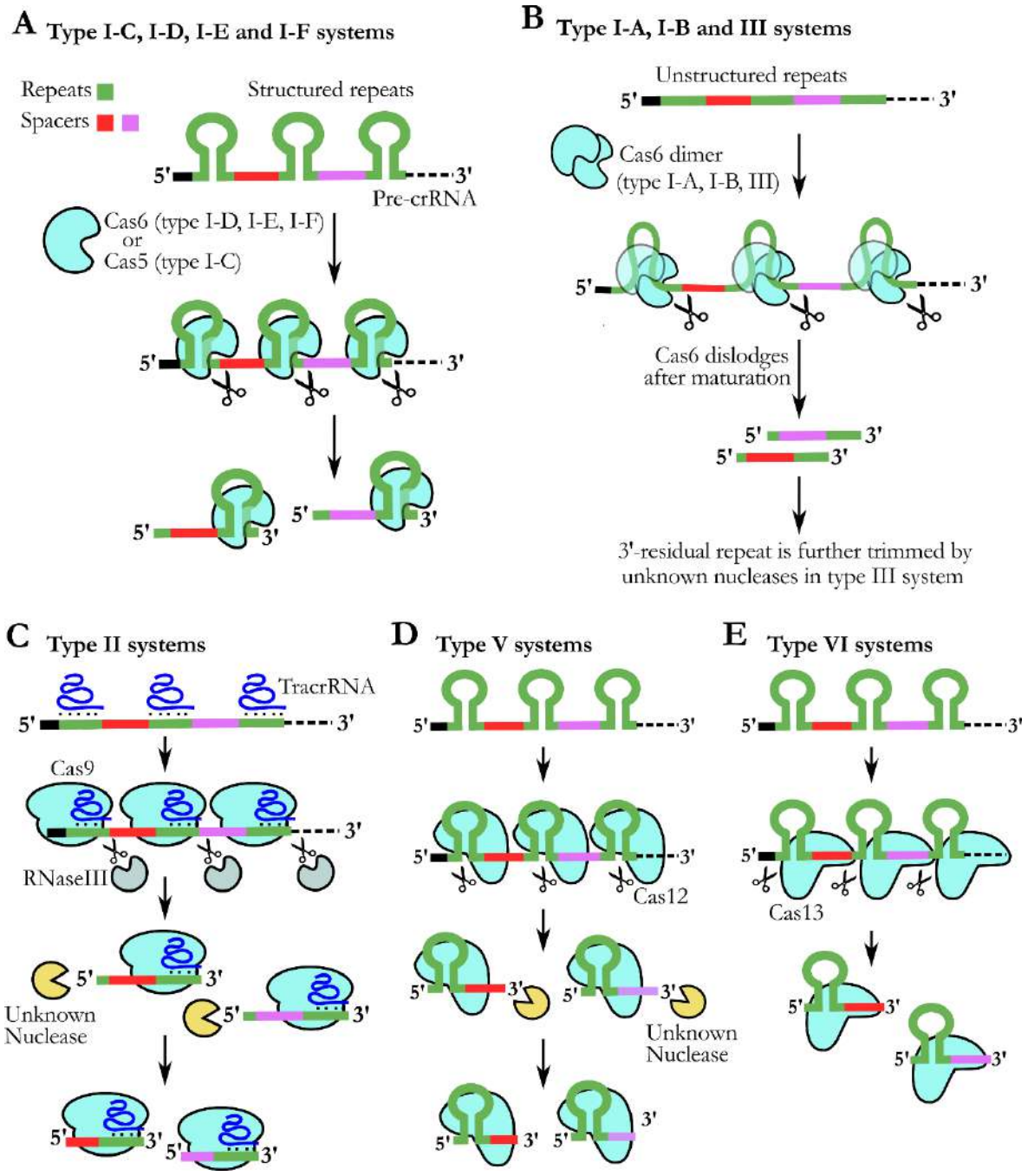


Figure 1.7: Mechanism of pre-crRNA processing in various CRISPR-Cas systems

Figure depicting the mechanism of maturation in CRISPR-Cas types I-VI.

(A) In type I-C, I-D, I-E and I-F, pre-crRNA contains structured repeats (Green). Here, Cas6 (I-D, I-E and I-F) or Cas5 (type I-C) (Cyan) molecule recognises the folded repeat region and cleaves the RNA towards 3'-end. Once processed, the nuclease remains bound to the crRNA.

(B) The pre-crRNA of type I-A, I-B and III encompass unstructured repeats. Here, a Cas6 dimer (Cyan) partially folds the repeat and cleaves towards 3'-end of the RNA. After

processing, the Cas6 nuclease dislodges from the crRNA. In type III systems, the 3'-residual repeats are further trimmed by unidentified RNases.
(C) Type II systems require a tracrRNA (Blue) for the maturation. Here, tracrRNA bound to Cas9 (Cyan) interacts with complement sequence on repeat regions of pre-crRNA. Host RNaseIII (Grey) recognises the duplexed RNA regions towards 3'-end and process them. The crRNA is further trimmed and even a part of spacer is cleaved off by unidentified nucleases (Yellow). Cas9-tracrRNA-crRNA complex remains stable and acts as an effector nuclease.
(D) In type V systems, Cas12 interacts with folded repeat regions of the pre-crRNA. Upon binding, Cas12 cleaves the RNA towards 5'-end. Next, crRNA is further trimmed on both 5'- and 3'-by unidentified nucleases. After processing, Cas12 remain bound to the crRNA.
(E) In type VI systems, Cas13 interacts with folded repeat regions of the pre-crRNA. Upon binding, Cas13 cleaves the RNA towards 5'-end at repeat-spacer boundary. Cas13-crRNA complex

1.5.6. CRISPR-Cas interference silences the invasion of MGEs

The targeted cleavage of foreign nucleic acids is the primary purpose of CRISPR-Cas defence pathway. In the interference stage, matured crRNA interacts with various Cas proteins and form an effector complex. Guided by the crRNA, effector complexes scout, identify and target the protospacer regions on MGEs. The interference machinery comprises of multiprotein-crRNA complex in class 1 systems and a multidomain protein-crRNA complex in class 2 systems (Makarova et al., 2020b).

Type I CRISPR-Cas systems deploy "CRISPR-associated complex for antiviral defense (Cascade)", a crRNA-multiprotein complex for surveillance to detect complementary targets. Here, the type I signature protein Cas3 acts as an interfering nuclease and cleaves the targets detected by the surveillance complex (Brouns et al., 2008). Amongst all type I variants, the cascade structure and function of type I-E system is well characterised. In addition, type I-E cascade contains all the known type I surveillance Cas proteins (Makarova et al., 2020b). Usually, there are five Cas family proteins involved in cascade formation (i.e., Cas5, Cas6, Cas7, Cas8 and Cas11). Among these, Cas5, Cas6 and Cas7 belong to RAMP family and they contain RRM in its structural core for interacting with RNA (Brouns et al., 2008; Jore et al., 2011). The type I-E cascade is seahorse-shaped (Figure 1.8A) (Jore et al., 2011; Wiedenheft et al., 2011). A single molecule of Cas6e is bound to the 3'-repeat derived stem-loop of crRNA. After maturation of pre-crRNA, Cas6e remains bound to its recognition region on the

repeat (i.e., 3'-stem-loop) (Gesner et al., 2011; Sashital et al., 2011). Here, Cas6e forms a head portion of seahorse-shaped cascade and this promotes the recruitment of other Cas subunits. The backbone of the complex is formed by six structurally interlocked palm-shaped Cas7e subunits. The 5'-repeat derived region of crRNA is bound by palm-shaped Cas5e (Jackson et al., 2014; Jore et al., 2011; Mulepati et al., 2014; Wiedenheft et al., 2011; Zhao et al., 2014). A thumb and palm-like structures of Cas5 and Cas7 are connected by a web-like fold. Thumb motifs of Cas5e and Cas7e introduce kinks in crRNA by flipping the nucleotide base in the reverse direction (Figure 1.8A). These kinks are introduced at regular intervals and are positioned on to the top of web-like fold. The flipping of nucleotides starts from the last residue of 5'-repeat (by Cas5e) and occur at every sixth nucleotide. Five residues between two kinks are held firmly by Cas7e subunits. This confers an efficient base pairing between protospacer and crRNA during target recognition step (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). Cas8e and Cas11e are the respective large subunit (LS) and small subunit (SS) of the cascade. Two Cas11e subunits interact with Cas7e at belly region of the cascade. A single copy of Cas8e recruited towards 5'-end of crRNA interacts with Cas5e, Cas7e and Cas11e to form the tail portion of the cascade (Figure 1.8A) (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014).

During the surveillance step, the target identification at protospacer is initiated by the recognition of PAM. Cas8 identifies the PAM region (Figure 1.8A) and initiates various structural changes that unwind the proximal DNA region and facilitate target recognition by the strong base pairing of crRNA with protospacer (Hayes et al., 2016). The interaction of PAM proximal protospacer residues (termed 'seed sequence') with crRNA is critical for the interference. Any mutations in PAM and seed motif (except the flipped base at the kink on crRNA) hampers the target recognition and interference (Datsenko et al., 2012; Semenova et al., 2011; Xiao et al., 2017a). The protospacer region complementary to crRNA is termed 'target strand' and the other one is called 'non-target strand'. During the cascade loading, various structural rearrangements allow the formation of R-loop (a three-stranded nucleic acid loop comprising of DNA:RNA hybrid and an unpaired DNA strand). The binding of the target strand to helical crRNA is firmly supported by Cas7e backbone and the free non-target strand is stabilised via interactions with Cas11e belly (Hayes et al., 2016; Xiao et al., 2017a). The architecture of cascade and the mechanism of target recognition is mostly conserved with few subtype-specific variations in type I (Hille et al., 2018; Makarova et al., 2020b). For example, the type I-F cascade lacks the belly region and displays a closed ring-like architecture, which

opens upon the target binding (Chowdhury et al., 2017; Guo et al., 2017). Except for type I-A and I-E, the cascades lack a separate LS or SS (Makarova et al., 2020b). Type I-C has a minimal cascade with the absence of Cas6 and fusion of Cas8-Cas11 (Hochstrasser et al., 2016; Nam et al., 2012; Punetha et al., 2014). A notable feature in type I-D is the presence of fused Cas10 (LS subunit in type III)-Cas11. Therefore, type I-D could be a potential connecting link between type I and III systems (Makarova et al., 2020b).

Once the bona fide target is identified by cascade, the resulting structural rearrangements in LS and SS allow the recruitment of Cas3 interfering nuclease (Hayes et al., 2016; Xiao et al., 2017a). The mechanism of target cleavage by Cas3 seems to be conserved in most variants of type I (Figure 1.8A) (Nimkar and Anand, 2020; Rollins et al., 2015). Once recruited at the protospacer, the HD domain of Cas3 makes a single nick on to the non-target strand. This results in proper positioning of Cas3 helicase domain and activation of ATP-dependent DNA reeling. Presence of any molecular blockade obstructs the translocation by Cas3 and stimulate HD domain to cleave the target. Cas3 helicase action and ssDNA cleavage releases the fragments and creates ssDNA gaps (Dillard et al., 2018; Nimkar and Anand, 2020; Westra et al., 2012; Xiao et al., 2018; Xiao et al., 2017a; Zhao et al., 2014). Usually, the DNA repair machinery of the host could reseal ssDNA gaps. Therefore, the mechanism by which the Cas3 could achieve complete interference is yet to be understood. Cas3 also demonstrates cascade-independent degradation of nucleic acids *in vitro* (Mulepati and Bailey, 2013; Nimkar and Anand, 2020; Sinkunas et al., 2011; Sinkunas et al., 2013). It would be interesting to probe if such activity by Cas3 could silence the MGE infection.

Type III-A and III-B utilise cascade like Csm and Cmr, respectively for target recognition and interference (Jackson et al., 2014; Makarova et al., 2020b; Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015). But, unlike type I, type III systems target both RNA and DNA (Deng et al., 2013; Elmore et al., 2016; Goldberg et al., 2014; Samai et al., 2015). The interference model in type III system is based on the understandings from the studies on type III-A and III-B. Other subtypes, III-C to III-F are discovered recently and remain uncharacterised. Cas6 in type III systems is not a part of the surveillance complex (Figure 1.8B). The crRNA in type III surveillance complex is held by Cas5 (III-A: Csm4; III-B: Cmr3) at 5'-end and Cas7 (III-A: Csm3 and Csm5; III-B: Cmr4, Cmr6, and Cmr1) at the spacer part as a backbone. Similar to the type I system, Cas7 family proteins introduce kinks in crRNA and hold the guide firmly. Cas10 (III-A: Csm1; III-B: Cmr2) and Cas11 (III-A: Csm2; III-B:

Cmr5) function as an LS and SS, respectively (Figure 1.8B) (Jackson et al., 2014; Makarova et al., 2020b; Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015). Once the target RNA transcript is bound by the surveillance complex, the Cas7 subunits cleave the target at every sixth base. In type III systems, cleavage of target DNA is conditional and can occur only during co-transcriptional recognition of the target ssRNA by Csm/Cmr complex. Here, HD domain of Cas10 introduce breaks on template DNA (Figure 1.8B) (Osawa et al., 2015; Samai et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014; Taylor et al., 2015). In addition to this, palm domain of Cas10 converts ATP to cyclic adenylates upon target recognition (Kazlauskiene et al., 2017; Niewoehner et al., 2017). The CARF domains of auxiliary protein Csm6 (type III-A) or Csx1 (type III-B) recognises the cyclic adenylates and activates the HEPN domain to display promiscuous RNase activity on cellular and invader transcripts and arrest the cell growth (Figure 1.8B) (Anantharaman et al., 2013; Deng et al., 2013; Hatoum-Aslan et al., 2014; Jiang et al., 2016b; Makarova et al., 2020a; Niewoehner and Jinek, 2016; Sheppard et al., 2016). Unlike type I systems, PAM is not required for target cleavage in type III. Contrastingly, the DNA targeting activity of Cmr2 of *P. furiosus* demonstrates the requirement of PAM sequence on target RNA (rPAM) (Elmore et al., 2016).

The type IV systems contain Cas5, Cas7 and Cas8 family proteins, but lack interference machinery (Özcan et al., 2019). It is predicted that the type IV system could share the interference machinery derived from the co-existing CRISPR-Cas systems in a host (Pinilla-Redondo et al., 2020). Structural and functional characteristics of interference components in this system are yet to be understood.

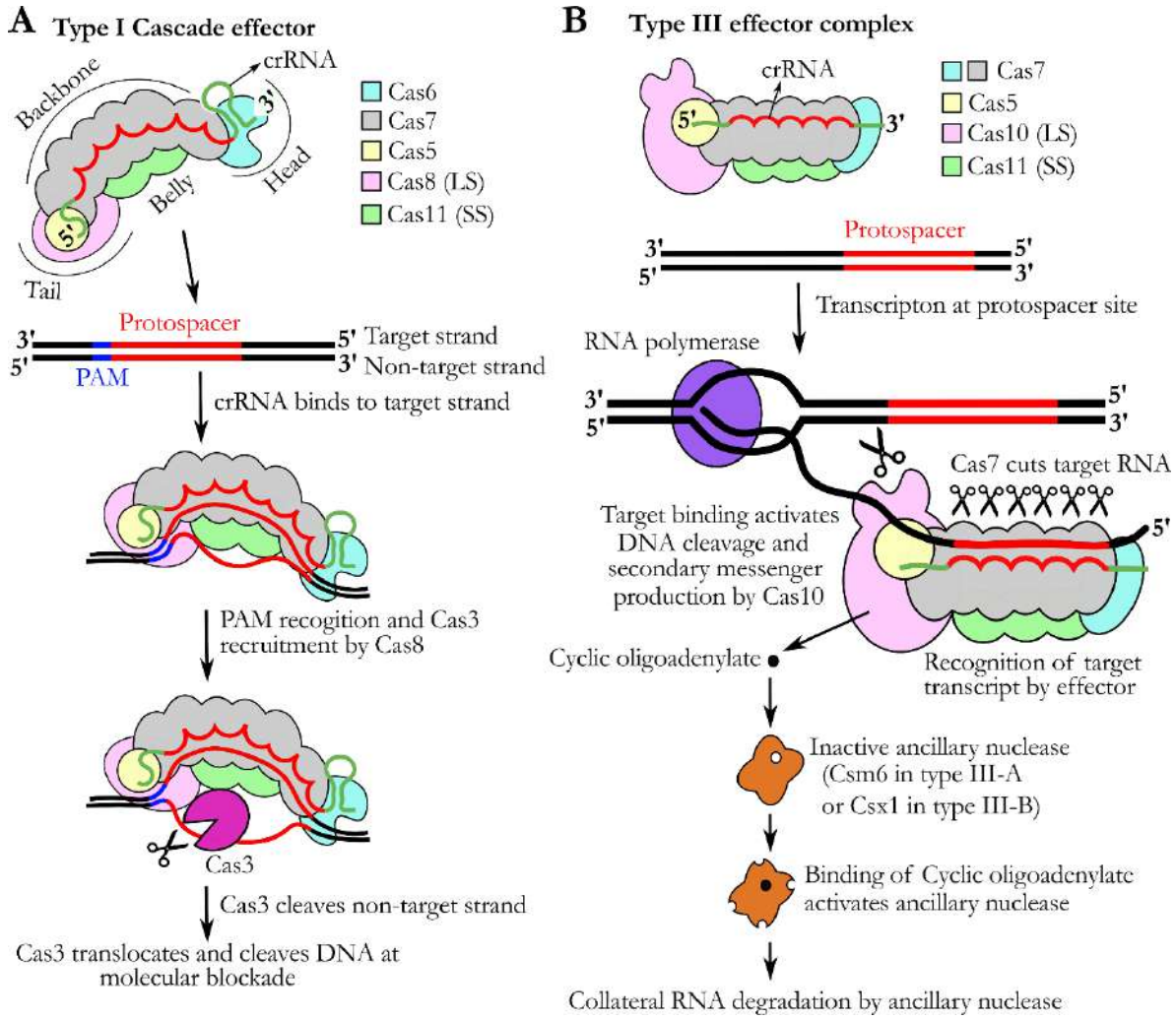


Figure 1.8: Mechanism of interference in class 1 CRISPR-Cas systems

(A) The type I systems have a seahorse-shaped Cascade complex. The head region of the complex contains Cas6 (Cyan) and holds the 3'-end of the repeat structure. Multiple Cas7 (Grey) molecules assemble to form backbone part of the complex. The Cascade backbone holds the spacer part of crRNA and introduces the kinks. Cas5 (Yellow) interacts with the 5'-end of the crRNA. The belly region of Cascade is formed by Cas11 (small subunit) (light green). The Cas8 (large subunit) (Pink) interacts with Cas5, Cas11 and backbone subunits to form the tail part of Cascade. During surveillance, Cas8 of the Cascade recognises the PAM (Blue) and target protospacer sequence (Red) is bound to the crRNA. These interactions signal Cas3 nuclease (Magenta) to bind and nick the non-target strand. Later, the helicase domain of Cas3 catalyses the translocation by pulling the non-target strand. During translocation, the nuclease domain of Cas3 cleaves the foreign DNA upon encountering the molecular blockade.

(B) The Cascade like effector complex of type III encompass Cas7 family proteins (Cyan and Grey), Cas5 (Yellow), Cas10 (LS) (Pink) and Cas11 (SS) (Light green). Type III effector complex recognises RNA as a target. During the transcription of the protospacer region, type III effector complex interacts with the complementary part of the RNA (Red). Once the complex is bound, Cas7 subunits cut the target RNA at multiple points. The

target binding also activates the DNase activity of the Cas10, which cleaves the foreign DNA at the proximity to the protospacer region. During the targeting step, Cas10 also produces secondary messengers. These molecules bind and activate ancillary HEPN nucleases (Brown) to cleave RNA indiscriminately.

A multidomain effector nuclease engenders the targeting activity in class 2 CRISPR-Cas systems. The type II interference model is based on the study of type II-A Cas9 activity from *S. pyogenes*. The effector complex of type II system comprises of Cas9, matured crRNA and a tracrRNA (Figure 1.9A) (Gasiunas et al., 2012; Jinek et al., 2012; Jinek et al., 2014; Nishimasu et al., 2015; Nishimasu et al., 2014). The recognition and nuclease lobes of Cas9 are connected by a positively charged bridge helix. The nuclease lobe contains HNH and RuvC nuclease domains to exert targeting activity. Binding of crRNA induces a conformational change in Cas9 and facilitate the PAM and protospacer identification. Bridge helix interacts and stabilises the crRNA (Jiang et al., 2016a; Jinek et al., 2014; Nishimasu et al., 2015; Nishimasu et al., 2014). During surveillance step, the C-terminal PAM interacting domain recognizes the PAM sequence on 3'-end of the non-target strand and induce the target DNA unwinding. Perfect base pairing of crRNA with seed region results in the further unwinding of protospacer by Cas9. The complete base pairing of crRNA induces structural changes to stabilise the R-loop and activate nuclease domains (Anders et al., 2014; Jiang et al., 2016a; Jinek et al., 2012; Mekler et al., 2017; Sternberg et al., 2014). Well-poised catalytic centres of HNH and RuvC domains cleave at the targeting and non-targeting DNA strands, respectively (Figure 1.9A). This cleavage results in a blunt DSB from the fixed distance of the PAM sequence (Garneau et al., 2010; Jinek et al., 2012; Sternberg et al., 2015). Due to the well-defined positioning of the cutting site by the Cas9, the research community is utilising CRISPR-Cas9 as a genome-editing tool to create targeted mutations in organisms belonging to diverse kingdoms of life. Molecular engineering and generation of single guide RNA (sgRNA) have been a major breakthrough to ease the utilisation of CRISPR-Cas9 tool in genome editing applications. The sgRNA is a chimeric RNA and constitute essential elements for target recognition from both tracrRNA and crRNA. The maturation step is avoided using sgRNA; here, Cas9 can directly bind with sgRNA and perform target cleavage (Jinek et al., 2012). The significant applications by CRISPR-Cas9 tool had encouraged the scientists to engineer Cas9 to enhance the cleavage efficiency, broaden the PAM specificity and reduce

the off-targeting (Anzalone et al., 2020; Kondrateva et al., 2020; Manghwar et al., 2020). Also, a major focus has been emphasised to discover and characterise new orthologues of Cas9. For example: compared to routinely used *S. pyogenes* Cas9 (1368 aa), the *Campylobacter jejuni* Cas9 is small (984 aa) and easily deliverable into eukaryotic nuclei (Kim et al., 2017). Though being larger in size, *Francisella novicida* Cas9 (1629 aa) demonstrates high-fidelity in target selection and cleavage (Acharya et al., 2019).

Interference in type V system is mediated by Cas12-crRNA effector complex (Figure 1.9B) (Makarova et al., 2020b). In addition to this, tracrRNA is also required for interference activity in a few variants like type V-B (Liu et al., 2017a; Shmakov et al., 2015). The nuclease lobe of Cas12 contains only RuvC domain. During target surveillance, PAM recognition and protospacer base pairing with crRNA brings in various structural changes and positions the RuvC domain for the cleavage. A single catalytic site of RuvC domain alone cut both the target and non-target strands of the protospacer DNA (Dong et al., 2016; Gao et al., 2016; Shmakov et al., 2015; Wu et al., 2017; Zetsche et al., 2015). Unlike blunt-ended cleavage of Cas9, Cas12 results in sticky DSB with 5-7 nt overhangs (Figure 1.9B). Cas12a even display robust and non-specific endonucleolytic degradation of ssDNA upon binding to the target DNA (Chen et al., 2018).

Type VI systems are the only RNA-guided RNA-targeting class 2 types. The interfering RNase activity is displayed by Cas13-crRNA effector complex (Figure 1.9C). Cas13 contain two HEPN domains in the nuclease lobe (Anantharaman et al., 2013; Makarova et al., 2020b); they not only catalyse the target RNA cleavage but also show collateral ssRNA cleavage activity upon target binding (Abudayyeh et al., 2016; East-Seletsky et al., 2017; Shmakov et al., 2015). Recruitment of crRNA onto Cas13 results in structural rearrangements to assist target binding (Liu et al., 2017b; Liu et al., 2017c). The maturation of crRNA seems to be dispensable for interference activity of type VI systems (East-Seletsky et al., 2017). Notably, Cas13 effector complex interacts with the conserved protospacer flanking sequence (PFS) on one (type VI-A) (Abudayyeh et al., 2016; East-Seletsky et al., 2017; Liu et al., 2017c) or both ends (type VI-B) (Smargon et al., 2017) of the target ssRNA. A perfect match between the crRNA and the central seed region of protospacer is critical for target recognition and degradation (Figure 1.9C). Upon latching onto the target, catalytic sites of two HEPN domains are positioned to close proximity (Liu et al., 2017b; Liu et al., 2017c). Unlike all other effector nucleases, the composite catalytic site of Cas13 is located on the exterior

surface. Activated Cas13 degrades both target RNA and collateral ssRNA (Figure 1.9C). This action not only silences the infection but can also lead to host dormancy and death (Abudayyeh et al., 2016; East-Seletsky et al., 2017; East-Seletsky et al., 2016; Smargon et al., 2017).

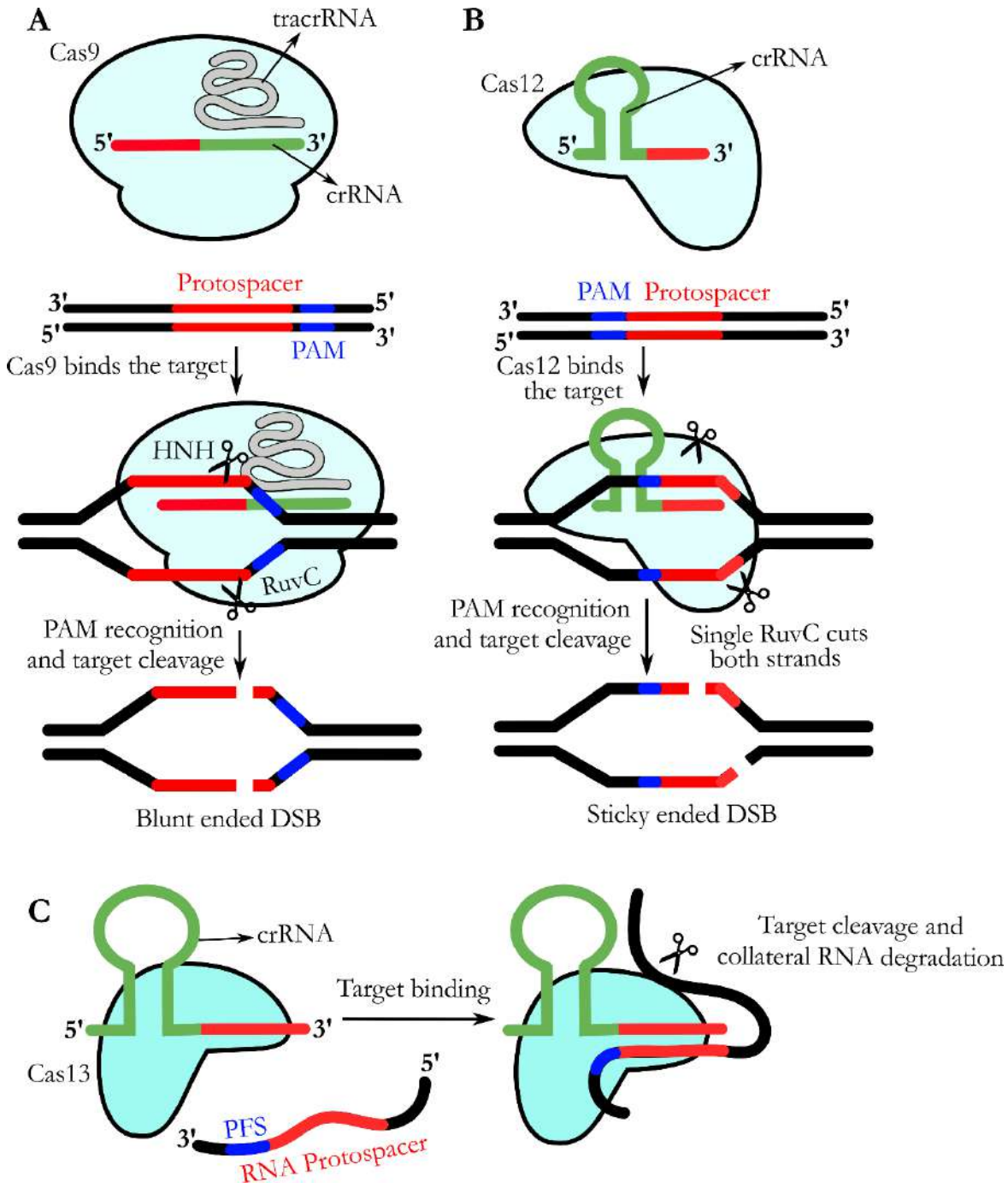


Figure 1.9: Mechanism of interference in class 2 CRISPR-Cas systems

(A) Type II effector complex comprises of Cas9 (Cyan), tracrRNA (Grey) and crRNA. During the targeting step, the effector complex identifies the protospacer (Red) on MGE via sequence complementarity with crRNA. Upon PAM (Blue) recognition and target binding, the HNH and RuvC nuclease domains of Cas9 nick the protospacer region of the target and non-target strands, respectively. Usually, the cleavage by Cas9 generates a blunt-ended DSB.

(B) Type V effector complex comprises of Cas12 (Cyan) and crRNA. Upon PAM and protospacer detection, Cas12 nicks both the target and non-target strands. Here, only a single RuvC nuclease domain is involved in the cleavage. Usually, Cas12 introduces DSB with a sticky cut.
(C) Cas13-crRNA complex acts as an effector nuclease in type VI system. Here, the interference occurs via RNA-guided RNA targeting. Upon recognising the PFS and protospacer region on RNA, the two HEPN domains on the surface of Cas13 gets activated. These HEPN domains catalyse the cleavage of target RNA and other RNA molecules indiscriminately.

1.5.6.1. Differentiation of self- versus nonself-targets during interference

The CRISPR array in the host contains a spacer sequence that is entirely complementary to the crRNA. Therefore, recognising the CRISPR DNA as a complementary target by the effector complex will give rise to the fatal autoimmune response in the host. Hence, interference by the CRISPR-Cas system is not solely dependent on the sequence complementarity between the target and crRNA. In most of the CRISPR-Cas systems, protospacers encompass a highly conserved PAM (or PFS in type VI) (Abudayyeh et al., 2016; Deveau et al., 2008; Mojica et al., 2009; Yosef et al., 2012; Zetsche et al., 2015). Usually, during spacer acquisition, full PAM sequence from the protospacer region is not integrated (Goren et al., 2012; Heler et al., 2015). Therefore, spacer sequences in the CRISPR array do not harbour any PAM. Relying on this difference, interference machinery seems to consider the crRNA matching region with a PAM in proximity as a legitimate feature of the nonself target site (Anders et al., 2014; Gleditzsch et al., 2019; Hayes et al., 2016; Westra et al., 2013).

Type III systems lack PAM sequences. Here, the effector complex follows a different set of molecular principles to avert autoimmunity. During maturation of crRNA, a remnant portion of repeat sequence remains on either end of the spacer sequence. The target protospacer on invader does not contain any adjacent sequences that resemble repeat. Therefore, the crRNA portion derived from the spacer region alone shows a complementarity at the target site on invaders. Whereas, at the spacer sites on CRISPR array, the complementarity is extended towards remnant repeat residues on mature crRNA. Such base pairing at 5'-end of crRNA deactivates the interference machinery and protects the host from an autoimmune response (Marraffini and Sontheimer, 2010).

1.5.7. Interference guided CRISPR adaptation enhances the immunity against mutated MGEs

In some CRISPR-Cas subtypes, target degradation by interfering nuclease generates shredded small nucleic acids from the invaders (Ex: Cas3 mediated digestion in type I systems) (Gong et al., 2014; Huo et al., 2014; Nimkar and Anand, 2020; Sinkunas et al., 2011). In *E. coli*, such remnant degraded products were observed to be uptaken and integrated as spacers by adaptation machinery (Künne et al., 2016). Such interference guided spacer uptake generates a feedback loop and results in effective immunity against recurring and mutated invaders (Datsenko et al., 2012; Staals et al., 2016). A natural fusion of Cas3 nuclease to the C-terminus of Cas2 adaption protein in type I-F infers the evolutionary advantage of interlinking interference and adaption stages (Fagerlund et al., 2017; Makarova et al., 2020b; Staals et al., 2016). In type II systems, target cleavage introduces only a DSB. In this regard, genome repair proteins such as RecBCD can shred the region further and generate a nucleic acid pool that could be potentially fuelled to the adaptation machinery (refer section 1.5.4.1) (Nussenzweig et al., 2019).

MGEs mutate their genomes often. These mutations help as a survival tactic against the assault of CRISPR-Cas interference machinery. Mismatches in PAM and seed regions on protospacers avert the target detection by effector complexes (Datsenko et al., 2012; Jackson et al., 2019). To counter such escape mutants, CRISPR-Cas system adopts a unique strategy termed ‘primed adaptation’. Type I systems are well known to display primed adaptation (Hille et al., 2018) and variants like type I-B acquire spacers by primed process alone (Li et al., 2014). During this step, though the interference is blocked, spacers were swiftly acquired from the bordering regions of the mutated protospacer site. In type I-E system, the spacers acquired by priming process mostly lie in the same strand direction as mismatched protospacer (Datsenko et al., 2012; Shmakov et al., 2014; Swarts et al., 2012), whereas, such bias is not observed in type I-B and I-F (Li et al., 2014; Richter et al., 2014). The primed acquisition requires all cascade components and Cas3 nuclease. Cascade conformational regulation by Cas8 upon interaction with legitimate or mismatched protospacer seems to dictate the occurrence of interference or primed adaptation, respectively (Xue et al., 2016). Preliminary

in vitro experiments suggest the association of cascade, adaptation complex and Cas3 initiate primed adaptation (Dillard et al., 2018; Redding et al., 2015). These models propose that the cascade can bind to mutated protospacer, but fails to recruit Cas3 for interference. Miscued interactions of cascade seem to direct the recruitment of Cas1-2 integrase at the R-loop, which is followed by Cas3 loading and translocation. The detailed mechanism of primed adaptation is still to be understood.

1.6. Definition of the problem

The adaptation stage spearheads the CRISPR-Cas immune response against lethal MGEs by acquiring a genetic memory of the infection. The spacer repository is inherited by the bacterial progeny, which in turn ensures their fitness against evolutionary pressures such as recurrent MGE attacks. During this process, a variety of spacers were acquired from foreign nucleic acids. CRISPR-Cas adaptation machinery of several organisms prefer prespacers that are bordered by a highly conserved PAM. Additionally, the length of the spacers incorporated into CRISPR locus surprisingly remains constant. In some bacteria, an exonuclease Cas4 assists the Cas1-2 integrase in PAM selection and also processes the prespacers for incorporation into the CRISPR array. However, Cas4 is non-ubiquitous and many organisms including *E. coli* (CRISPR subtype I-E) lack this nuclease. Here, an unknown sequence of events seems to dictate the PAM specificity and prespacer sizing.

The spacers located in the proximity of the leader region represent a recently acquired infection memory in response to the prevailing threat in the immediate neighbourhood. Preferential transcription of these spacers and their maturation into crRNAs are known to provide immediate response against the infections. Given this importance, the adaptation machinery incorporates new spacers derived from fresh phage invasions at the leader proximal repeat (amidst the presence of numerous other repeats). Despite being such a vital step in the CRISPR mediated immune response, the molecular events guiding the directionality of spacer insertion remained elusive. In *E. coli*, prespacer integration is site-specific *in vivo* that preferentially takes place at leader proximal repeat; however, purified adaptation complex integrates prespacers at all the repeats *in vitro*. This observation suggests the involvement of a mysterious host factor to confer specificity for prespacer integration.

When the author embarked his thesis work, these were the two major research questions – the mechanism of (i) prespacer capture and sizing and (ii) polarised prespacer integration – that remained elusive. In order to address these lacunae, the current work attempts to understand the mechanistic details of prespacer processing and polarised integration of prespacers in *E. coli* (type I-E).

1.7. Objectives of the study

- Identification of the factors that confer PAM selection and prespacer processing
- Development of a strategy to identify the host factors interacting with CRISPR locus
- Deciphering the role of host factors in the CRISPR adaptation process
- Unravelling the mechanism of directional prespacer integration

Chapter II

Deciphering the molecular principles of spacer selection in CRISPR adaptation

2. Chapter II

2.1. Introduction

The spacer acquisition stage constitutes the cornerstone of CRISPR-Cas adaptive response by expanding the compendium of infection memory (Barrangou et al., 2007; Deveau et al., 2008; Jackson et al., 2017; McGinn and Marraffini, 2019; Yosef et al., 2012). Despite the extreme variability in the nucleotide sequence, length of spacer remains conserved (approximately equivalent to that of repeat length) (Bolotin et al., 2005; McGinn and Marraffini, 2019; Mojica et al., 2005). In addition, majority of the organisms (CRISPR types I, II and V) display conservation of PAM at spacer origin site in MGE (Heler et al., 2015; McGinn and Marraffini, 2019; Yosef et al., 2012; Zetsche et al., 2015). CRISPR adaptation is primarily dictated by two highly conserved Cas1 and Cas2 (Jackson et al., 2017; Makarova et al., 2018; Yosef et al., 2012). *E. coli* encompasses 33 bp spacers and abut a 5'-AAG-3' PAM (where 'G' is destined to be the first residue of the spacer) (Datsenko et al., 2012; Mojica et al., 2009; Swarts et al., 2012; Wang et al., 2015; Yosef et al., 2012). The CRISPR adaptation machinery of *E. coli* (type I-E) derives its spacers predominantly from the DNA debris generated during the action of multi-subunit RecBCD DNA repair complex (discussed in section 1.5.4.1) (Levy et al., 2015). RecBCD generates ssDNA fragments ranging from tens to thousands of nucleotides in length (Dillingham and Kowalczykowski, 2008; Muskavitch and Linn, 1982a). But, structural studies have demonstrated that Cas1-2/I-E efficiently binds to partial duplex prespacers that are 33 bp in length (Nunez et al., 2015a; Wang et al., 2015). The existence of such spacer sized DNA fragments is infinitesimal among the RecBCD products. Moreover, a previous study also demonstrated the incorporation of 33 bp spacers that are directly acquired from electroporated longer DNA duplexes (63 bp) (Shipman et al., 2016). These findings augment the involvement of an additional DNA trimming step to generate befitting substrates for CRISPR adaptation.

Recent studies in various type I systems (I-A, I-B, I-C, I-D and I-G) highlighted the indispensable role of Cas4 nuclease in PAM selection and prespacer processing (Almendros et al., 2019; Kieper et al., 2018; Lee et al., 2019; Lee et al., 2018; Liu et al., 2017d; Rollie et al., 2018; Shiimori et al., 2018; Zhang et al., 2019b). The occurrence of Cas4 is prevalent in type I CRISPR-Cas systems except in subtypes I-E and I-F (Makarova et al., 2018). An

unorthodox type I-E Cas2 with extended DnaQ exonuclease at C-terminus in *Streptococcus thermophilus* DGCC7710 was shown to trim prespacers (Drabavicius et al., 2018). But *E. coli* (type I-E) do not harbour DnaQ fusion or Cas4 (Makarova et al., 2018) and the mechanism of prespacer processing is yet to be characterised. Though recent studies envisage the involvement of exonucleases during spacer acquisition in *E. coli* (Radovcic et al., 2018), the molecular events guiding PAM selectivity and prespacer processing remain obscure.

The inadequacy of lacking Cas4 or Cas2-DnaQ variant in *E. coli* does not appear to hinder PAM selection or spacer size preference (Mojica et al., 2009; Shipman et al., 2016; Swarts et al., 2012; Yosef et al., 2012; Yosef et al., 2013). Intrigued by these observations, we sought to understand how prespacers are selected and tailored to the appropriate size for CRISPR adaptation. In this chapter, we demonstrate that the PAM directed interactions with longer DNA fragments signals Cas1-2 to demarcate potential prespacer boundaries. Upon supplementing the reaction with exonucleases to mimic the cellular environment, we found that Cas1-2-DNA nucleoprotein complex could protect DNA fragments of ~33 bp length. These findings demystify the mechanism by which *E. coli* efficiently scales the fragments of the foreign DNA to generate prespacers of the desired length, in contrast to other CRISPR-Cas subtypes that possess dedicated prespacer processing nucleases such as Cas4.

2.2. Materials and Methods

2.2.1. Construction of plasmids

Lists of plasmids, strains and oligonucleotides used in this study are detailed in Appendix Table 1, Table 2 and Table 3, respectively.

Genes encoding Cas1 and Cas2 were amplified using *E. coli* K-12 MG1655 genomic DNA as a template. Expression vector p13SR-Cas1 was generated by inserting an amplicon encoding Cas1 at SspI site of plasmid p13SR. Expression vector pMS-Cas2 was created by introducing an amplicon encoding Cas2 between BamHI/HindIII sites of plasmid pMS (Guerrero et al., 2015), respectively. Bicistronic cassettes (*cas1-cas2*) expressing 6X Histidine tagged WT and 5M Cas1 were amplified using plasmid pCas1-2[K] (Diez-Villasenor et al., 2013) and plasmid pMut89 (Shipman et al., 2016) as a template, respectively. Amplified fragments encoding WT and 5M Cas1-2 were inserted between NcoI/NotI sites of pCas1-2[K] to generate plasmids pCas1-2H and p5M, respectively. PCR based mutagenesis was used to create plasmids pY22A and pΔC that express Y22A and ΔC variants of Cas1, respectively.

Gibson assembly protocol was utilised for generating all the recombinant vectors (Gibson et al., 2009) and the resultant constructs were verified by Sanger sequencing (Sanger et al., 1977).

2.2.2. Expression and purification of proteins

To express Cas1, *E. coli* BL21(DE3) harbouring p13SR-Cas1 was grown until 0.6 OD_{600} at 37 °C, 180 rpm in Auto-induction LB medium (Himedia) supplemented with 100 µg/ml Spectinomycin. After that, growth and induction were continued for 16 hrs more at 16 °C, 180 rpm. Subsequently, cells were harvested and resuspended in Buffer 1A (20 mM HEPES–NaOH pH 7.4, 500 mM KCl, 10 % Glycerol and 1 mM DTT) containing 1 mM PMSF and lysed by sonication. The clarified soluble cell extract was loaded on to 5ml StrepTrap HP column, which was then washed with Buffer 1A. Proteins were eluted with Buffer 1A containing 2.5 mM d-desthiobiotin. Eluted protein fractions were dialysed against Buffer 1B (20 mM HEPES–NaOH pH 7.4, 50 mM KCl, 10 % Glycerol and 1 mM DTT) and loaded onto

5 ml HiTrap Heparin HP column. Protein loaded columns were washed with Buffer 1B and bound proteins were eluted with a linear gradient of 0.05–2 M KCl in Buffer 1B. Purified fractions were pooled up and dialysed against Buffer 1A. The resulting sample was concentrated, snap-frozen and stored at −80 °C until required.

To purify Cas2, *E. coli* BL21(DE3) harbouring pMS-Cas2 was grown in Auto-induction LB medium supplemented with 100 µg/ml Kanamycin at 37 °C, 180 rpm. Upon reaching 0.6 $OD_{600,}$ temperature was shifted to 16 °C and the growth and induction were continued for 16 hrs at 180 rpm. Subsequently, cells were harvested and washed 2X times with Buffer 2A (20 mM HEPES–NaOH pH 7.4, 500 mM KCl and 10 % Glycerol). The bacterial pellet was resuspended in Buffer 2A containing 1 mM PMSF and the cells were lysed by sonication. Here, Cas2 encompasses 6X Histidine tagged MBP-SUMO as an N-terminal fusion and a Strep-II tag on the C-terminal end. The clarified fraction of the lysate was applied to a 5 ml MBPTrap HP column (GE Healthcare) and was followed by a washing step with Buffer 2A. After that, the bound proteins were eluted with Buffer 2A containing 10 mM Maltose. Eluted fractions were mixed with SUMO protease ($Ulp1_{403-621}$) (in 400:1 ratio of His-MBP-SUMO-Cas2-strep:$Ulp1_{403-621}$) (Guerrero et al., 2015) and the incubation was continued for 60 mins at 25 °C. Following this, the mixture was loaded onto 5 ml HiTrap IMAC HP column (GE Healthcare) 5X times to facilitate binding of Histidine tagged MBP-SUMO-Cas2-strep, MBP-SUMO and $Ulp1_{403-621}$. Column flow-through containing Cas2-strep was concentrated using a centrifugal membrane filter (Sartorius). To remove any trace protein contaminants, the concentrated sample was loaded onto HiLoad Superdex 200 pg gel filtration column (GE Healthcare), that is pre-equilibrated with Buffer 2B (20 mM HEPES–NaOH pH 7.4, 150 mM KCl and 10 % Glycerol). Eluted fractions containing Cas2-strep were pooled, concentrated, snap-frozen in liquid nitrogen and stored at −80 °C until required.

Integrase complex comprising of untagged Cas1 and C-terminal 6X Histidine tagged Cas2 was expressed and purified as described before (Moch et al., 2017) with minor modifications. Here, *E. coli* BL21(DE3) transformed using pCas1-2H was grown in 2XYT broth supplemented with 100 µg/ml Spectinomycin at 37 °C, 180 rpm till 0.6 OD_{600}. After that, the protein expression was induced by addition of 0.7 mM IPTG and the growth was continued at 25 °C for 24 hrs. Simultaneously, cells were harvested and washed 2X times with Buffer 2A (20 mM HEPES–NaOH pH 7.4, 150 mM KCl, 10 % Glycerol and 30 mM Imidazole). The pellet was resuspended in Buffer 2A containing 1 mM PMSF and cells were

lysed by sonication. After that, the lysate was clarified and loaded onto a 5 ml HiTrap IMAC HP column (GE Healthcare) and was followed by a washing step with Buffer 2A. A linear gradient of Imidazole (0.03–0.5 M) in Buffer 2A was applied to elute the proteins that were bound to the column resin. The purified fractions that contain the complex of Cas1-2 were pooled and concentrated using a centrifugal membrane filter (Sartorius). To remove trace protein contaminants and un-complexed Cas2, the concentrate was further purified using HiLoad Superdex 200 pg gel filtration column (GE Healthcare) that is pre-equilibrated with Buffer 2B (20 mM HEPES–NaOH pH 7.4, 150 mM KCl and 10 % Glycerol). Eluted fractions containing Cas1-2 integrase were pooled, concentrated, snap-frozen in liquid nitrogen and stored at −80°C until required. A similar procedure was implemented to purify 5M, ΔC and Y22A Cas1 variants of Cas1-2 from the IPTG induced *E. coli* BL21(DE3) cells that harbour p5M, pΔC and pY22A, respectively.

2.2.3. Electrophoretic mobility shift assays

To generate various prespacers (P33, P23[3'-5], P23[5'-5], P23[3'-10], P63, P63mPAM and their 5'-FAM labelled variants), respective oligonucleotides (Appendix Table 3) were mixed in a buffer containing 10 mM Tris-Cl pH 8.5. These mixtures were heated to 95°C and gradually allowed to cool to room temperature to facilitate the formation of duplex and partial-duplex prespacers.

The binding of Cas1-2 with various prespacers was monitored using electrophoretic mobility shift assays (EMSA). Here, 100 nM of desired 5'-FAM labelled prespacers (P23[3'-5], P23[5'-5], P33 and P23[3'-10]) were incubated with increasing concentration of WT Cas1-2 (0.1, 0.15, 0.2, 0.25, 0.45, 0.6, 0.8, 1, 1.5, 2 and 3 µM) in prespacer binding buffer (20 mM HEPES–NaOH pH 7.4, 125 mM KCl, 10 mM $MgCl_2$ and 1 mM DTT) for 30 mins at 37 °C. Subsequently, all the samples were directly loaded on 0.8 % agarose gel and electrophoresed in 1X TAE at 4 °C. Bound fraction for each sample in the gel was estimated by quantifying the amount of DNA at each band using densitometric analysis (Bound fraction of prespacer (%) at X µM Cas1-2 = [(Amount of DNA in the absence of Cas1-2 – Amount of unbound DNA at X µM Cas1-2) / (Amount of DNA in the absence of Cas1-2)] * 100. To estimate dissociation constants (K_D), the resulting plots of bound fraction (%) against Cas1-2

concentration was fitted to a non-linear equation: $y = B_{max} * x / (K_D + x)$ (where x, y, B_{max} and K_D represents Cas1-2 concentration (μM), bound fraction (%), maximum concentration of Cas1-2 bound to prespacer and dissociation constant, respectively). In EMSA that involves 5'-FAM labelled P63 or P63mPAM prespacers, 100 nM of DNA was incubated with 0.2 to 5 μM of WT / ΔC / Y22A or 0.2 to 20 μM of 5M Cas1-2 variants. All the binding experiments were independently repeated thrice.

To further verify the formation of Cas1-2 nucleoprotein complex, the release of prespacers was monitored upon Proteinase K treatment of Cas1-2 in each assay. To achieve this, an aliquot of the sample containing prespacer and 3 μM Cas1-2 was mixed with 1 mg/ml Proteinase K and incubated at 37 °C for 15 mins.

2.2.4. Exonuclease treatment of Cas1-2 bound DNA fragments

Exonuclease treatment was performed to identify the extent of protection conferred by binding of Cas1-2 on to a long DNA fragment. 40 μl of 0.5 μM P63 and 6 μM of either Cas1 or Cas2 or Cas1-2 in prespacer binding buffer were incubated at 37 °C for 45 mins. Subsequently, 20 μl aliquots of these samples were supplemented with 3 units of either T5 exonuclease (NEB) or Exonuclease III (NEB), or 3 units of the mixture containing both exonucleases and incubation was continued for 60 mins at 37 °C. After that, all the samples were mixed with an equal volume of denaturation buffer that contains 200 mM Tris-Cl pH 8.3, 200 mM Boric acid, 20 mM EDTA, 0.05 % SDS and 8 M Urea followed by heating at 95°C for 15 mins. These samples were loaded onto pre-heated 20 % denaturing polyacrylamide gels that were maintained at 50 °C and electrophoresed in 1X TBE. Subsequently, gels were stained with EtBr and visualised using a gel documentation system.

2.2.5. Exonuclease footprinting

T5 exonuclease mediated footprinting was performed to identify the interaction boundaries of Cas1-2 on longer prespacer DNA fragments. Here, 40 μl of 0.5 μM of desired fluorescein labelled P63 variant (P63T*, P63B*, P63mPAMT* and P63mPAMB*) was mixed

with 6 µM of WT or one of the mutant variants of Cas1-2 (Y22A, ΔC and 5M) in prespacer binding buffer and incubated for 45 mins at 37 °C. Subsequently, 20 µl aliquots of these samples were supplemented with 3 units of T5 exonuclease and incubation was continued for 60 mins at 37 °C. Thereafter, all the samples were mixed with an equal volume of denaturation buffer that contains 200 mM Tris-Cl pH 8.3, 200 mM Boric acid, 20 mM EDTA, 0.05 % SDS and 8 M Urea followed by heating at 95 °C for 15 mins. These samples were loaded onto pre-heated 20 % denaturing polyacrylamide gels that were maintained at 50 °C and electrophoresed in 1X TBE. Subsequently, gels were directly visualised using a gel documentation system.

2.2.6. Spacer acquisition assays

The *in vivo* spacer acquisition assays were performed as described previously (Yoganand et al., 2017; Yosef et al., 2012) with minor modifications. After transformation using plasmids (pCas1-2H, p5M, pY22A and pΔC), *E. coli* IYB5101 that expresses WT or a mutant of Cas1-2 was subjected to three cycles of growth and induction in LB medium supplemented with 100 µg/ml Spectinomycin, 0.2 % L-arabinose and 0.1 mM IPTG for 16 hrs at 37 °C. After each cycle, cultures were diluted to 1:300 times with fresh LB medium containing the aforementioned supplements and the growth was continued for 16 hours. After that, genomic DNA was isolated according to manufacturer's protocol (HiPurA bacterial genomic DNA purification kit, Himedia) and this was used as a template for PCR to monitor the spacer integration at CRISPR 2.1. All the PCR amplified samples were resolved on 1.5 % agarose gels to identify the DNA bands corresponding to parental and expanded arrays (parental array + n x 61 bp), where n is a positive integer. DNA quantities corresponding to parental and expanded array were quantified by densitometric analysis. Utilising these values, percentage of spacer integration for each Cas1-2 variant was estimated (% integration = [(Amount of expanded array) / (Amount of parental array + Amount of expanded array)] * 100.

2.2.7. High-throughput sequencing and analysis

To understand the effect of Cas1-2 mutants on prespacer scaling and PAM selectivity, high-throughput sequencing was performed to derive the sequences of newly incorporated prespacers. Expanded CRISPR arrays corresponding to the expression of each Cas1-2 variant were extracted from the agarose gels (QIAquick Gel Extraction Kit, Qiagen). Approximately 200 ng of each PCR product was further purified using HighPrep magnetic beads (MAGBIO). These purified samples were subjected to DNA end-repair and adaptor ligation using Illumina-compatible NEXTflex Rapid DNA sequencing kit (BIOO Scientific, Austin, Texas, U.S.A.). Subsequently, the ligated DNA products were purified with HighPrep magnetic beads and further enrichment was achieved by 8 cycles of PCR with Illumina-compatible primers (NEXTFlex DNA sequencing kit). These amplicons were subjected to an additional step of purification with HighPrep magnetic beads and were sequenced on a Miseq 300 paired-end platform.

The paired-end reads were subjected to several pre-processing steps as described below. Firstly, both F and R reads with a Phred score less than 20 were removed by utilising fastq_quality_trimmer from the FASTX-toolkit-version-0.0.13. The remaining F and R reads were trimmed in paired-end mode to remove F [5'-AGATCGGAAGAGCACACGTCTGA ACTCCAGTCA-3'] and R [5'-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3'] adapter sequences using Cutadapt-1.18 (Martin, 2011). Following these, the leader proximal spacer sequence (S0) was selectively retrieved in FASTA format. These S0 sequences that were derived from *E. coli* expressing WT and mutants of Cas1-2 were searched against plasmid (pCas1-2[K]) (Diez-Villasenor et al., 2013) and *E. coli* K-12 MG1655 genome (GenBank assembly accession: GCA_000005845.2), respectively, using BLASTN (Altschul et al., 1997). From the BLAST hits, we identified the location of spacer sequences on plasmid and *E. coli* K-12 MG1655 genome, respectively and extracted the triplet sequence corresponding to PAM. The conservation of PAM was analysed using WebLogo (Crooks et al., 2004). For all sequence manipulations including the extraction of S0 and PAM sequences, we employed custom-written python codes utilising the Biopython library (Cock et al., 2009).

2.2.8. Analysis of Cas1 C-terminal tail across CRISPR-Cas type I systems

Locus ID corresponding to Cas1 genes of type I-A, I-B, I-C and I-E were derived from the previous study (Makarova et al., 2015). Utilising these identifiers, we extracted Cas1 protein sequences of 36 type I-A, 150 type I-B, 129 type I-C and 116 type I-E organisms from NCBI protein database via Batch Entrez. After this, we performed multiple sequence alignments in the T-COFFEE web server (Notredame et al., 2000) for Cas1 proteins of each subtype separately. Utilising Cas1 crystal structures of *Archeoglobus fulgidus* (type I-A; PDB ID: 4N06), *Pyrococcus horikoshii* (type I-B; PDB ID: 4WJ0) and *E. coli* (type I-E; PDB ID: 5DQZ) as a reference, C-terminal tail residues of each Cas1 was extracted from their multiple sequence alignments. Owing to the absence of type I-C Cas1 crystal structure in PDB, Cas1 (BH0341) structure of *B. halodurans* was predicted by I-TASSER web server (Zhang, 2008) using structures corresponding to PDB ID: 3LFX, 4N06 and 2YZS as threading templates. I-TASSER predicted five different models for *B. halodurans* Cas1. Among these, the model with the highest confidence score (C-score = 1.25) was used as a structural reference for predicting the C-terminal tail residues from multiple sequence alignment of 129 type I-C Cas1 proteins.

Utilising ESpript server (Robert and Gouet, 2014), various secondary structural element positions were mapped on to the multiple sequence alignments of type I-A, I-B, I-C and I-E Cas1 proteins. All the protein structural representations were generated using ChimeraX (Goddard et al., 2018).

2.3. Results

2.3.1. Cas1-2 foothold protects the potential prespacer regions during exonuclease action

Spacers in *E. coli* are routinely derived from the remnant ssDNA fragments generated by the action of RecBCD mediated double-stranded break repair (Levy et al., 2015). Upon reannealing to their complementary sequences, variably sized DNA fragments that encompass blunt ends, 3'- or 5'- overhangs could be generated. Hence, we sought to understand if purified Cas1-2 (Figure 2.1A) can interact with such DNA fragments. To simulate these conditions *in vitro*, we prepared various types of DNA fragments such as P33 (33 bp duplex), P23[3'-5] (23 bp duplex with 5 nt 3'-overhangs), P23[5'-5] (23 bp duplex with 5 nt 5'-overhangs), P23[3'-10] (23 bp duplex with 10 nt 3'-overhangs), P63 (63 bp blunt duplex with PAM (5'-AAG-3')) and P63mPAM (63 bp blunt duplex without PAM) and monitored Cas1-2 binding by EMSA (Figure 2.1B).

In the EMSAs that employed 3'- and 5'- tailed duplex prespacers (Figure 2.1C-E), a slow migrating Cas1-2-prespacer complex was observed with increasing concentrations of Cas1-2. However, at higher Cas1-2 concentrations, a supershift of DNA in the wells presumably due to the accumulation of DNA-protein aggregates was seen (Figure 2.1C-E). In the binding assays that employ blunt prespacers P33, P63 and P63mPAM (Figure 2.1F-H), a reduction in prespacer band intensity was observed with increasing concentrations of Cas1-2. In line with previous studies (Radovcic et al., 2018), only DNA-protein aggregates were detected in the wells when blunt prespacers were incubated with Cas1-2 (Figure 2.1F-H). To further verify the presence of Cas1-2-prespacer complex, aliquots of the samples that contain the mixture of 100 nM prespacer and 3 µM Cas1-2 was treated with Proteinase K and the release of intact prespacer was detected (Lane PK in Figure 2.1C-H). These experiments revealed that the Cas1-2 interacts with different forms of DNA, albeit with varied affinities. Here, Cas1-2 displayed a stronger affinity to 3'-tailed substrates (P23[3'-5] (K_D = 648.5 ± 136.2 nM) and P23[3'-10] (K_D = 748.5 ± 163.7 nM)) in comparison to 5'-tailed (P23[5'-5] (K_D = 1.278 ± 0.25 µM)) or blunt (P33 (K_D = 2.885 ± 0.613 µM), P63 (K_D = 2.842 ± 0.372 µM) and P63mPAM (K_D = 6.478 ± 1.65 µM)) substrates (Figure 2.1C-H).

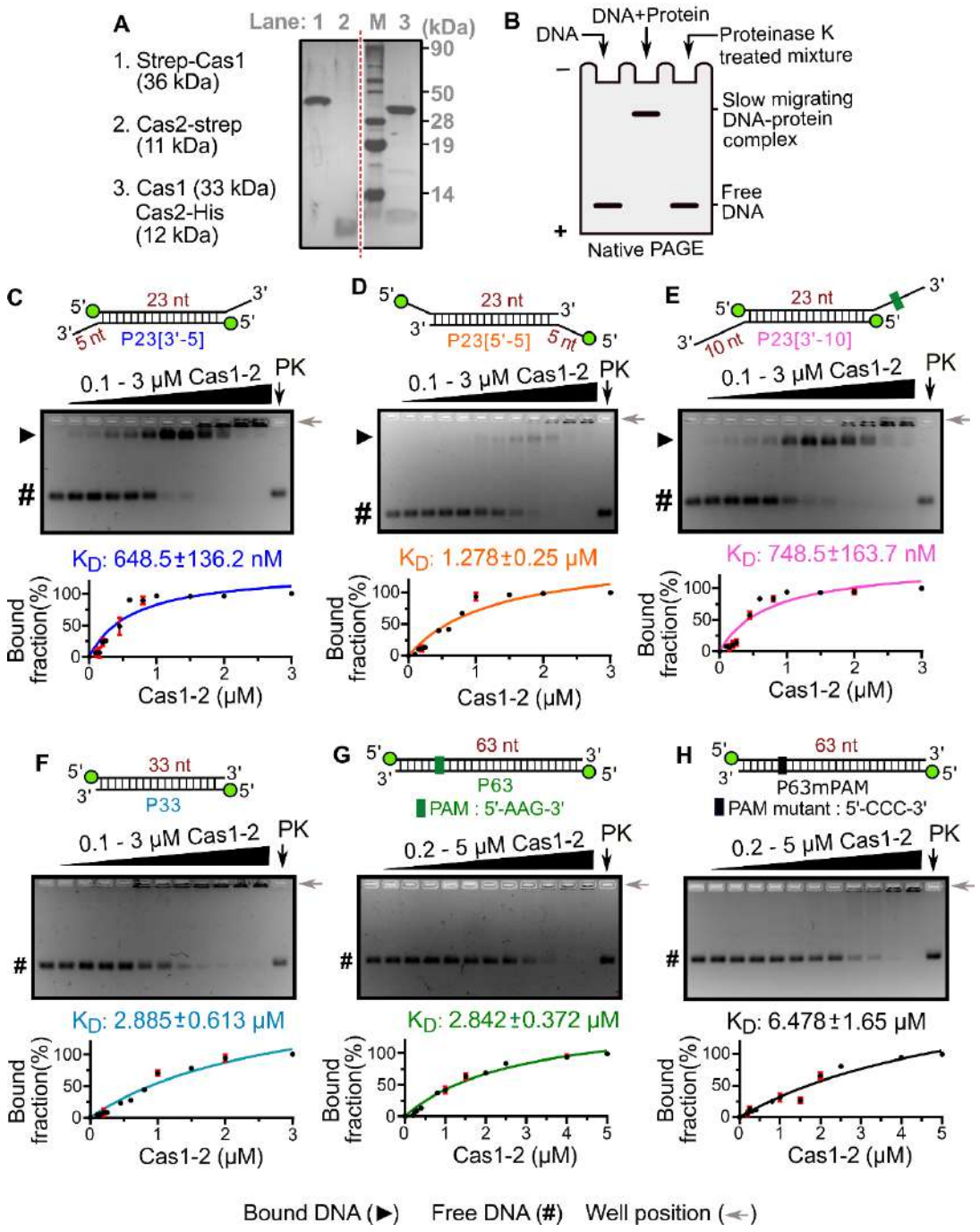


Figure 2.1: Cas1-2 interacts with prespacers of varied lengths

(A) Gel displaying the SDS-PAGE of purified Cas1 (Lane1), Cas2 (Lane2) and Cas1-2 (Lane 3) proteins. Protein marker (M) positions (in kDa) and the molecular weights corresponding to proteins in each lane are shown on the right.

(B) Picture representing the mechanism of EMSA. Generally, the interaction of nucleic acid binding proteins with DNA results in the formation of higher molecular weight nucleoprotein complexes. Such complex formation can be traced by resolving the reaction mixtures by native PAGE and comparing the mobility differences. To verify the nucleoprotein complex formation, the reaction mixtures can be supplemented with Proteinase K and the release of DNA upon protein digestion can be traced.

(C-H) Representative agarose gels depicting the interactions of Cas1-2 with 5'-FAM labelled prespacers P23[3'-5] (C), P23[5'-5] (D), P23[3'-10] (E), P33 (F), P63 (G) and P63mPAM (H). Schema of the prespacer highlighting the duplex and overhang length are displayed at the corresponding gels. Positions of bound prespacer (►), free prespacer (#) and wells (Grey arrow) are indicated in each gel picture. The lane that is loaded with Proteinase K treated Cas1-2 + prespacer mixture is indicated as PK. 100 nM of each prespacer DNA (P23[3'-5], P23[5'-5], P23[3'-10] and P33) was incubated with increasing concentrations of Cas1-2 (0, 0.1, 0.15, 0.2, 0.25, 0.45, 0.6, 0.8, 1, 1.5, 2 and 3 µM). In case of prespacers P63 and P63mPAM, 100 nM of DNA was incubated with 0, 0.2, 0.25, 0.3, 0.4, 0.8, 1, 1.5, 2, 2.5, 4 and 5 µM of Cas1-2. Plots of the bound fraction of prespacer (%) against Cas1-2 concentration (µM) and the estimated equilibrium disassociation constant values ($K_D \pm$ SD) from the binding experiments (in triplicates) are depicted at the bottom of the respective gels.

CRISPR adaptation in *E. coli* results in expansion of CRISPR array by incorporation of spacers that are precisely 33 bp in length (Swarts et al., 2012; Yosef et al., 2012). To achieve this, long DNA fragments generated during RecBCD activity have to be trimmed further by nuclease action. Of the multitude of proteins encoded by the Cas operon, only Cas1 and Cas2 contribute to naive spacer acquisition in *E. coli* (Yosef et al., 2012). Hence to understand how the integration compatible prespacers are generated, the longer DNA substrates P23[3'-10] and P63 (with an effective length of 43 nt and 63 nt, respectively) were incubated with increasing concentrations of Cas1-2 (Figure 2.2A). In these cases, even at the highest Cas1-2 concentration (75 µM), no DNA trimming was noted (Figure 2.2A).

The type I-E system is devoid of prespacer processing exonuclease Cas4 and its deficit cannot be complemented by Cas1-2 alone (Figure 2.2A). Hence, we predicted the involvement of cytoplasmic exonucleases in trimming the longer prespacer to a suitable length. Having established that Cas1-2 indeed binds DNA fragments of variable length (Figure 2.1C-H), we sought to test the fate of Cas1-2 bound DNA fragments against exonucleases. Here, Cas1-2-P63 nucleoprotein complex was treated with a mixture containing 5'→3' acting T5 exonuclease (T5exo) and 3'→5' acting Exonuclease III (ExoIII) (Figure 2.2B). To our

surprise, we identified a smear of protected DNA fragments (P63exo+) that ranged from 30-40 nt in the sample containing both P63 and Cas1-2 (Lane 8 in Figure 2.2B), whereas, such protection was not observed when we treated P63 in the absence of Cas1-2 or in the presence of either Cas1 or Cas2 (Lanes 5, 6 and 7 in Figure 2.2B). Coincidentally, the length of the protected fragments corresponded to legitimate spacer size in *E. coli* (~33 nt). Additionally, it was noted that this protection was absent when Cas1-2-P63 was treated with ExoIII alone (Lane 11 in Figure 2.2B). The nuclease generates 5'-overhangs with which Cas1-2 binds weakly (compare P23[3'-5] and P23[5'-5] in Figure 2.1C and D, respectively). Therefore, it appears that ExoIII seems to dislodge the weakly bound Cas1-2 from its position on the prespacer. In contrast, the treatment of Cas1-2-P63 with 5'→3' acting T5exo resulted in incompletely digested fragments that were predominantly of higher length (~45 nt) than that of P63exo+ (compare lanes 8 and 13 in Figure 2.2B). The protection of cognate spacer sized fragments (P63exo+) during T5exo+ExoIII digestion by Cas1-2 (Lane 8 in Figure 2.2B) indicates that the Cas1-2 mediated binding of large DNA fragments secure the boundaries of suitable prespacers from the exonucleolytic action of cellular nucleases in *E. coli* (Figure 2.2C).

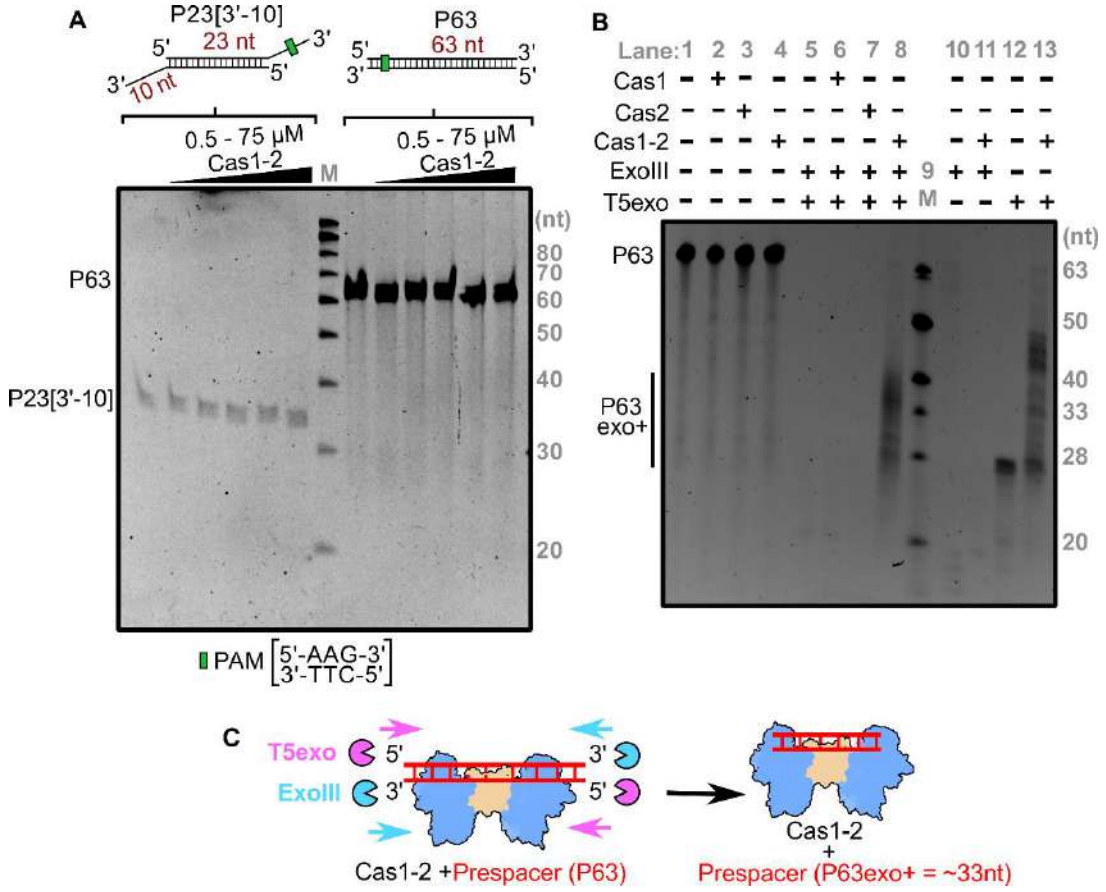


Figure 2.2: Tailoring of Cas1-2 bound longer DNA substrates by exonucleases generates spacer sized nucleic acid fragments

(A) Gel displaying denaturing PAGE of 0.5 µM of prespacer DNA (P23[3'-10] and P63) incubated with increasing concentrations of Cas1-2 (0, 0.5, 5, 10, 25 and 75 µM). Drawings of P23[3'-10] and P63 is displayed above their respective lanes. Positions corresponding to each of the prespacer DNA and oligo marker (M) are shown on the sides of the gel.

(B) Gel displaying denaturing PAGE of nuclease treated Cas1-2 bound P63 DNA. Presence (+) or absence (-) of each reaction component is labelled on top of each lane. Positions corresponding to the substrate (P63) and T5exo/ExoIII digested DNA fragments (P63exo+) are indicated on the left. Oligo marker (M) positions are shown on the right.

(C) Schema illustrating the mechanism of Cas1-2 mediated protection of prespacer boundaries is displayed. Cas1-2 (in Blue and Brown), T5exo (Magenta pie), ExoIII (Cyan pie) and prespacer P63 (Red ladder) are portrayed.

2.3.2. PAM directed binding of Cas1-2 defines the boundary for prespacers

Since Cas1-2 binding was shown to mark the spacer boundaries (Figure 2.2), we sought to identify and map these regions. As no protected DNA fragments appeared upon treatment of Cas1-2-P63 complex with ExoIII alone (Lane 11 in Figure 2.2B), we utilised T5exo digestion (Lane 13 in Figure 2.2B) to demarcate the prespacer boundaries defined by Cas1-2. To accomplish this, we employed variants of 63 bp blunt-ended prespacers that encompass fluorescein labelled 3'-end (6-FAM) either on the top (P63T*) or on the bottom (P63B*) strand (Figure 2.3A). Further, to identify the footprints of Cas1-2 on these 3'-end labelled prespacers, we incubated Cas1-2 bound prespacer complex with T5exo. Here, the Cas1-2 binding on the prespacer acts as a roadblock that stalls the 5'→3' progression of T5exo. The length of the resultant labelled fragments specifies the stalling points of the exonuclease, which in turn, indicates the binding position of Cas1-2 on the prespacer. Utilising this approach, we mapped the cleavage termination points on the top and bottom strands of the prespacer. After T5exo treatment of P63T*, we could observe ~28 nt labelled fragment. This is indicative of an inherent nuclease stalling point (in the absence of roadblocks such as Cas1-2) around the 28th nt position from the labelled end in P63T* (Lane 2 in Figure 2.3B; Lane 12 in Figure 2.2B). However, owing to complete exonucleolytic cleavage of the bottom strand, we did not observe such stalling on P63B* (Lane 10 in Figure 2.3B). Upon T5exo treatment of Cas1-2 bound P63T* complex, we noticed a shift in the nuclease stalling point to ~45 nt position from the labelled end (Lane 4 in Figure 2.3B). This maps the Cas1-2 binding position to be around 45 nt from the labelled end (P63T* in Figure 2.3A). Coincidentally, this binding position of Cas1-2 on P63T* is localised around a cognate PAM sequence (5'-AAG-3' ranging from 47 nt to 49 nt upstream of labelled position) (P63T* in Figure 2.3A).

Prompted by this finding, we were interested in identifying the extent of protection that Cas1-2 could confer on the bottom strand upon its binding at the PAM region. To accomplish this, we treated the Cas1-2 bound P63B* complex with T5exo. Here, the resulting length of the protected fragments upon exonuclease treatment indicated that Cas1-2 complex interaction could guard a region spanning ~45 nt from the labelled end in P63B* (Lane 12 in Figure 2.3B and P63B* in Figure 2.3A). As the PAM residues are positioned at 14 nt from the labelled end of the P63B* (Figure 2.3A), the effective length of the protected prespacer

from the PAM is ~30 nt. Overall, these results suggest a PAM dependent mechanism by which Cas1-2 could selectively acquire 33 nt prespacers.

To test the role of PAM in defining the protospacer boundary, we employed mutated P63 DNA fragments (P63mPAMT* and P63mPAMB* in Figure 2.3A) that are devoid of any cognate *E. coli* PAM sequence (5'-AWG-3', where W=A/T). These labelled fragments were incubated with Cas1-2 and later treated with T5exo. Here, we found extended smears and multiple bands upon employing P63mPAMT* (Figure 2.3A and Lane 8 in Figure 2.3B) and P63mPAMB* (Figure 2.3A and Lane 16 in Figure 2.3B). The varied length of resultant labelled fragments is indicative of numerous stalling points on these P63 mutants (Figure 2.3A) that occurred due to Cas1-2 binding. These results highlight that Cas1-2 gets specifically recruited towards PAM containing region that plays a crucial role in defining prespacer boundaries. Whereas, such specificity is lost when the DNA fragments lack PAM. Here the promiscuous interaction of Cas1-2 results in the generation of illicit prespacers that defy the productive length of the prespacers for integration (Lanes 8 and 16 in Figure 2.3B; P63mPAMT* and P63mPAMB* in Figure 2.3A).

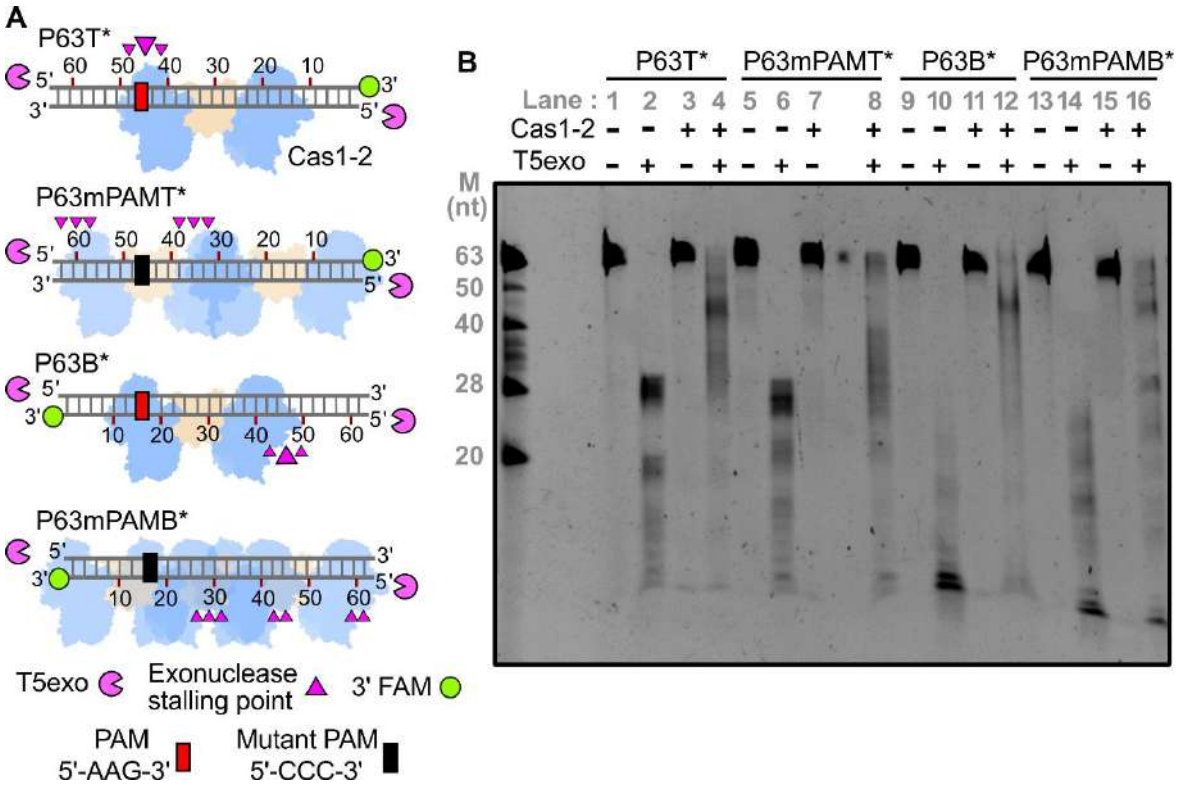


Figure 2.3: Cas1-2 complex is predominantly localised around PAM region

(A) Schematic representation of various fluorescein labelled prespacer substrates (P63T*, P63mPAMT*, P63B* and P63mPAMB* as Grey ladder) used in the assay is shown. Labelled 3'-end of each prespacer is highlighted as Green circles, whereas, PAM and its mutated variant are depicted as Red and Black boxes, respectively. Numbering on the DNA represents the distance (in nt) of particular position from the labelled end. T5exo (Magenta pie) is positioned at susceptible 5'-ends of DNA substrate. Positions of T5exo stalling points (Magenta triangles) and binding sites of Cas1-2 (Blue and Brown blobs) that are estimated from nuclease footprinting assay performed in (B) are displayed.

(B) Gel displaying the denaturing PAGE of T5exo treated Cas1-2 bound fluorescein labelled P63 variants (P63T*, P63mPAMT*, P63B* and P63mPAMB*). Presence (+) or absence (-) of Cas1-2 and T5exo is labelled on top of each lane. Positions corresponding to the DNA fragments of oligo marker (M) are shown on the left.

2.3.3. Intrinsic specificity of Cas1-2 circumvents the requirement of Cas4 during PAM selection in *E. coli*

Having found the role of PAM mediated interaction of Cas1-2 in selecting the prespacers for uptake, we attempted to understand the intrinsic molecular principles that confer precision to Cas1-2 in PAM selectivity and prespacer scaling. Previous structural studies of CRISPR adaptation complex in *E. coli* suggested that the extended Cas1 C-terminal tail of apoCas1-2 complex (Nuñez et al., 2014) gets organised around the PAM residues upon binding to prespacer DNA (compare C-terminal tail in Figure 2.4A and B) (Wang et al., 2015). In particular, Q287 and I291 residues of this proline-rich C-terminal tail make direct contacts with the nucleotides of PAM (Figure 2.4B), thus possibly imparting the PAM specificity. In another striking feature, a pair of Y22 residues that are derived from two different Cas1 protomers scales a 23 bp duplex region of prespacer via stacking interactions at either end (Figure 2.4B) (Nunez et al., 2015a; Wang et al., 2015). This gating mechanism at the Cas1-2 platform seems to scale the spacer length by facilitating the positioning of 3'-overhang at the catalytic groove for integration (Figure 2.4B). To validate these observations, we generated Cas1-2 variants that encompass either deletion of Cas1 C-terminal tail (ΔC - ΔP279-S305) or a Cas1 point mutant Y22A (Figure 2.4C). As a control, we also used a Cas1 variant (5M - Q24H, P202Q, G241D, E276D and L297Q) (Figure 2.4B) that was previously shown to abrogate PAM selectivity (Shipman et al., 2016).

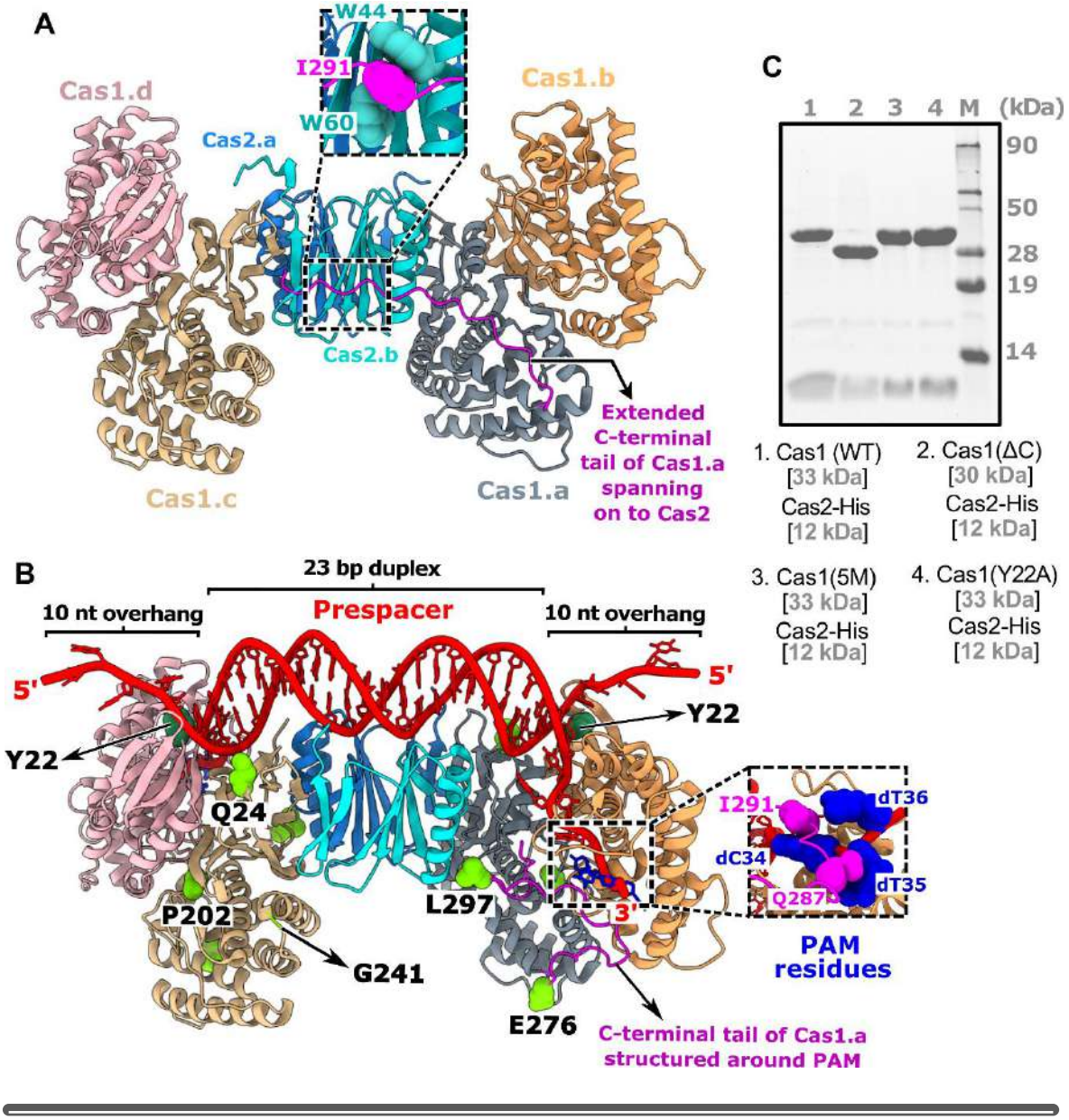
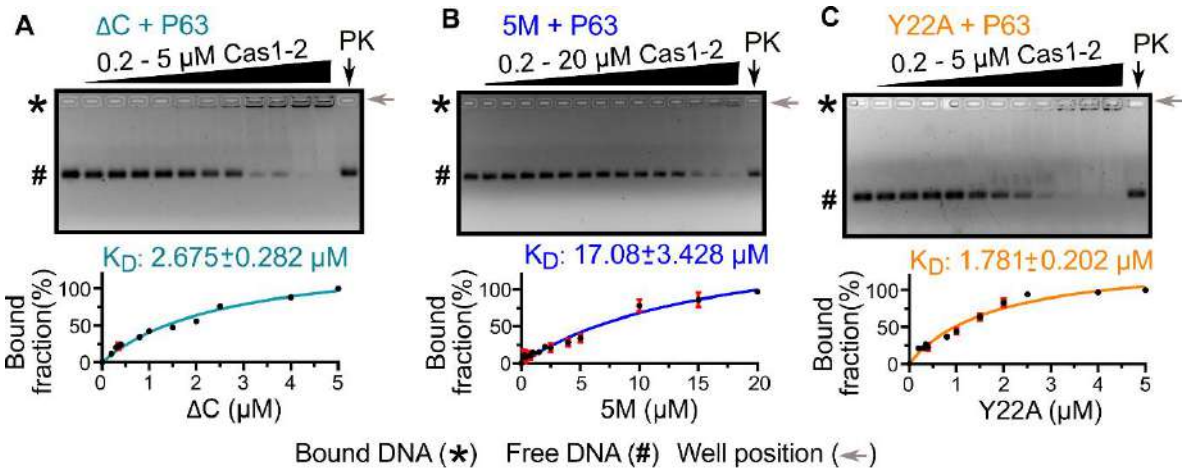


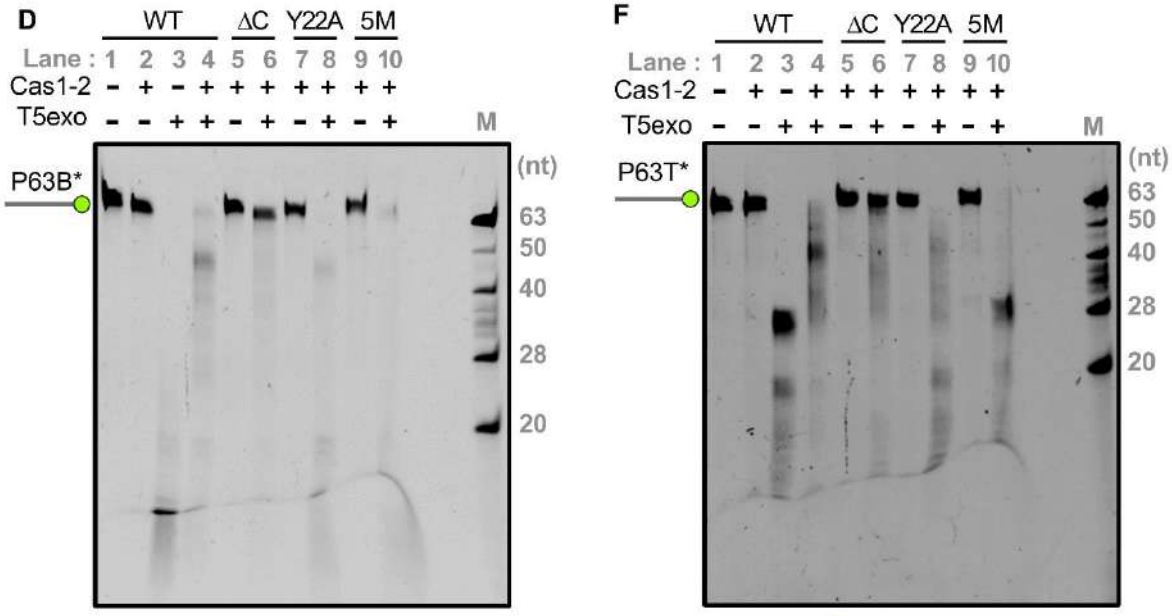
Figure 2.4: Structural features of Cas1-2 that determine the prespacer selection

(A-B) Structural comparison of apo-Cas1-2 integrase (PDB ID: 4P6I) (A) and Cas1-2-prespacer complex (PDB: 5DQZ) (B). Four protomers of Cas1 (indicated as Cas1.a-d) and two protomers of Cas2 (indicated as Cas2.a-b) are shown in different colours. The stacking of Cas2 tryptophan residues (W44 and W60 in Cyan) with I291 of Cas1.a C-terminal tail (G275-S305 in Magenta) in apo-Cas1-2 (close-up view in (A)) and the interactions of Cas1.a I291 and Q287 residues with PAM residues (in Navy blue) Cytosine 34 (dC34) and Thymidine 35 (dT35) in Cas1-2-prespacer complex (close-up view in (B)) are shown. The Cas1 Y22 residues (in Dark green spheres) that stack the nucleic acid bases at the prespacer duplex ends are denoted (B). Amino acid residues corresponding to Cas1 mutations (Q24H, P202Q, G241D, E276D and L297Q) in 5M variant are displayed (Green spheres) as part of the Cas1.a and Cas1.c protomers. For clarity, E276 and L297 of Cas1.a and Q24, P202 and G241 of Cas1.c are labelled at their respective positions.

(C) Gel displaying SDS-PAGE of purified WT, ΔC, 5M and Y22A is shown. Molecular weight (kDa) corresponding to proteins in each lane are shown on the bottom and protein molecular weight marker (M) positions are shown on the right.

Next, we employed purified Cas1-2 variants (Figure 2.4C) in footprinting assays and analysed their binding positions on P63B* and P63T* (Figure 2.5D-G). As like in the case of wild type Cas1-2 (WT), T5exo treated Y22A-P63B* and Y22A-P63T* generated a nuclease stalling point at ~45 nt albeit with reduced efficiency (compare lanes 4 and 8 in Figure 2.5D and F, respectively). These findings indicate that Cas1 Y22A does not result in the altered PAM specificity (WT and Y22A in Figure 2.5E and G). Despite having a stronger affinity for P63 (Y22A K_D = 1.781 ± 0.202 μM *versus* WT K_D = 2.842 ± 0.372 μM) (Figure 2.5C and Figure 2.1G), the absence of Y22 mediated stacking interaction seems to reduce the prespacer protection ability of Y22A against nuclease action. Additionally, a shift in the nuclease stalling point to 60 nt from the labelled ends was observed for ΔC and 5M variants (Lanes 6 and 10 in Figure 2.5D and F). Due to its low affinity, 5M seems to display reduced protection of prespacers from nuclease action (ΔC K_D = 2.675 ± 0.282 μM *versus* 5M K_D = 17.08 ± 3.428 μM) (Figure 2.5A and B). The shift in this nuclease stalling point indeed indicates that the ΔC and 5M variants displayed an impaired PAM specificity and were randomly interacting at the ends of P63B* and P63T* (Figure 2.5E and G, respectively).





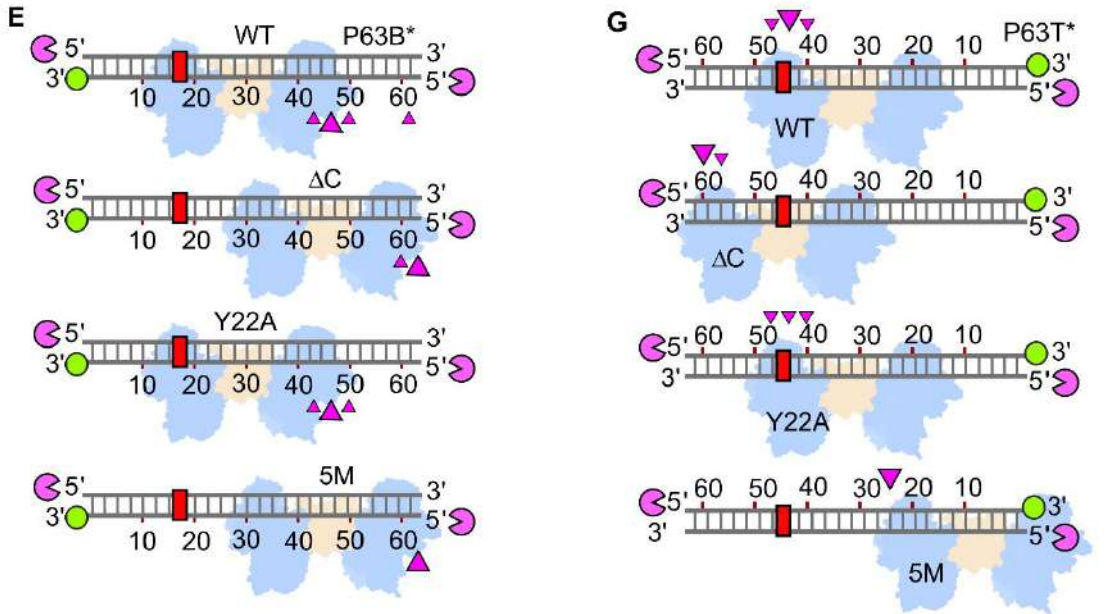


Figure 2.5: Cas1-2 variants display differing specificities towards prespacers

(A-C) Representative agarose gels depicting the interactions of ΔC (A), 5M (B) and Y22A (C) with 5'-FAM labelled prespacer P63. The variant of Cas1-2 employed and the positions of bound prespacer (*), free prespacer (#) and wells (Grey arrow) are represented in each gel image. 100 nM of P63 was incubated with increasing concentrations of ΔC (A) or Y22A (C) (0.2, 0.3, 0.35, 0.4, 0.8, 1, 1.5, 2, 2.5, 4 and 5 μM) or 5M (B) (0.2, 0.3, 0.35, 0.4, 0.8, 1, 1.5, 2, 2.5, 4, 5, 10, 15 and 20 μM). The lane that is loaded with Proteinase K treated Cas1-2 mutant + prespacer mixture is indicated as PK. Plots of the bound fraction of prespacer (%) against Cas1-2 concentration (μM) and the estimated equilibrium disassociation constant values ($K_D \pm$ SD) from the binding experiments (in triplicates) are depicted at the bottom of the respective gels.

(D & F) Gels depicting the denaturing PAGE of T5exo treated Cas1-2 (WT (lanes 1-4) or ΔC (lanes 5-6) or Y22A (lanes 7-8) or 5M (lanes 9-10)) bound fluorescein labelled P63B* (D) and P63T* (F). Presence (+) or absence (-) of each reaction component is indicated on top of each lane. Positions of labelled DNA fragments P63B* (D) and P63T* (F) are shown on the left. Oligo marker (M) positions are indicated on the right.

(E & G) Schematic illustrations of the footprinting assays performed in (D) and (F). DNA substrate P63B* and P63T* (Grey ladder in (E) and (G), respectively), positions of 3' fluorescein label (Green circle) and PAM region (Red rectangle) are represented. Numbering on the DNA represents the distance (in nt) of particular position from the labelled end. T5exo (Magenta pie) is positioned at susceptible 5'-ends of DNA substrate. Positions of T5exo stalling points (Magenta triangles) and binding sites of each variant of Cas1-2 (WT or ΔC or Y22A or 5M in Blue and Brown blobs) that are estimated from nuclease footprinting assays performed in (D) and (F) are displayed.

To fortify our observations on Cas1-2 directed PAM selection, we sought to understand the impact of these Cas1-2 mutations on the PAM composition of spacers acquired *in vivo*. To probe this quest a robust spacer acquisition assay system is a prerequisite. Hence, we utilised an established strategy for assessing the spacer acquisition in *E. coli* IYB5101 (Yosef et al., 2012) (Figure 2.6). Upon expression of Cas1-2, new spacers of 33 bp in length (S0) are integrated at the leader-repeat junction in *E. coli*. This homing step is facilitated by concomitant duplication of leader proximal repeat (R0) (Datsenko et al., 2012; Goren et al., 2012; Yosef et al., 2012). These events during CRISPR adaptation expands the length of the CRISPR array by 61 bp (Lane 2 in Figure 2.6C). In the spacer acquisition assay, an episomal copy of IPTG inducible (T7-LacO expression system) Cas1-2 bicistron was employed. The expansion of CRISPR locus during the spacer integration was probed using PCR (Figure 2.6).

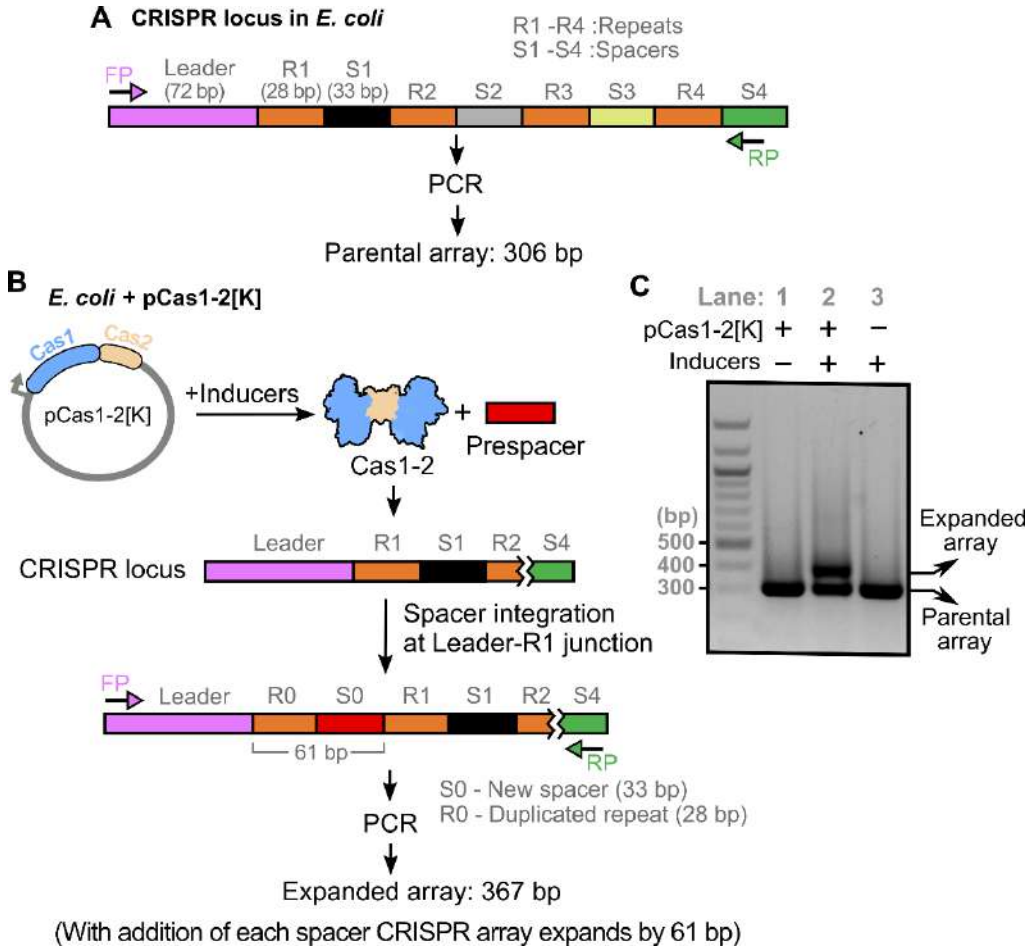


Figure 2.6: Spacer acquisition assay detects the incorporation of new spacers into CRISPR locus.

(A) Diagram of *E. coli* CRISPR locus is depicted. CRISPR DNA elements, *viz.*, leader (Purple), repeats (R1-R4 in Orange) and spacers (S1-S4 in Black, Grey, Yellow and Green, respectively) are labelled. The annealing positions of forward and reverse primers (FP and RP, respectively) used in PCR amplification of 306 bp Parental CRISPR array is shown.

(B) Outline of spacer acquisition assay performed in *E. coli* is presented. Upon induction, Cas1-2 captures and integrates the prespacer (Red) at the target site (Leader-R1 junction) in CRISPR locus. Due to spacer homing (S0) and simultaneous repeat duplication (R0), the CRISPR array expands by 61 bp. This increment in length was traced by PCR to monitor new spacer integration.

(C) Agarose gel displaying the result of spacer acquisition assay performed in *E. coli*. Absence (-) or presence (+) of pCas1-2[K] and inducers is indicated on top of each lane. Positions of parental and expanded arrays are denoted on the right and the positions corresponding to DNA marker are shown on the left.

Next, we employed plasmids encoding Cas1-2 variants (WT, ΔC, 5M and Y22A) and monitored *in vivo* spacer uptake. We observed that the mutations in ΔC and 5M have partly reduced the spacer incorporation efficacy of Cas1-2 (compare Lanes 2, 4 and 6 in Figure 2.7A). Surprisingly, Y22A displayed a drastic reduction of spacer uptake *in vivo* (compare Lanes 2 and 8 in Figure 2.7A). Expanded CRISPR arrays corresponding to the expression of each mutant were purified, and the sequences of newly incorporated spacers were derived from high-throughput sequencing. In line with previous studies, we observed that the spacers originated from both genome and plasmid (Shipman et al., 2016). Irrespective of the mutations in Cas1-2, the length of the incorporated spacers is strictly conserved (i.e., 33 nt) (Figure 2.7B). This finding suggests that Y22 mediated stacking interaction with prespacer or the Cas1 C-terminal restructuring is dispensable for the scaling of prespacers. These spacer sequences were mapped onto the plasmid and genome to identify the PAM. Despite the display of precise prespacer scaling by Cas1-2 variants, the specificity towards PAM region appears to be profoundly altered (Figure 2.7C). In concurrence with previous studies (Shipman et al., 2016; Yosef et al., 2013), we observed that most of the spacers acquired by WT Cas1-2 encompass a conserved PAM region (5'-AAG-3', where 'G' indicates +1 position of 33 nt spacer) (Figure 2.7C). In line with the nuclease protection assay (Figure 2.5D-G), we did not observe any preference towards the PAM region when we employed 5M or ΔC (Figure 2.7C). This finding bolsters the involvement of Cas1 C-terminal tail in PAM selectivity. Despite the reduced efficiency in spacer acquisition *in vivo* (Figure 2.7A), surprisingly, Y22A has displayed a remarkable precision for PAM selectivity, suggesting that this mutation bestowed high fidelity with respect to PAM recognition (Figure 2.7C).

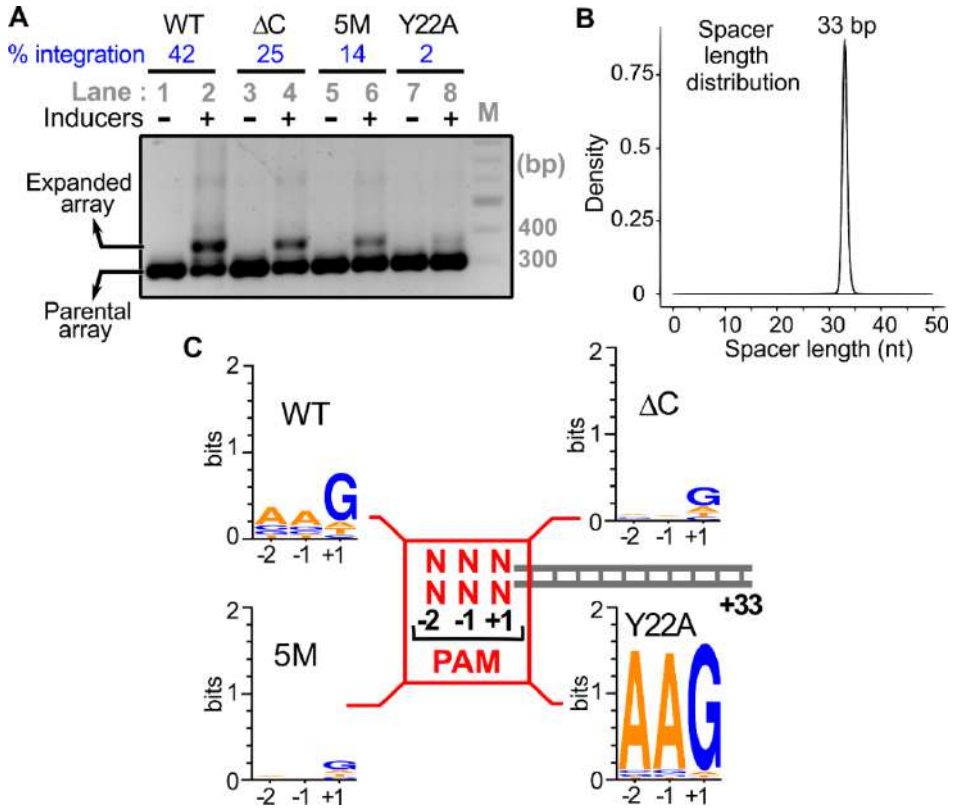


Figure 2.7: Intrinsic specificity of Cas1-2 integrase directs the uniformity in spacer length and PAM preference during CRISPR adaptation

(A) Agarose gel depicting the PCR products from spacer acquisition assay performed in *E. coli* harbouring the plasmids that express the Cas1-2 variants (WT (lanes 1–2), ΔC (lanes 3–4), 5M (lanes 5–6) and Y22A (lanes 7-8)). Absence (–) or presence (+) of inducers is indicated on top of each lane. Positions corresponding to parental and expanded arrays (CRISPR 2.1 array) are indicated on the left. The percentage of integration is displayed on top of the respective lanes (indicated in Blue). DNA marker (M) positions are represented on the right.

(B) Overlay of plots depicting length distribution of the newly acquired spacers that are incorporated into CRISPR 2.1 array by Cas1-2 variants (WT, ΔC, 5M and Y22A) during *in vivo* integration assay (A) is shown. X-axis depicts length of the spacer (nt) whereas normalised frequency (density) is indicated on the Y-axis.

(C) Illustration depicting the PAM preference of Cas1-2 variants (WT, ΔC, 5M and Y22A) during *in vivo* integration assay (A). +1 to +33 sequence of each spacer (Grey ladder) was extracted from high-throughput sequencing data. Subsequently, sequence information of -1 and -2 positions of each spacer was derived from the respective plasmid/genome sequence. The conservation profile of PAM sequences (-2, -1 and +1 in red) corresponding to the respective Cas1-2 variant is shown as a sequence logo.

2.4. Discussion

CRISPR system in *E. coli* (type I-E) displays precise scaling of prespacer length and stringent selectivity of PAM (Shipman et al., 2016; Swarts et al., 2012; Yosef et al., 2013). Here, we attempted to uncover the elusive molecular events that drive the generation of competent substrates for homing at the leader-repeat junction by CRISPR adaptation complex. During naïve adaptation, prespacers are predominantly foraged from the DNA fragments generated by RecBCD during DNA repair (Levy et al., 2015) or by Cas3 during primed adaptation (Datsenko et al., 2012; Künne et al., 2016; Semenova et al., 2016). Being helicase-nuclease enzymes, RecBCD and Cas3 action generates DNA fragments of varied length (Dillingham and Kowalczykowski, 2008; Sinkunas et al., 2011). Though these fragmented DNA acts as a source for prespacers, the mechanism by which the prespacers are sized remained elusive.

A previous study has shown that an *in vitro* reconstituted Cas1-2 can solely process the prespacers with 10 nt 3'-overhangs to legitimate spacer size (Wang et al., 2015). In contrast, our experiments show the absence of such prespacer processing even with increasing Cas1-2 concentrations, despite the presence of PAM (P23[3'-10] in Figure 2.2A). We posit that these variations could be due to differences in the purification strategies of Cas1-2 complex in both studies. Here, we employed Cas1-2 that was generated as a complex *in vivo* and was further purified extensively, whereas, the previous study (Wang et al., 2015) utilized *in vitro* reconstituted Cas1-2 that potentially contained unassembled, free Cas1 protomers. The utilization of high concentrations of the *in vitro* reconstituted Cas1-2 and thus the increased presence of free Cas1, a known endonuclease (Babu et al., 2011), could have resulted in the cleavage of the prespacer overhangs (Wang et al., 2015). Alternatively, because the nuclease activity was only seen at µM concentration of *in vitro* reconstituted Cas1-2 (Wang et al., 2015), despite the fact that the integrase activity requires just nM concentration (Nunez et al., 2015b; Yoganand et al., 2017), it points to the possibility of trace cellular nuclease contamination.

Previous structural studies of *E. coli* Cas integrase complex reveals that 33 bp length of DNA can be exactly accommodated in between two active site regions of Cas1-2 (Nunez et al., 2015a; Wang et al., 2015). This hints at the fact that the Cas1-2 foothold can only mask 33 bp region and the rest is exposed to potential nuclease action. Mimicking such conditions,

we incubated Cas1-2 bound longer DNA fragments (P63) with 3'→5' (ExoIII) and 5'→3' (T5exo) acting exonucleases. These reactions resulted in the generation of fragments that are in the range of cognate *E. coli* spacer size (Figure 2.2B). Likewise, in *B. halodurans* (type I-C) and *S. thermophilus* DGCC7710 (type I-E), the nuclease action of Cas4 and DnaQ (an auxiliary domain of Cas2), respectively, on Cas1-2 bound DNA, generate integration competent prespacers (Drabavicius et al., 2018; Lee et al., 2018). This implies that Cas1-2 can solely catalyse spacer integration, but the generation of productive prespacers involves the action of an additional nuclease. The lack of a known prespacer processing enzyme such as Cas4 in *E. coli* led us to hypothesise that the trimming action could be complemented by other cellular nucleases. Unlike in other CRISPR variants (Almendros et al., 2019; Kieper et al., 2018; Lee et al., 2018; Liu et al., 2017d; Rollie et al., 2018; Shiimori et al., 2018; Zhang et al., 2019b), the productive pruning of prespacers by non-Cas nucleases is not sequence-specific (Figure 2.2). Moreover, a parallel study on prespacer generation in *E. coli* also independently demonstrated that the generic nucleases (such as DNA polymerase III or Exonuclease T) are sufficient for trimming the prespacers upon PAM recognition by Cas1-2 (Kim et al., 2019b). These explain why the involvement of specific nucleases such as Cas4 is precluded for prespacer processing in *E. coli* (see below).

In addition to spacer length conservation, most prokaryotes display selective uptake of phage-origin prespacers bordered by a PAM (McGinn and Marraffini, 2019). Recent *in vivo* studies in various type I organisms (I-A, I-B, I-C, I-D and I-G) have underscored the indispensible requirement of Cas4 in PAM selection as well as in prespacer processing (Almendros et al., 2019; Kieper et al., 2018; Liu et al., 2017d; Shiimori et al., 2018; Zhang et al., 2019b). Moreover, *in vitro* studies performed with adaptation complex of *B. halodurans* (type I-C) and *S. solfataricus* (type I-A) revealed that Cas4 nuclease avoids the processing of free DNA ends that are devoid of PAM sequence (Lee et al., 2018; Rollie et al., 2018). This preferential activity of Cas4 seems to act as a critical checkpoint in ensuring the productive uptake of infection memory by Cas1-2 in the hosts.

A previous study in *E. coli* showed that upon expression of Cas1-2, 33 nt prespacer bordered by PAM originated from the longer electroporated DNA (P63) (Shipman et al., 2016). Interestingly, we observed the protection of the same region by Cas1-2 when we performed a nuclease footprinting assay on P63 (Figure 2.3). These experiments highlight that the Cas1-2 complex alone is sufficient to recognise PAM in *E. coli*. The footprinting

experiments also demonstrate the binding of substrates at multiple points when PAM residues of P63 were mutated (Figure 2.3). These non-specific interactions of Cas1-2 could generate a heterogenous population of protected prespacers. This explains how the adaptation complex of various type I organisms infrequently uptake prespacers with erroneous PAM (WT in Figure 2.7C) (Datsenko et al., 2012; Jackson et al., 2019; Li et al., 2017; Musharova et al., 2018; Rao et al., 2017; Savitskaya et al., 2013; Shmakov et al., 2014).

Structural analysis of Cas1-2-prespacer complex highlights the features that could lead to precise scaling and PAM selection of prespacers. A platform formed by the interaction of a Cas2 dimer with two Cas1 dimers on either side houses the 23 bp duplex region of prespacer (Figure 2.4) (Nunez et al., 2015a; Wang et al., 2015). Stationed at either end of this duplex is the aromatic ring of Y22 residue that stacks the prespacer at the border of the Cas1 catalytic groove and directs the 3'-overhang to position it's 5^{th} nt at the catalytic site (Figure 2.4). Thus, Y22 guided meticulous placement of DNA substrate seems to dictate the length of prespacer (Nunez et al., 2015a; Wang et al., 2015). Furthermore, the flexible C-terminal tail of Cas1 is moulded around the PAM region (Figure 2.4B). The absence of such molecular architecture upon mutating the PAM hints at the role of C-terminal tail in PAM recognition (Wang et al., 2015). Deployment of Cas1-2 variants that encompass either deletion of Cas1 C-terminal tail (ΔC) or Y22A in spacer integration assays helped to unveil the role of these structural entities in determining the PAM selection (Figure 2.7). As shown here, the deletion of C-terminal tail resulted in impaired PAM recognition (ΔC in Figure 2.5D-G) and led to an uptake of prespacers that were lacking PAM (ΔC in Figure 2.7C).

A comparison of the structures of Cas1 from various type I organisms (Figure 2.8) revealed a striking contrast between Cas1 C-terminal tail of type I-E and other subtypes. The C-terminal tail of *E. coli* Cas1 is noticeably longer with 31 amino acid (aa) than the shorter 12 aa tails of *A. fulgidus*, *P. horikoshii* and *B. halodurans* (Figure 2.8A-D) (Kim et al., 2013; Wang et al., 2015; Zhang, 2008). Additional comparative analysis reinforced these observations that type I-E encompasses the most extended C-terminal tail with an average length of 29 aa (Figure 2.8E and Appendix Figure 1-4). Coincidentally, CRISPR-Cas subtypes with shorter Cas1 C-terminal tails such as type I-A, I-B and I-C encompass Cas4 (Makarova et al., 2018) and previous studies suggest the indispensability of Cas4 in promoting PAM specificity (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018; Shiimori et al., 2018). In

contrast to these systems, it appears that the extended C-terminal tail of Cas1 in *E. coli* (type I-E) compensates for the lack of Cas4 by guiding the PAM selection.

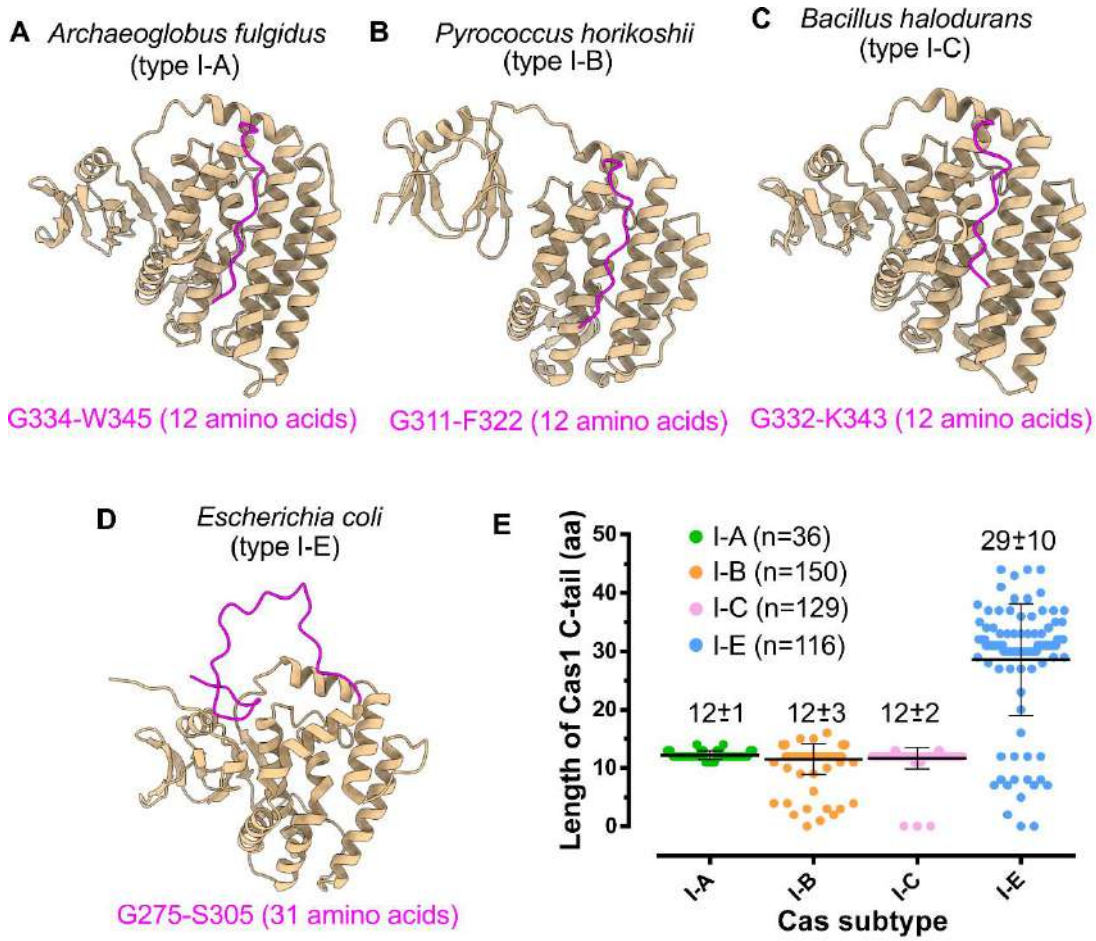


Figure 2.8: Cas1/I-E harbours extended C-terminal tail

(A-D) Structures highlighting Cas1 C-terminal tail (in Magenta) of *A. fulgidus* (type I-A; PDB ID: 4N06) (A), *P. horikoshii* (type I-B; PDB ID: 4WJ0) (B), *B. halodurans* (type I-C; predicted model) and *E. coli* (type I-E; PDB ID: 5DQZ) (D). The amino acids corresponding to the start and end position of C-terminal tail are displayed at the bottom of the respective structures.

(E) Scatter plot representing the length differences among Cas1 of type I-A, I-B, I-C and I-E is displayed. Each Cas subtype is shown in a different colour and the average length of amino acids in C-terminal tail (Mean ± SD) of each subtype is indicated at the respective position. 'n' corresponds to the number of Cas1 sequences from each subtype that are considered for the analysis (*vide* Section 2.2.8).

In line with the previous observations (Nunez et al., 2015a), a rampant decrease in spacer integration efficiency was observed in Y22A variant (Figure 2.7A). In addition to this, Y22A also conferred reduced prespacer protection against the nucleases in comparison to WT (Y22A in Figure 2.5D and F). These observations highlight the critical role of Y22 residue in providing the WT with a better grip on bound DNA (WT in Figure 2.5D-G). As Y22A could lack such interactions with its substrates, nucleases might seamlessly dislodge it from the bound prespacer (Y22A in Figure 2.5D and F). This action appears to limit the substrate availability and impede spacer integration *in vivo* (Figure 2.7A). Despite the reduction in spacer acquisition potential, Y22A showed high fidelity towards "AAG" PAM selection than WT (Figure 2.7C). As discussed above, Y22A displays a reduced grip on the prespacers during nuclease mediated processing. This weak prespacer binding could be further disrupted in the absence of PAM due to the loss of interactions with Cas1 C-terminal tail. Therefore, only the presence of cognate PAM (AAG) is likely to allow Y22A to retain the hold on DNA during prespacer processing leading to selective enrichment (Figure 2.7C). The strategic position of Cas1 Y22 in adaptation complex appears to have a key role in defining the prespacer length (Figure 2.4B). Interestingly, our experiments with Y22A resulted in the uptake of prespacers that were predominantly 33 bp in length (Figure 2.7B). These findings negate the involvement of Y22 stacking interactions in deciding the prespacer boundary. Recent studies in type V-C demonstrated that a mini integrase complex constituted by Cas1 tetramer prefers short (18 bp) spacers (Wright et al., 2019), likewise in *E. coli*, Cas1-2 structural framework alone appears to be a critical parameter in gauging the length of spacers (Nunez et al., 2015a; Wang et al., 2015).

Our work in conjunction with the previous studies allows us to propose an updated model for prespacer capture in type I systems (Figure 2.9). During CRISPR adaptation, the dispensability of sequence-specific auxiliary nucleases such as Cas4 seems to be contingent on the type of molecular players that are involved in PAM selection. Though Cas1-2 integrase catalyses the prespacer homing, in the majority of type I systems, PAM selection and prespacer processing require Cas4 (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018; Shiimori et al., 2018). In contrast, in type I-E system, the intrinsic affinity of Cas1-2 integrase alone is sufficient to recognise cognate PAM. This lineage-specific remarkable adaptation of Cas1-2/I-E offsets the requirement of PAM specifying Cas4 nuclease. Instead, generic cellular non-Cas nucleases are co-opted to trim the exposed DNA ends of Cas1-2-prespacer complex for generating the legitimate prespacers for integration.

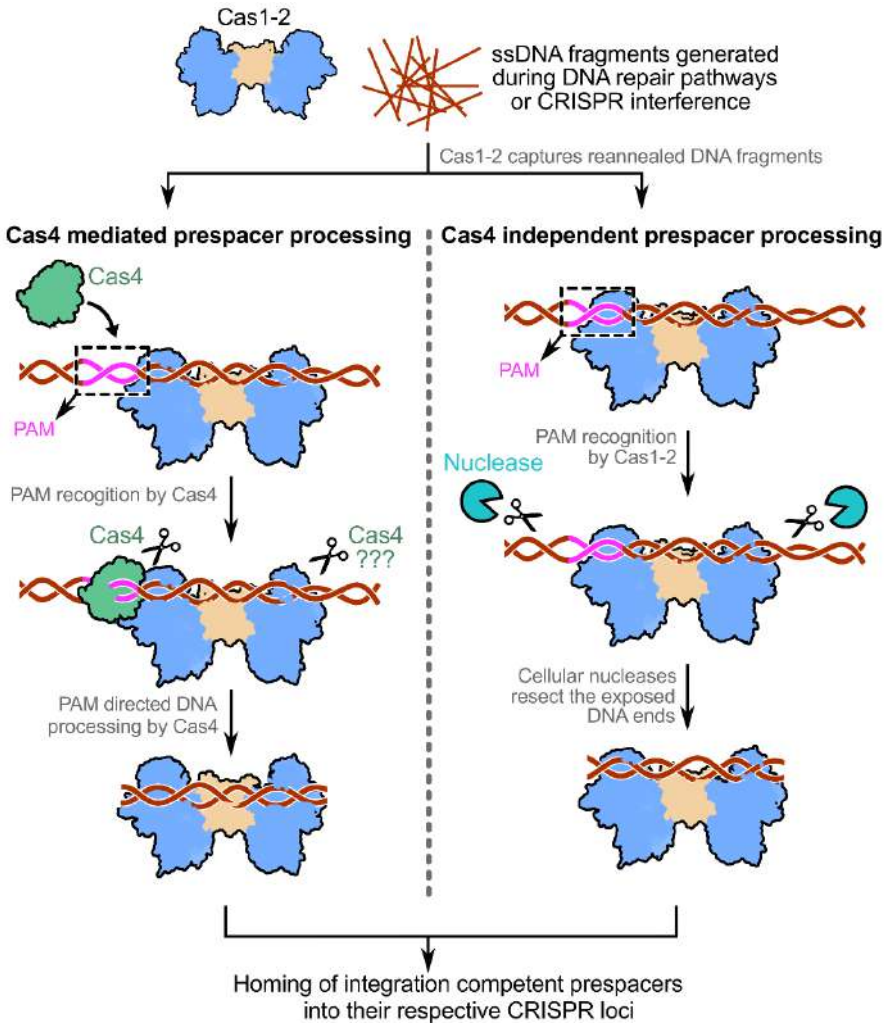


Figure 2.9: Model depicting the mechanism of Cas4 dependent and independent prespacer processing in type I CRISPR-Cas systems

The prespacer production in CRISPR systems encompasses a subset of two events: (i) PAM directed prespacer capture and (ii) processing of selected prespacer to a defined length for integration at CRISPR locus. Generally, Cas1-2 captures long dsDNA fragments that are produced by reannealing of ssDNA products (in Brown) derived from DNA repair pathways (such as RecBCD in *E. coli*) or CRISPR interference. Though in type I-E it is clear that the DNA capture by Cas1-2 precedes the prespacer processing event, the order of DNA capture and processing is yet to be understood in other CRISPR-Cas subtypes. The Cas4 (in Green) in CRISPR-Cas subtypes I-A, I-B, I-C and I-D or the Cas4 domain of Cas4-Cas1 fusion in type I-G trims the DNA upon recognising the PAM region (in Magenta) (Almendros et al., 2019; Kieper et al., 2018; Lee et al., 2019; Lee et al., 2018; Liu et al., 2017d; Rollie et al., 2018; Shiimori et al., 2018; Zhang et al., 2019b). The second copy of Cas4 in type I-B was shown to trim the non-PAM end upon recognising a short motif (Shiimori et al., 2018), whereas, in other subtypes, it is not clear whether Cas4 processes this end. Here, the dual role of PAM recognition and prespacer processing by Cas4 propels CRISPR adaptation. Unlike this, in type I-E system, the intrinsic affinity of Cas1-2 integrase towards PAM itself is sufficient to define the potential prespacer regions. This aspect of Cas1-2/I-E precludes

the involvement of any PAM specific Cas nucleases (such as Cas4) in prespacer selection. As the Cas1-2/I-E protects the prespacer boundaries efficiently upon recognising the PAM, any common cellular non-Cas nucleases (in Cyan) could trim the exposed ends to generate aptly sized prespacers for integration into the CRISPR locus.

Upon generation of suitable prespacers, CRISPR-Cas adaptation machinery catalyses their integration into CRISPR locus. This event is polarised and always occur at leader adjoining repeat. Though Cas1-2 integrase complex is known to champion the spacer homing, the factors that direct site-specificity remains uncomprehended. Hence, in the upcoming chapters, we attempted to characterise the steps involved in directional homing of spacers into CRISPR locus.

2.5. Summary

In the current chapter, utilising EMSA and nuclease protection assays, we identified that the Cas1-2 interacts with larger DNA fragments and protects the prespacer boundaries. Through DNA footprinting assays, we determined that the PAM sequence directs the Cas1-2-prespacer interaction. By analysing the available Cas1-2-prespacer complex structures, we predicted that the interaction of Cas1 C-terminal tail with the prespacer could confer PAM specificity to Cas1-2. We performed footprinting and spacer integration assays utilising a Cas1-2 variant that is devoid of Cas1 C-terminal tail (ΔC). In line with our assumptions, we found that ΔC did not interact at PAM regions and also acquired the spacers that are devoid of PAM. A comprehensive comparison of Cas1 proteins from various type I organisms indicated that the Cas1 C-terminal tail of type I-E candidates is longer in comparison to representatives of other types (I-A, I-B and I-C). Strikingly, Cas4, the prespacer processing nuclease that confers the PAM specificity in type I-A, I-B and I-C is also absent in *E. coli* (type I-E). Our observations in the current chapter indicated that the lack of Cas4 in *E. coli* is compensated by Cas1 C-terminal tail mediated PAM identification and subsequent trimming of exposed ends in Cas1-2-prespacer complex by non-Cas cellular exonucleases.

Chapter III

Identification of an accessory host factor for CRISPR adaptation

3. Chapter III

3.1. Introduction

The peculiar architecture of repeat-spacer array is the hallmark of the CRISPR-Cas system. Various conserved motifs in the leader and repeat regions of CRISPR locus mandate the productive immune response against MGEs (Arslan et al., 2014; Goren et al., 2016; Hille et al., 2018; Nunez et al., 2016; Yoganand et al., 2017; Yosef et al., 2012). Despite the presence of several repeat-spacer units in the CRISPR array, the site of integration of prespacer has always been at the leader-repeat junction resulting in the integration of the prespacer and concomitant duplication of the first repeat (Diez-Villasenor et al., 2013; Swarts et al., 2012; Yosef et al., 2012). In general, the integrase complex constituted by Cas1 and Cas2 directs the spacer uptake (Nunez et al., 2015a; Wright and Doudna, 2016; Xiao et al., 2017b). In *E. coli*, among multitude of type I-E Cas proteins (Makarova et al., 2018), Cas1-2 complex alone is sufficient for directional spacer integration *in vivo* (Yosef et al., 2012). Intriguingly, despite having intrinsic sequence specificity towards short sequence motifs at leader-repeat junction (Rollie et al., 2015), the Cas1-2 was shown to integrate prespacers at all CRISPR repeats *in vitro* (Nunez et al., 2015b). Such contrasting differences in the choice of integration sites by Cas1-2 during *in vitro* and *in vivo* spacer acquisition had hinted us the possible involvement of accessory factors to bring in specificity towards prespacer integration at leader-repeat junction.

To unravel the mechanism that dictates the fidelity of prespacer homing in *E. coli*, we directed our efforts in identifying the host proteins that interact at the CRISPR locus. Here, we utilised the dCas9 based immunoprecipitation principle (Fujita and Fujii, 2013) and designed an in-house CRISPR/dcas9 based molecular tool to discern the factors that bind at the CRISPR region. Upon this, we identified that a nucleoid-associated protein called Integration Host Factor (IHF) as an essential accessory factor in spacer acquisition (Yoganand et al., 2017) (independently the research team led by Jennifer Doudna also identified the essentiality of IHF in CRISPR adaptation (Nunez et al., 2016)).

3.2. Materials and Methods

3.2.1. Construction of bacterial strains and plasmids

Descriptions of the strains, plasmids and oligonucleotides are listed in Appendix Table 1, Table 2 and Table 3, respectively. *E. coli* IYB5101 (referred to as WT) (Yosef et al., 2012) was used as parental strain for all the genomic manipulations unless specified otherwise. Knock-out strains of *ihfα* (ΔIHFα) and *ihfβ* (ΔIHFβ) were created using λ Red recombineering (Datsenko and Wanner, 2000). Keio collection strains (Baba et al., 2006) carrying deletions of *ihfα* and *ihfβ* were used as templates for amplification of Kanamycin resistant cassettes along with 100 - 130 bp flanking sequence. Amplified cassettes were used to transform λ Red recombinase expressing *E. coli* IYB5101/pKD46 to create ΔIHFα and ΔIHFβ strains.

Plasmid pdCas9-bacteria (Qi et al., 2013) was modified with the construct encoding 3XFLAG-dCas9-StrepII (dCas9: nuclease null variant of *Streptococcus pyogenes* Cas9 (D10A, H840A)). Overlap extension PCR was used to generate a 166 bp DNA fragment encoding a sgRNA complementary to a region that is 86 bp upstream of first CRISPR repeat in *E. coli* BL21-AI (NCBI accession: NC_012947.1, nucleotide positions: 1002800-1003800). This region was inserted in between SpeI and HindIII sites of the plasmid pgRNA-bacteria (Qi et al., 2013) to create the plasmid pgRNA-leader.

E. coli K-12 MG1655 genomic DNA was used as a template to amplify genes encoding IHFα and IHFβ. To generate expression vector p8R-IHFαβ, a bicistronic cassette encoding IHFα and IHFβ was amplified and inserted at the SspI site of plasmid p8R.

Gibson assembly protocol was utilised for generating all the recombinant vectors (Gibson et al., 2009). After this, the resultant constructs were verified by Sanger sequencing (Sanger et al., 1977).

3.2.2. dCas9 mediated immunoprecipitation

E. coli BL21-AI was transformed with p3XF-dCas9, pgRNA-leader and pCas1-2[K] (Diez-Villasenor et al., 2013) and was allowed to grow in a shaker operated at 180 rpm till

OD_{600}=0.6 at 37 °C in LB media supplemented with 0.2 % L-arabinose, 0.1 mM IPTG, 25 µg/ml Chloramphenicol, 100 µg/ml Ampicillin and 50 µg/ml Spectinomycin. 100 ng/ml Anhydrotetracycline was added to induce the expression of 3X FLAG-tagged dCas9 and growth was continued for four more hours to allow dCas9-gRNA complex to anchor on its target site i.e., the upstream region of CRISPR leader. Chemical cross-linking and cell lysis were performed as described previously (Waldminghaus and Skarstad, 2010) with few modifications. Formaldehyde was added to a final concentration of 1 % to cross-link proximally interacting nucleic acids and proteins. Cross-linking was continued for 20 mins at 25 °C with gentle rocking. Glycine was added to a final concentration of 0.5 M and incubation was continued for 5 mins at 25 °C to quench the cross-linking reaction. 10 ml cells were centrifuged at 2500 g at 4 °C for 5 mins and the pellet was washed twice with an equal volume of buffer W (20 mM Tris-Cl pH 7.5 and 150 mM NaCl). Pelleted cells were resuspended in 1ml buffer L (10 mM Tris-Cl pH 8.0, 20 % Sucrose, 50 mM NaCl, 10 mM EDTA, 10 mg/ml Lysozyme) and incubated at 37 °C for 30 mins. Lysate was resuspended in 4 ml of buffer R (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 % Triton-X 100, 0.1 % Sodium deoxycholate, 1 mM PMSF and 0.1 % SDS). The cells were subjected to sonication for 4 rounds of 15 X 1 second pulses with 2 mins pause between each round in Vibra-cell probe sonicator that was set at 33 % amplitude. Clarified supernatant containing sheared DNA-protein complex was separated by centrifugation. 800 µl of supernatant was mixed with 200 µl of Dynabeads Protein G (Life technologies) conjugated with 20 µg of Anti-FLAG M2 antibody (Sigma) and rocked gently at 4 °C overnight. Incubated beads were separated by centrifugation and washed twice each with 1 ml of Low Salt Wash Buffer (20 mM Tris-Cl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1 % TritonX-100, 0.1 % SDS), High Salt Wash Buffer (20 mM Tris-Cl pH 8.0, 500 mM NaCl, 2 mM EDTA, 1 % TritonX-100, 0.1 % SDS), LiCl Wash Buffer (10 mM Tris-Cl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5 % Nonidet P-40 (NP-40), 0.5 % Sodium deoxycholate) and TBS Buffer (50 mM Tris, pH 7.5, 150 mM NaCl) with 0.1 % NP-40 as described previously (Fujita and Fujii, 2013). In the final step, beads were separated by centrifugation and resuspended in 100 µl of buffer containing 20 mM Tris-Cl pH 8 and 150 mM NaCl. 30 µl of resuspended beads were mixed with 10 µl of 4X SDS sample buffer and heated at 95 °C for 30 mins to reverse cross-link and denature the proteins. The heated mixture was loaded on to SDS-PAGE and electrophoresed to enter stacking gel. The part of stacking gel containing the proteins was sliced and analysed by mass spectrometry for the identification of protein factors in the sample.

3.2.3. Spacer acquisition assays

In vivo acquisition assays were performed as described earlier (Yosef et al., 2012). Briefly, three cycles of growth and induction was performed with *E. coli* IYB5101 (WT) or its variants (ΔIHFα and ΔIHFβ) carrying pCas1-2[K] (Diez-Villasenor et al., 2013) at 37 °C for 16 hours in LB media supplemented with 50 µg/ml Spectinomycin, 0.2 % L-arabinose and 0.1 mM IPTG. In between each cycle, cultures were diluted to 1:300 times with fresh LB media containing the aforementioned supplements and growth was continued for 16 hours. For IHF complementation experiments, ΔIHFα and ΔIHFβ strains were transformed with p8R-IHFαβ and pCas1-2[K] and 3 cycles of inductions were performed as discussed above. To monitor CRISPR array expansion, 200 µl of induced cells were collected after cycle 3 and washed thrice and resuspended in distilled water. These cells were used as a template for PCR to monitor CRISPR array expansion in CRISPR 2.1 array. Primers for the PCR were designed to anneal at 72 bp from the first repeat in upstream and fourth spacer sequence in the downstream. All the PCR amplified samples were separated on 1.5 % agarose gels to identify parental and expanded arrays (parental array + 61 bp).

3.3. Results

3.3.1. CRISPR/dCas9 based immunoprecipitation detects the participation of IHF as an accessory factor for adaptation *in vivo*

To identify the potential host factors that are likely to promote the directional insertion of prespacer fragment, CRISPR/Cas9 based immunoprecipitation was employed (Fujita and Fujii, 2013). Here, Cas1-2 complex was expressed along with the inactive form of FLAG-tagged Cas9 and the sgRNA that is targeted towards the leader region of the CRISPR array in *E. coli* BL21-AI (NCBI accession: NC_012947.1, nucleotide positions: 1002800-1003800). After the chemical cross-linking, the DNA bound protein factors that are localised into the leader region were selectively pulled down using the anti-FLAG coated beads against the FLAG-tagged dCas9 (Figure 3.1A). Upon analysing the pull-down fractions in SDS-PAGE, a protein band corresponding to the molecular weight of the dCas9 was noted (~160 kDa band in Lane 3 of Figure 3.1B).

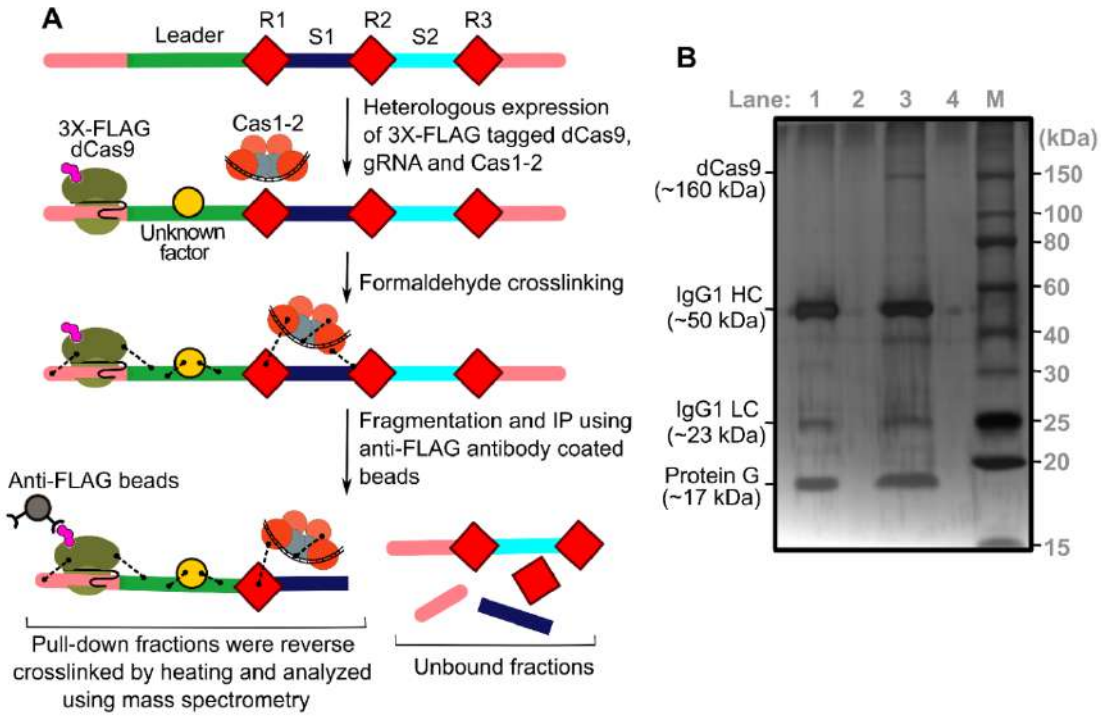


Figure 3.1: dCas9 based immunoprecipitation for identification of CRISPR associated host factors

(A) Schema of dCas9 directed immunoprecipitation assay to capture the CRISPR associated host factors during spacer acquisition step is displayed. CRISPR DNA encompassing leader (Green), repeats (R1-R3 in Red), spacers (S1-S2 in Blue and Cyan, respectively) and upstream and downstream regions (in pink) is shown. Heterologously expressed Cas1-2 complex (Orange and Grey) bound to the prespacer DNA (in Black) is displayed. 3X FLAG-tagged dCas9-gRNA complex (Green ovals with Magenta tag) is shown bound to the upstream region of the leader. During the active spacer acquisition process, Cas proteins and the unknown host factors (Yellow) are cross-linked (Black dotted lines) to the proximal DNA region by the addition of Formaldehyde. After the isolation and fragmentation of this cross-linked protein bound DNA, the CRISPR regions were selectively pulled down using Anti-FLAG antibody coated beads (Grey circles) against FLAG-tagged dCas9. Purified CRISPR DNA-protein complexes were reverse cross-linked and analysed by mass spectrometry for identification of the unknown host factors.

(B) Gel depicting the SDS-PAGE of the sample that is pulled down using Protein G beads coated with Anti-FLAG M2 antibody against FLAG-tagged dCas9. The immunoprecipitated sample is shown in lane 3 and the sample corresponding to untreated beads that act as control is shown in lane 1. Protein marker (M) positions are indicated on the right, whereas, positions corresponding to IgG1 antibody (heavy chain (HC) and light chain (LC)), dCas9 and Protein G are shown on the left.

Further the immunoprecipitated fractions were reverse cross-linked by heating at 95 °C and electrophoresed. After this, in-gel trypsin digestion, peptide extraction and mass spectrometry were performed (Mass Spectrometry facility at C-CAMP, Bangalore). The identified peptides from the mass spectrometry analysis were mapped to the proteome of *E. coli*. Most of these identified peptides belonged to cellular housekeeping machinery (Figure 3.2Figure 2.2 and Appendix Table 4). Remarkably, a few of the identified peptides were mapped to Cas1 and Cas2 (S/N 16 and 55 in Appendix Table 4). Among others, DNA architectural proteins such as H-NS and IHF and DNA repair proteins such as RecA were observed (S/N 7, 36, 62 and 91 in Appendix Table 4). As IHF is known to facilitate site-specific integrases (Chalmers et al., 1998; Friedman, 1988; Miller et al., 1980) and the Cas1-2 functions like an integrase (Nunez et al., 2015b), we were tempted to probe the involvement of IHF in spacer acquisition.

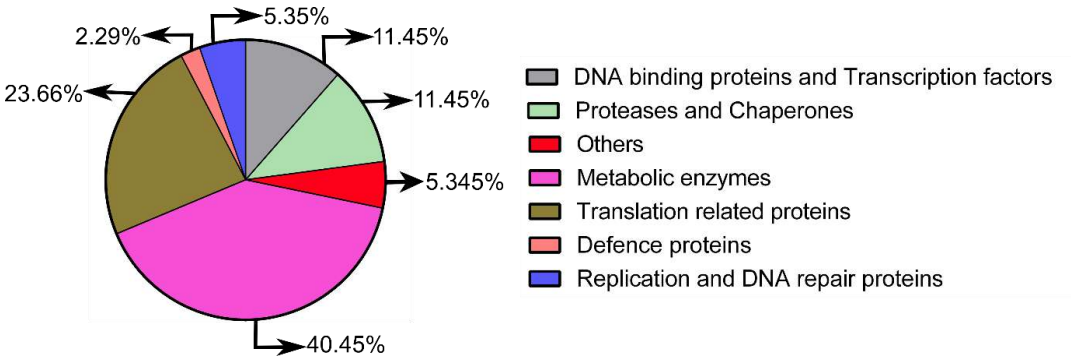


Figure 3.2: Interactome captured by immunoprecipitation of CRISPR DNA

Proteins identified by mass spectrometry of dCas9 directed pull-down of CRISPR DNA (Appendix Table 4) were categorised based on predicted or documented activity. The percentage distribution of each representative protein class is represented as a Pie chart.

3.3.2. IHF is essential for prespacer acquisition into the CRISPR locus *in vivo*

Here we employed the previously established *in vivo* spacer acquisition assay to validate the essentiality of predicted host factors during CRISPR adaptation (Figure 2.6) (Yosef et al., 2012). In the assay, appearance of expanded CRISPR array in Cas1-2 induced *E. coli* IYB5101 (WT) (Lane 2 in Figure 2.6C) signifies the existence of all the indispensable factors for spacer integration. As IHF was supposed to be a potentially suitable host protein (among the immunoprecipitated host factors) that could support spacer integration, acquisition assays were attempted with IHF null mutant strains to test its essentiality.

IHF is a heterodimer comprising of α and β subunits (Figure 3.3A) (Friedman, 1988; Rice et al., 1996). Hence, a null mutant of IHF devoid of either α or β subunit in *E. coli* IYB5101 was generated and spacer acquisition assay was performed. Surprisingly, in these mutants, no expansion in CRISPR array was seen (compare Lane 2 with 4 and 6 in Figure 3.3B). This indicates the abrogation of spacer acquisition in the strains that lack either α or β subunit of IHF. To reinforce these observations, null mutants were complemented with plasmid-borne IHFαβ expression. Upon complementation, restoration in the expansion of

CRISPR locus was noticed (compare Lane 4 with 8 and Lane 6 with 10 in Figure 3.3B). These observations strengthened our conjecture that the acquisition of prespacer requires the participation of IHF *in vivo* (Figure 3.3B and C).

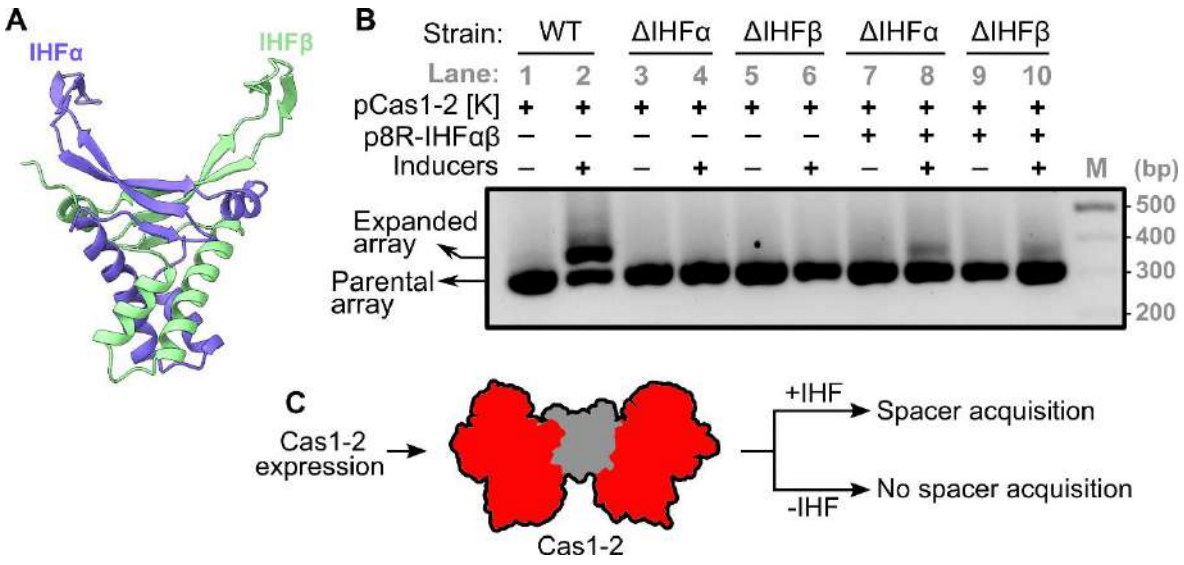


Figure 3.3: IHF triggers spacer uptake in *E. coli*

(A) Structure of IHF heterodimer (PDB ID: 1IHF) highlighting the α and β protomers (Blue and Green, respectively).

(B) Agarose gel displaying the result of spacer acquisition assay performed in WT (Lanes 1 and 2), ΔIHFα (Lanes 3, 4, 7 and 8) and ΔIHFβ (5, 6, 9 and 10) strains. Absence (-) or presence (+) of pCas1-2[K], p8R-IHFαβ and inducers is shown on top of each lane. Positions of parental and expanded arrays are denoted on the left and the positions corresponding to DNA marker are shown on the right.

(C) Schema highlighting the observations made in spacer acquisition assay. Abrogation of spacer integration in IHF null mutants underscores the essentiality of IHF in CRISPR adaptation.

3.4. Discussion

Polarised spacer integration is a notable feature of CRISPR adaptation. However, owing to intrinsic target sequence specificity towards short motifs (Rollie et al., 2015), Cas1-2 seems to integrate spacers at off-targets *in vitro* (Nunez et al., 2015b). Such random spacer insertions could potentially lead to mutations and genomic instability *in vivo*; this, in turn, could switch-off critical metabolic pathways and endanger the viability of the cell. Also, previous research study had demonstrated that the hosts trigger a spontaneous immune response if the invader's protospacers were matching the spacers that were more proximal to the leader (McGinn and Marraffini, 2016b). To avert the perils of off-target integration by Cas1-2 and to avail the fitness against pathogenic invaders, CRISPR-Cas systems seem to have developed an elegant mechanism to favour site-specific spacer integration (at the leader-repeat junction). Among all the Cas proteins of *E. coli* (type I-E) only Cas1 and Cas2 are required for spacer integration (Yosef et al., 2012). Hence, we strongly presumed that some of the host proteins in *E. coli* might be acting as a specificity determining factors.

The cytoplasm of *E. coli* harbours thousands of proteins that catalyse hundreds of biochemical pathways. Therefore, it is spatiotemporally challenging to identify the factors that interact with CRISPR/Cas components during spacer integration at the target site on the genomic DNA. To explore this task, a molecular tool that detects the CRISPR/Cas interactome is the need of the hour. Previous studies had demonstrated the utility of nuclease null dCas9-sgRNA (*Streptococcus pyogenes* type II-A CRISPR interference complex) as a biomolecular tool to bind at the specified DNA sequence around the promoters of various genes and tune their expression (in prokaryotes and eukaryotes) (Gilbert et al., 2013; Qi et al., 2013). Also, the sequence directed anchoring ability of dCas9-sgRNA was harnessed to isolate desired genomic regions from eukaryotic cells and the interacting proteins were successfully identified (Fujita and Fujii, 2013). We adopted this strategy in *E. coli* and detected various proteins that interact around the target site of dCas9-sgRNA complex (i.e., CRISPR locus of *E. coli*). Cas1 and Cas2 are the factors that are known to interact at the CRISPR locus. Thus, the appearance of peptides related to Cas1 and Cas2 in the pulled down fractions bolsters the utility of this approach. Moreover, a vast majority of the immunoprecipitated proteins were mapped to ribosomal components and transcription machinery – an aspect characteristic of their omnipresence due to their housekeeping functions (Figure 3.2 and Appendix Table 4).

As Cas1-2 complex shows integrase-like activity, we hypothesised that the presence of host factors that are previously characterised to facilitate the integration of DNA elements would be prospective candidates. We filtered out the factors that were previously shown not to be involved in CRISPR adaptation (Levy et al., 2015; Swarts et al., 2012; Westra et al., 2010) or were functionally unrelated such as chaperones, proteases, metabolic enzymes, etc. For example, though the DNA architectural protein H-NS was identified with a higher score than that of Cas1 and Cas2 (compare S/N 16 and 55 with 7 in Appendix Table 4), it was previously demonstrated that H-NS was inessential for CRISPR adaptation and that it acted as a repressor of *cas* operon in *E. coli* (Pul et al., 2010; Swarts et al., 2012). Therefore, we did not pursue further with H-NS and a similar rationale was exercised to exclude other factors. On the other hand, though another architectural protein IHF scored lower than H-NS (compare S/N 62 and 91 with 7 in Appendix Table 4), we chose IHF as a suitable candidate, as it was shown to facilitate the site-specific recombination by λ integrase (Friedman, 1988; Moitoso de Vargas et al., 1989).

Though we identified IHF as an indispensable accessory factor (Figure 3.3) for CRISPR adaptation in *E. coli*, the mechanism by which IHF stimulates the spacer integration remained elusive. IHF is a nucleoid-associated protein (NAP) and it is known to restructure the DNA and regulate various processes such as replication, transcription regulation, transposition and recombination (de Lorenzo et al., 1991; Miller et al., 1980; Moitoso de Vargas et al., 1989; Ryan et al., 2002). Henceforth, in the following chapters, we sought to understand if IHF could interact at CRISPR and facilitate the spacer integration.

3.5. Summary

In the current chapter, we established the CRISPR/dCas9 based immunoprecipitation strategy in *E. coli*. Utilising this molecular tool, we identified the proteins that interact at the CRISPR locus. Among these, we selected IHF as a probable candidate that could support spacer integration. Utilising *in vivo* spacer acquisition assay, we noticed the abolishment of spacer integration in *ihf* knock-out strains. Upon expressing IHF episomally in these knock-out strains, we observed the restoration of spacer acquisition. Hence, we determined IHF as an essential accessory factor for CRISPR adaptation.

Chapter IV

Unravelling the mechanistic role of IHF in CRISPR adaptation

4. Chapter IV

4.1. Introduction

In the previous chapter, we substantiated the involvement of IHF during spacer integration, but the mechanism by which IHF guides CRISPR adaptation in *E. coli* remained elusive. Being a NAP, IHF shares high sequence and structural resemblance with another NAP, HU. Despite these similarities, unlike HU, IHF displays sequence-specific DNA binding (Dillon and Dorman, 2010; Friedman, 1988; Rice et al., 1996). Apart from directing different viral recombinases and integrases via DNA sequence guided architectural rearrangement, IHF is also shown to regulate gene expression in bacteria (de Lorenzo et al., 1991; Martinez-Santos et al., 2012). Hence, we wondered if IHF interacts with CRISPR locus and regulates spacer acquisition or it could indirectly regulate CRISPR adaptation by tuning any other host protein expression. While addressing this quest, in the current chapter, we identified a motif in the leader region that matches to the IHF binding consensus. Utilising *in vivo* spacer integration assays and *in vitro* biochemical assays we proved that this motif is critical for spacer acquisition and the interaction of IHF at this motif induces a sharp bending in CRISPR DNA. Further, we went on to establish an *in vitro* integration assay and demonstrated that the IHF induced structural deformation prompts the spacer integration into the CRISPR locus.

4.2. Materials and Methods

4.2.1. Construction of bacterial strains and plasmids

Descriptions of the strains, plasmids and oligonucleotides are listed in Appendix Table 1, Table 2 and Table 3, respectively. Plasmid pCSIR-T (Diez-Villasenor et al., 2013) was used as a template to amplify WT array and IHF binding site mutants (IBS and ΔIBS). WT and mutant array amplicons were individually inserted in between KpnI/PstI sites in plasmid pOSIP-CT (St-Pierre et al., 2013) and subsequently integrated into Phi 21 (P21) locus of *E. coli* IYB5101 strain by a one-step process of cloning and integration into attB locus termed as "clonetegration".

To generate plasmid pBend-WT, 81 bp complementary oligos encompassing 69 bp of WT leader sequence was annealed and end filled by PCR. This DNA construct was phosphorylated using T4 polynucleotide kinase and inserted into plasmid pBend5 using HpaI site (Zwieb and Adhya, 2009).

E. coli K-12 MG1655 genomic DNA was used as a template to amplify genes encoding IHFα and IHFβ. To generate expression plasmid p1R-IHFαβ, a bicistronic cassette encoding IHFα and IHFβ was amplified and inserted into plasmid p1R using SspI site.

Gibson assembly protocol was utilised for generating all the recombinant vectors and clonetegrated strains (Gibson et al., 2009). After this, the resultant constructs were verified by Sanger sequencing (Sanger et al., 1977).

4.2.2. Expression and purification of proteins

E. coli BL21(DE3) harbouring p1R-IHFαβ was grown in Terrific broth supplemented with 100 μg/ml Kanamycin at 37 °C till the OD_{600} reaches 0.6. At this point, IHF expression was induced with the addition of 0.5 mM IPTG and the cells were allowed to grow for 4 hrs at 37 °C. After that, the cells were harvested and resuspended in IHF binding buffer (20 mM Tris-Cl pH 8, 150 mM NaCl, 10 % Glycerol and 6 mM β-ME) containing 1 mM PMSF. The cells were then subjected to lysis by sonication and clarified soluble extract was loaded on to

5 ml StrepTrap HP column (GE Healthcare). After loading, the column was washed with IHF binding buffer and proteins were eluted with IHF binding buffer containing 2.5 mM D-desthiobiotin (Sigma). Eluted protein fractions were pooled up and loaded on to 5 ml HiTrap Heparin HP column (GE Healthcare). The column was washed with IHF binding buffer and bound proteins were eluted with a linear gradient of 0.15 – 2 M NaCl in IHF binding buffer. Purified fractions were pooled and dialysed against IHF binding buffer. Dialysed protein was concentrated, flash frozen and stored at −80 °C until required.

Episomally expressed Cas1 and Cas2 were affinity purified as per the protocol in Section 2.2.2.

4.2.3. Spacer acquisition assays

In vivo spacer integration assay for the strains that encompass mutation in IHF binding sites (IBS and ΔIBS) was performed as described in Section 2.2.6. As the WT, IBS and ΔIBS CRISPR variants were integrated into P21 attB site in *E. coli* IYB5101 genome via pOSIP-CT (Section 4.2.1), primers annealing to upstream of KpnI site and downstream of PstI site in pOSIP-CT were utilised to monitor CRISPR array expansion by PCR.

4.2.4. Electrophoretic Mobility Shift Assays

WT or mutant leader DNA (IBS and ΔIBS) was PCR amplified from the strain carrying the respective construct that is integrated into P21 locus. 14 nM of amplified DNA was incubated with increasing concentration of purified IHF (0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.2, 1.4 and 1.6 µM) in buffer containing 0.5X TBE (50 mM Tris-Cl pH 8.3, 50 mM Boric acid and 1mM EDTA), 100 mM KCl, 10 % Glycerol and 5 µg/ml BSA for 30 mins at 25 °C. Post-incubation samples were directly loaded on 8 % native polyacrylamide gel and electrophoresed in 1X TBE at 4 °C. Gels were post-stained with Ethidium bromide (EtBr) and DNA bands were visualised in the gel documentation system (Bio-Rad).

4.2.5. FRET-based monitoring of DNA bending

A 35 bp DNA encompassing leader sequence (-4 to -38 from the leader-repeat junction) of WT (or WT without quencher or ΔIBS) was assembled from three oligos by annealing. This DNA construct contains 6-FAM and Iowa Black as 3'- and 5'- end labels, respectively (IDT). 222 nM of DNA probe was incubated with increasing concentrations of purified IHF (0, 0.3, 0.6, 0.9, 1.2, 1.5 and 1.8 µM) in buffer containing 0.5X TBE, 100 mM KCl, 10 % Glycerol and 5 µg/ml BSA for 20 mins at 25 °C. Post-incubation samples were excited at 495 nm and emission was monitored from 500-600 nm, with averaging over three scans in FluoroMax-4 spectrofluorometer (Horiba Scientific, Edison, NJ). The slit width used for excitation and emission was 2 nm and 7 nm, respectively. After background correction, the fluorescence intensity of DNA in the presence of IHF is normalised relative to that of DNA alone. To further ascertain that the enhanced quenching is due to IHF mediated DNA bending; a fluorescence recovery assay was designed. In this assay, buffer (0.5X TBE, 100 mM KCl, 10 % Glycerol and 5 µg/ml BSA) containing 222 nM DNA was excited at 495 nm and emission was captured for 200 seconds at 520 nm. To this sample, IHF was added to a final concentration of 1.8 µM and fluorescence emission was recorded till 600 seconds. Thereafter, IHF degradation and DNA release were initiated by the addition of Proteinase K to a final concentration of 1 mg/ml and emission was monitored for another 400 seconds. The temporal change in fluorescence emission was plotted by normalising the fluorescence intensity at each time point to that of fluorescence intensity of DNA at 0^{th} second.

4.2.6. Estimation of bending angles by circular permutation gel retardation assay

Plasmid pBend-WT was digested with HindIII and EcoRI to produce a 329 bp DNA fragment. This fragment was gel purified as per the manufacturer's instruction (Qiagen) and digested with BamHI, KpnI, SspI, EcoRV, SpeI, BglII and MluI in separate reactions. All the digested DNA samples were further purified (Qiagen) and 21 nM of each DNA was incubated individually with 0.7 µM IHF in buffer containing 0.5X TBE, 100 mM KCl, 10 % Glycerol and 5 µg/ml BSA for 30 mins at 25 °C. Post-reaction samples were directly loaded on 8 % native polyacrylamide gel and electrophoresed in 1X TBE at 4°C. Gels were post-stained with

EtBr and DNA bands were visualised in the gel documentation system. IHF bending angles were calculated as described previously (Papapanagiotou et al., 2007). Mobilities of IHF bound DNA complex (R_b) and the respective free DNA (R_f) were calculated for all the restriction-digested fragments. R_b values were normalised to the respective R_f values and were plotted against flexure displacement (length from the middle of the binding site to the 5' end of the restriction fragment/total restriction fragment length). The resulting plot was fitted to a quadratic equation: $y = ax^2 - bx + c$, where x and y denotes flexure displacement and R_b/R_f, respectively. The bending angle (α) was calculated using the relationship $a = -b = 2c(1 - cos\alpha)$. Here, we have represented the bending angle (α) as the average value that was calculated from the parameters a and b.

4.2.7. *In vitro* integration assay

WT or mutant leader DNA (IBS and ΔIBS) was PCR amplified from the strain carrying the respective construct that is integrated into P21 locus. Prespacer DNA (P23[3'-5]: 23 bp duplex and 5 nt 3'-overhangs) is prepared by annealing the complementary oligos. *In vitro* integration assays employing Cas1 and Cas2 were performed as previously described (Nunez et al., 2015b) with few modifications. 210 nM of Cas1 and Cas2 were mixed and incubated at 4 °C for 15 mins. 550 nM of P23[3'-5] was added to the mixture and incubation at 4 °C was continued for another 15 mins. To this complex, 21 nM of CRISPR DNA substrate (WT or Mutant leader) was added along with 0.7 µM IHF in duplicates and incubated at 37 °C for 60 mins in buffer containing 20 mM HEPES-NaOH pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ and 1 mM DTT. The first set of reaction mixtures were directly loaded and electrophoresed on 8 % native polyacrylamide gel in 1X TBE at 4 °C. Whereas, the second set of reaction mixtures were treated with 1 mg/ml Proteinase K for 30 mins at 37 °C before electrophoresis. Electrophoresed gels were post-stained with EtBr and imaged in the gel documentation system.

4.2.8. Spacer disintegration assay

The reaction mixture from integration assay was purified using the PCR purification kit (Qiagen) as per the manufacturer's instruction. 210 nM of Cas1 and Cas2 were mixed and incubated at 4 °C for 15 mins. To this complex, 21 nM purified integration product was mixed with or without 0.7 µM IHF and incubated at 37 °C for 60 mins in buffer containing 20 mM HEPES-NaOH pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ and 1 mM DTT. Subsequently, Proteinase K was supplemented to a final concentration of 1 mg/ml concentration and incubated for 30 mins at 37 °C. The sample was mixed with 6X DNA loading dye and electrophoresed on 8 % native polyacrylamide gel in 1X TBE at 4 °C. Electrophoresed gels were post-stained with EtBr and imaged in the gel documentation system.

4.2.9. Sequence comparison of CRISPR leader derived from type I-E individuals

The multiple sequence alignment corresponding to the leader region for type I-E CRISPR system was obtained from the CRISPRleader database (Alkhnbashi et al., 2016). The conservation profile was generated using WebLogo 3 (Crooks et al., 2004).

4.3. Results

4.3.1. CRISPR leader encompasses an IHF binding motif

IHF heterodimer displays a sequence-specific binding that is targeted to the consensus sequence 5'-WATCAANNNNTTR-3' (where W – A/T, N – A/T/G/C, R – A/G) (Chalmers et al., 1998; Leong et al., 1985; Moitoso de Vargas et al., 1989). Therefore, we searched for potential IHF binding site abutting the CRISPR 2.1 locus in *E. coli* IYB5101 as well as in related strains. This search led to the identification of a putative IHF binding site upstream of the first CRISPR repeat (Figure 4.1).

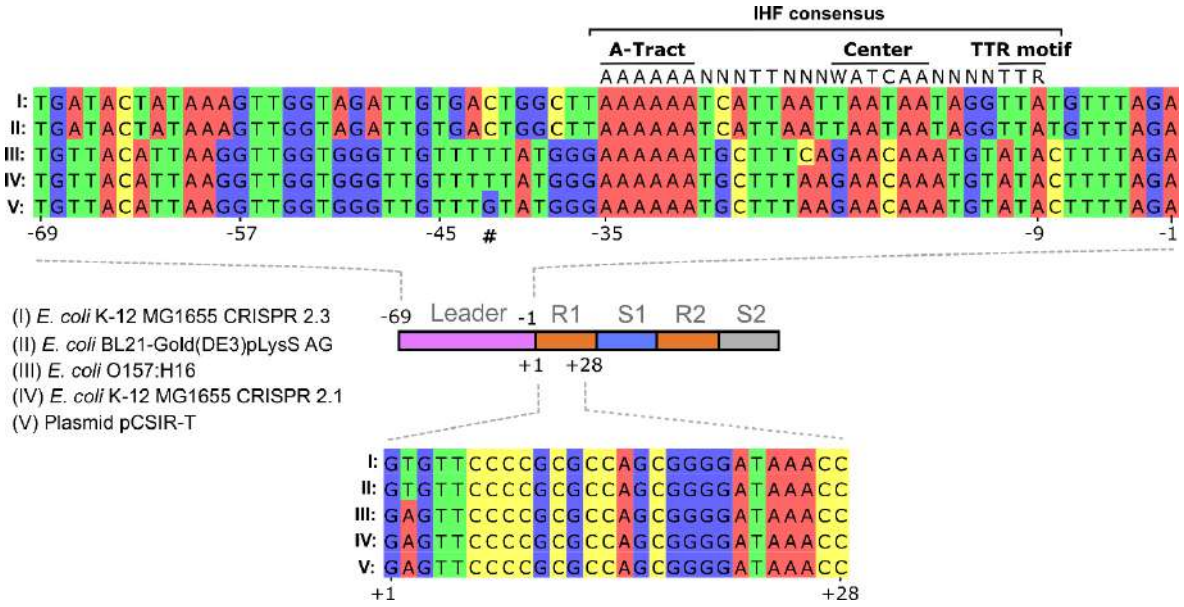


Figure 4.1: Sequence comparison of CRISPR leader and repeat from related *E. coli* strains

Diagram of CRISPR locus is depicted. CRISPR DNA elements, viz., leader, repeats (R1-R2) and spacers (S1-S2) are labelled. Alignments of sequences harbouring the leader and repeat1 from *E. coli* strains O157:H16, K-12 (CRISPR array 2.3 and 2.1), BL21-Gold(DE3) plysS AG and plasmid pCSIR-T are shown. Positions of the residues (with respect to leader-repeat junction) are labelled on the bottom of the alignments. A single mutation present in CRISPR 2.1 array of plasmid pCSIR-T with respect to K-12 2.1 array is represented by '#'. IHF binding site consensus in the leader region is represented at the top of the alignment.

Intrigued by identifying IHF binding consensus, we sought to understand if this region could regulate spacer integration. To test this, we partially deleted the IHF binding site (ΔIBS in Figure 4.2A) in *E. coli* IYB5101 and assayed for the spacer acquisition. Interestingly, no expansion of the array was noted (compare Lane 2 with 6 in Figure 4.2B). Similarly, mutation of the key putative IHF binding nucleotides (IBS in Figure 4.2A) also abolished the acquisition (compare Lane 2 with 4 in Figure 4.2B). These observations suggest that the putative IHF binding site indeed impacts the spacer acquisition.

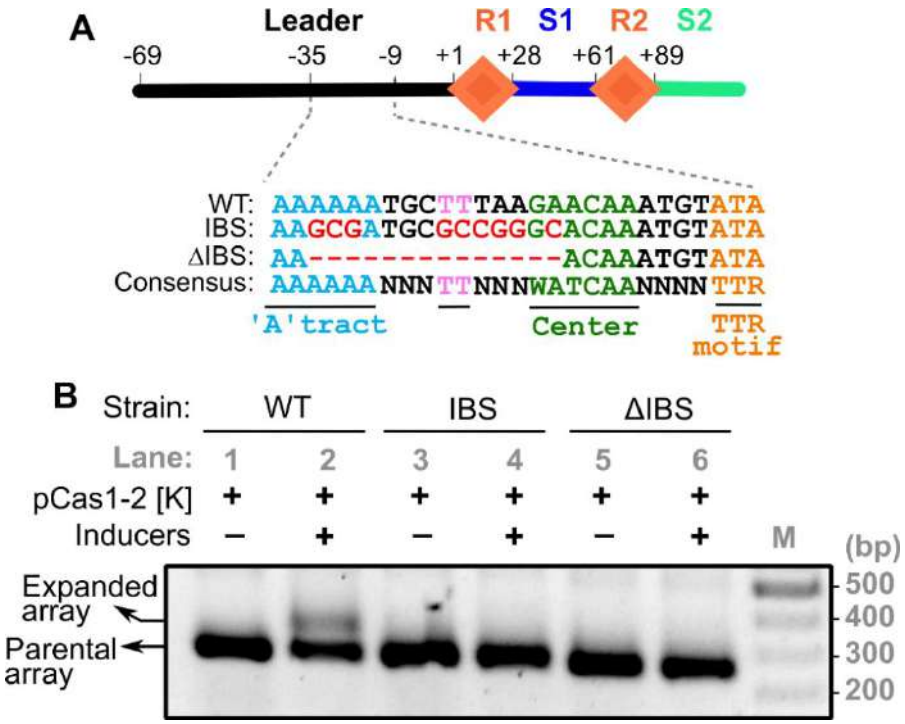


Figure 4.2: Mutations in putative IHF binding site of CRISPR leader abolishes the spacer uptake *in vivo*

(A) CRISPR locus encompassing leader (Black), repeats (R1-R2 as Orange diamonds) and spacers (S1-S2 in Blue and Green, respectively) is shown. Numbering on top of the schema represents the position of nucleotides with respect to the leader-repeat junction. Putative IHF binding site corresponding to the wild type (WT), mutated (IBS) and deleted (ΔIBS) region is shown. Conserved submotifs were underlined and highlighted in different colours.

(B) PCR products from spacer acquisition assay performed in *E. coli* strains harbouring CRISPR variants (WT (Lanes 1-2), IBS (Lanes 3-4) and ΔIBS (Lanes 5-6)) at P21 attB site are shown. Absence (-) or presence (+) of plasmid pCas1-2[K] and inducers is shown on top of each lane. Positions corresponding to parental and expanded arrays are indicated on left. DNA marker (M) positions are represented on the right.

Electrophoretic mobility shift assay (EMSA) was performed to test the binding of IHF at the putative interaction site. In this assay, the differences in the mobility of free DNA and higher molecular weight DNA-protein complex was monitored to determine if the protein interacts with the nucleic acids (Figure 4.3A). To assess the IHF-CRISPR locus interaction by EMSA, amplified CRISPR DNA was incubated with increasing concentration of purified IHF complex (Figure 4.3B). These reaction mixtures were electrophoresed and the mobility differences were noted (Figure 4.3C-E). Prominent retardation in DNA mobility was observed with the presence of IHF in case of WT (Figure 4.3C). This observation indeed indicates the binding of IHF to CRISPR leader. To reinforce these findings, EMSA was repeated with mutated variants of CRISPR leader (IBS and ΔIBS in Figure 4.3D and E, respectively). In both cases, a drastic reduction in IHF binding was noted and only smears were seen in the lanes that contained higher concentrations of IHF (compare Figure 4.3D and E with C) (at higher concentrations, IHF was shown to interact with DNA non-specifically (Holbrook et al., 2001; Lin et al., 2012)). Overall, these results give a clue that the disruption of IHF binding at CRISPR leader in IBS and ΔIBS strains could have potentially contributed to the inhibition of spacer uptake (Figure 4.2B).

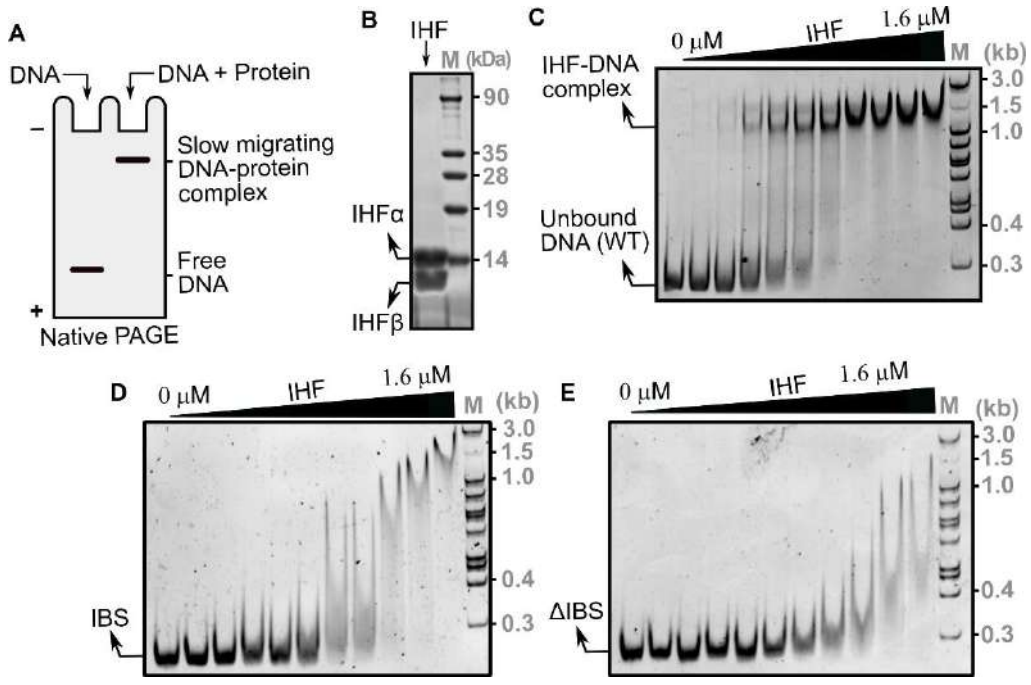


Figure 4.3: IHF interacts with CRISPR leader at the predicted binding region

(A) Picture depicting the outcome of EMSA for identifying the DNA-protein complex formation. Generally, the interaction of nucleic acid binding proteins with DNA results in

the formation of higher molecular weight nucleoprotein complexes. This complex formation can be traced by resolving the reaction mixtures by native PAGE and comparing the mobility differences.

(B) Gel displaying the SDS-PAGE of purified IHF. Positions of N-terminally Strep tagged IHFα and untagged IHFβ are indicated on the left. Protein molecular weight marker (M) positions are shown on the right.

(C-E) Native polyacrylamide gel displaying the results of EMSA performed using IHF and CRISPR DNA variants (WT (C), IBS (D) and ΔIBS (E)). 14 nM of CRISPR DNA was incubated with increasing concentration of IHF (0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.2, 1.4 and 1.6 μM). The corresponding positions of unbound and IHF bound DNA are represented on the left. DNA marker (M) positions are shown on the right.

4.3.2. IHF interaction prompts bending of the leader region

The structure of IHF-DNA complex shows that the IHFα and β form an intertwined compact body from which two β structures protrude out clamping the DNA (Rice et al., 1996). This induces bending of DNA by about 160° leading to the reversal of the direction of DNA (Figure 4.4A). Motivated by the IHF binding to the CRISPR leader (Figure 4.3), we wondered whether the binding leads to bending of the DNA. To assess this, we designed a FRET-based assay wherein one end of the IHF binding region is tagged with a fluorophore (6-FAM) and the other end with the quencher (Iowa Black). In the linear DNA, the fluorophore and the quencher will be sequestered and hence this will not quench the fluorescence. However, if IHF bends the DNA, this brings both the fluorophore and quencher into proximity leading to quenching of the fluorescence (Figure 4.4D). Upon incubation of labelled DNA fragments derived from WT and ΔIBS CRISPR loci (-4 to -38 from the leader-repeat junction) with increasing concentrations of IHF, we observed steady decrement in fluorescence intensity of WT DNA alone (Figure 4.4B). Owing to the enfeebled interaction of IHF at the mutated binding site in ΔIBS (Figure 4.3E), no apparent reduction in fluorescence was noted (Figure 4.4C and D). These observations indicate that the interaction of IHF at binding site results in fluorescence quenching. To further corroborate this, a fluorescence recovery assay was designed. In this assay, Proteinase K was supplemented to the sample containing IHF-DNA complex and the change in fluorescence emission was recorded during the degradation of IHF (Figure 4.4E). A drastic reduction in the fluorescence intensity was noted upon IHF addition to WT (Figure 4.4E). Upon addition of protease to this sample, a steady increment in fluorescence emission intensity was observed (Solid line in Figure 4.4E). On the contrary, a

similar experiment performed with the 6-FAM labelled DNA, albeit without the quencher, showed that despite the addition of IHF the intensity of the fluorescence remained constant (dotted line in Figure 4.4E). These experiments reaffirm that IHF indeed bends the leader region and brings the two ends into proximity.

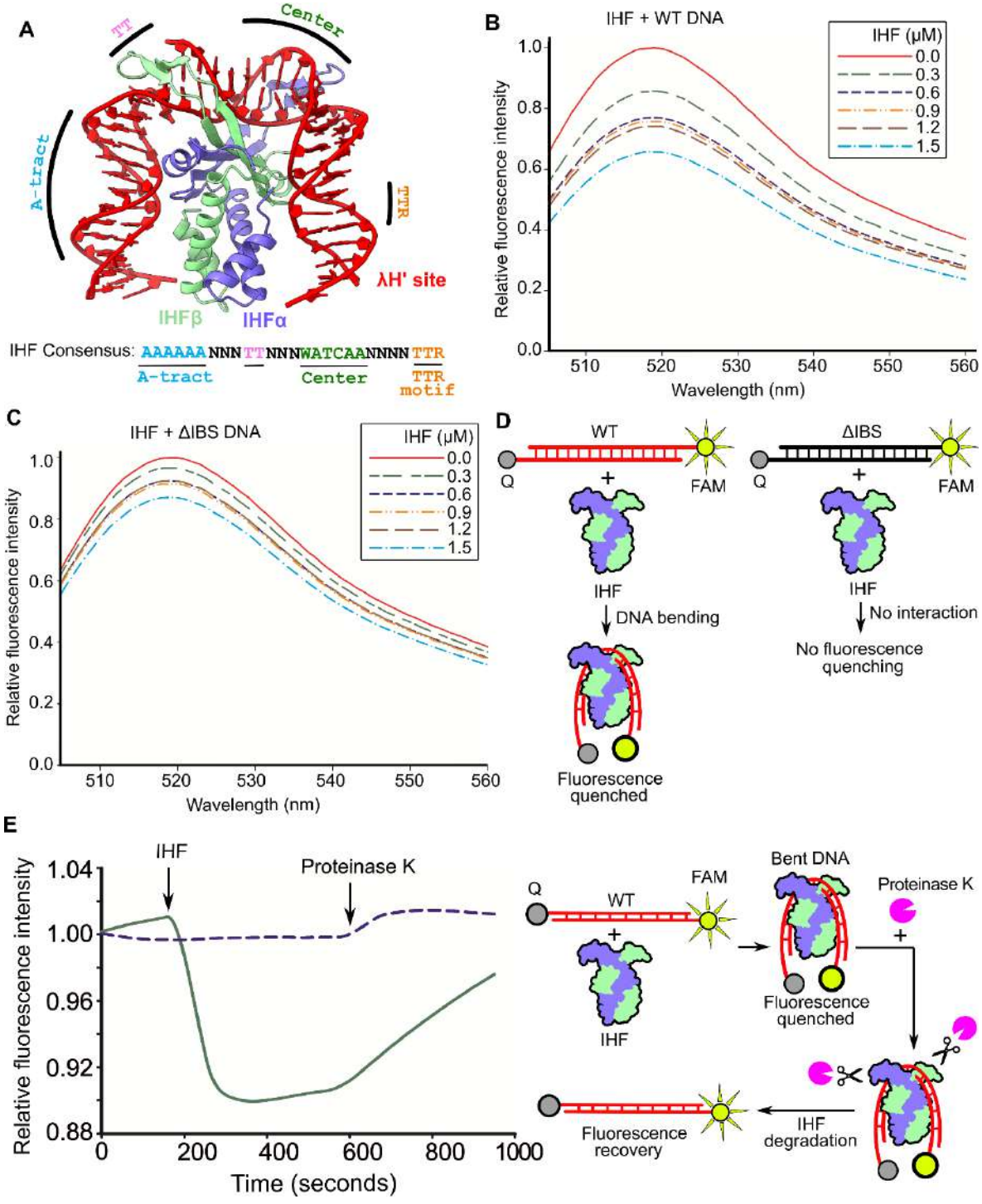


Figure 4.4: Monitoring of IHF induced bending by FRET

(A) Structure of IHF in complex with λ DNA H' site (PDB ID: 1IHF). IHFα and IHFβ are presented in Blue and Green ribbons, respectively. λ DNA H' strands are coloured Red. IHF binding site consensus is shown below the structure and the corresponding positions of submotifs (A-tract, TT, Center and TTR) in the IHF binding region of λ DNA H' site are labelled.

(B-C) Plots displaying the change in fluorescence emission intensities of WT (panel B) and ΔIBS (panel C) DNA probes (normalised to the intensity of the sample that contains only DNA) with an increase in IHF concentration (inset panel). 222 nM of labelled DNA probe that encompasses -4 to -38 region of either WT (Red ladder in panel D) or ΔIBS (Black ladder in panel D) was used in the assay.
(D) Illustration of fluorescence quenching experiments performed using WT (panel B) and ΔIBS (panel C). Interaction of IHF (Blue and Green) bends the WT (red ladder) and positions the Iowa Black labelled end (Q in Grey) close to FAM labelled end (Yellow), thus quenching the fluorescence. Unlike this, the mutations in ΔIBS (Black ladder) disrupt the IHF interactions and therefore hampers the quenching by Iowa Black.
(E) Plot depicting the fluorescence recovery upon degradation of IHF. One end of the WT DNA (in red) was labelled with FAM (Yellow) and the other end with Iowa Black (Q in Grey) or left unlabelled. IHF (Blue and Green) and Proteinase K (Magenta) that were added at various time points are indicated. Addition of IHF to DNA leads to quenching of fluorescence, which was later restored by the addition of Proteinase K. Solid line shows intensities corresponding to fluorescent DNA with quencher whereas the dotted line indicates intensities of fluorescent DNA without a quencher.

Having established the fact that IHF indeed bends the leader region, we were interested in investigating the extent to which IHF bends the leader DNA. To address this, we utilised the bending vector pBend5, which contains circularly permuted duplicated restriction sites (Zwieb and Adhya, 2009). Cloning of the IHF binding site (IBS) into pBend5 and subsequent digestion using the restriction enzymes ensure fragments with the same length but with the binding site distributed to different positions, either in the middle or towards the end (Figure 4.5A). When the DNA undergoes bending due to protein binding, the fragment that harbours the binding site in the middle migrates slower than the one with the binding site at the end (Figure 4.5A). From these mobility differences, it is possible to estimate the bending angle, which is defined as the angle by which the DNA deviates from the linearity (Section 4.2.6). The mobility differences of free restriction digested DNA with that of IHF-DNA complex was estimated from the electrophoresed native gels (Figure 4.5B). Utilising this data, we estimated that IHF bends DNA by 122°, suggesting that the sharp deformation could result in the reversal of the DNA direction (Figure 4.5C).

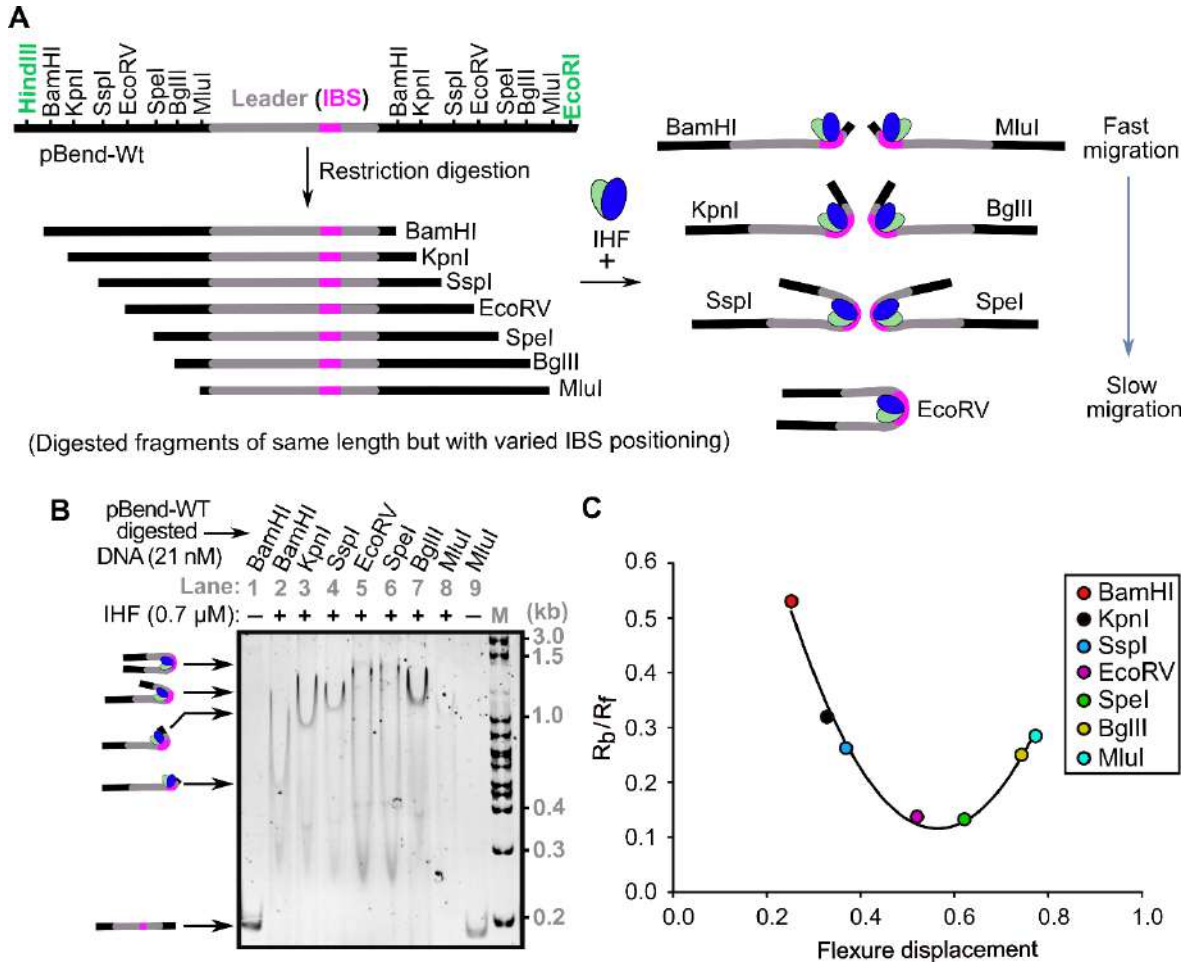


Figure 4.5: IHF interactions deform the CRISPR leader by 120°

(A) Schematic representation of circular permutation gel retardation assay. Seven restriction enzymes (Black) were used to generate equal-sized DNA fragments carrying the CRISPR leader (Grey) and IHF binding site (Magenta) in different positions. EcoRI and HindIII (Green) were used to excise the cassette from pBend-WT. Cartoon on the right displays different nucleoprotein complex conformations that could result from the interaction of IHF with different restriction digested DNA. In comparison to the complexes that have IBS towards the end (BamHI and MluI fragments), the migration of complex that has IBS at the centre (EcoRV) migrates slower on the native PAGE.

(B) The migration of IHF bound DNA is shown. Enzymes labelled on each lane were used to generate the respective restriction DNA fragments from plasmid pBend-WT. Absence (-) or presence (+) of IHF is indicated on top of each lane. Positions of bent nucleoprotein complexes are shown on the left. DNA marker (M) positions are labelled on the right.

(C) A plot of the relative mobilities (R_b/R_f) of restricted fragments against the flexure displacement for the assay performed in (B). Positions corresponding to R_b/R_f values of the respective restriction fragments are indicated.

4.3.3. IHF induced bending of the linear DNA facilitates prespacer integration

Upon identifying that IHF interacts with the CRISPR leader and induces architectural rearrangements, we sought to examine the conservation of this binding site among the organisms harbouring type I-E system. This analysis identified high conservation of the IHF binding site (-9 to -35 nt; boxed in solid line in Figure 4.6) along with another region (-44 to -59 nt; boxed in dotted line in Figure 4.6) across other species as well.

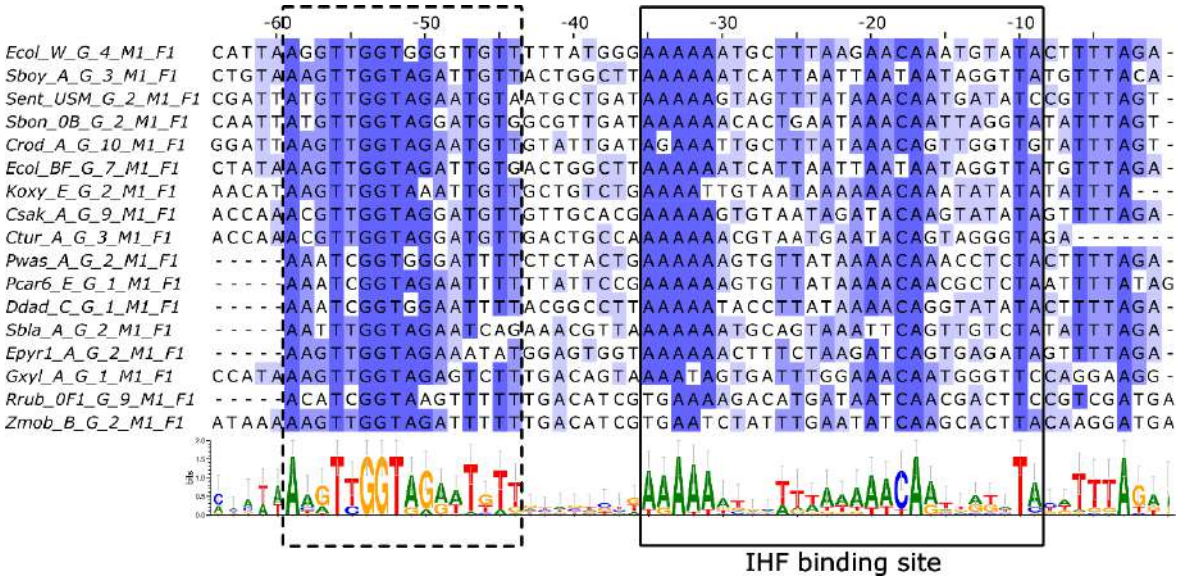


Figure 4.6: IHF binding site is conserved across type I-E individuals

Multiple sequence alignment of CRISPR leader sequences derived from 17 representative type I-E organisms that belong to F1B1 cluster (Alkhnbashi et al., 2016). The annotations mentioned on the left side of the alignment were derived from CRISPRmap (Lange et al., 2013). Positions with respect to the leader-repeat junction are labelled on top of the sequences. Residues in the alignment are highlighted in blue based on the extent of conservation using Jalview (Waterhouse et al., 2009). Conservation profile depicted in the lower section is generated using WebLogo 3 (Crooks et al., 2004). Two conserved regions in the CRISPR leader are highlighted using dotted and solid boxes (IHF binding site).

Motivated by the fact that the IHF binding site is highly conserved in type I-E organisms (Figure 4.6) and this region being indispensable for the spacer acquisition in *E. coli* (Figure 4.2), we aimed to decipher the mechanism by which IHF promotes the prespacer

integration at the CRISPR locus. Therefore, we performed an assay to monitor prespacer integration into linear CRISPR DNA. In this assay, a CRISPR DNA substrate that encompasses conserved regions of the leader with two sets of repeat-spacer units was incubated with prespacer P23[3'-5] (right panel in Figure 4.7A) and purified IHF, Cas1 and Cas2 proteins (left panel in Figure 4.7A). Upon addition of IHF to the CRISPR DNA, a single slow migrating band was noted (Lane 4 in Figure 4.7B). This observation suggests that the IHF induced bending retards the mobility of the CRISPR DNA. Subsequent addition of Cas1-2 complex and prespacer fragment resulted in the appearance of a super-shifted band (Lane 12 in Figure 4.7B). Strikingly, this band was not seen in the absence of IHF (Lane 11 in Figure 4.7B). When the DNA bound proteins were removed using Proteinase K treatment, a slow migrating band that seemed to be larger than the CRISPR DNA was spotted (Lane 12 in Figure 4.7C). Remarkably, this band appeared only from the Proteinase K treated reaction mixture consisting of CRISPR DNA, prespacer, IHF and Cas1-2 (compare Lanes 1-11 with Lane 12 in Figure 4.7C). These findings suggest the possibility that the slow migrating band represents the prespacer integrated into the CRISPR DNA.

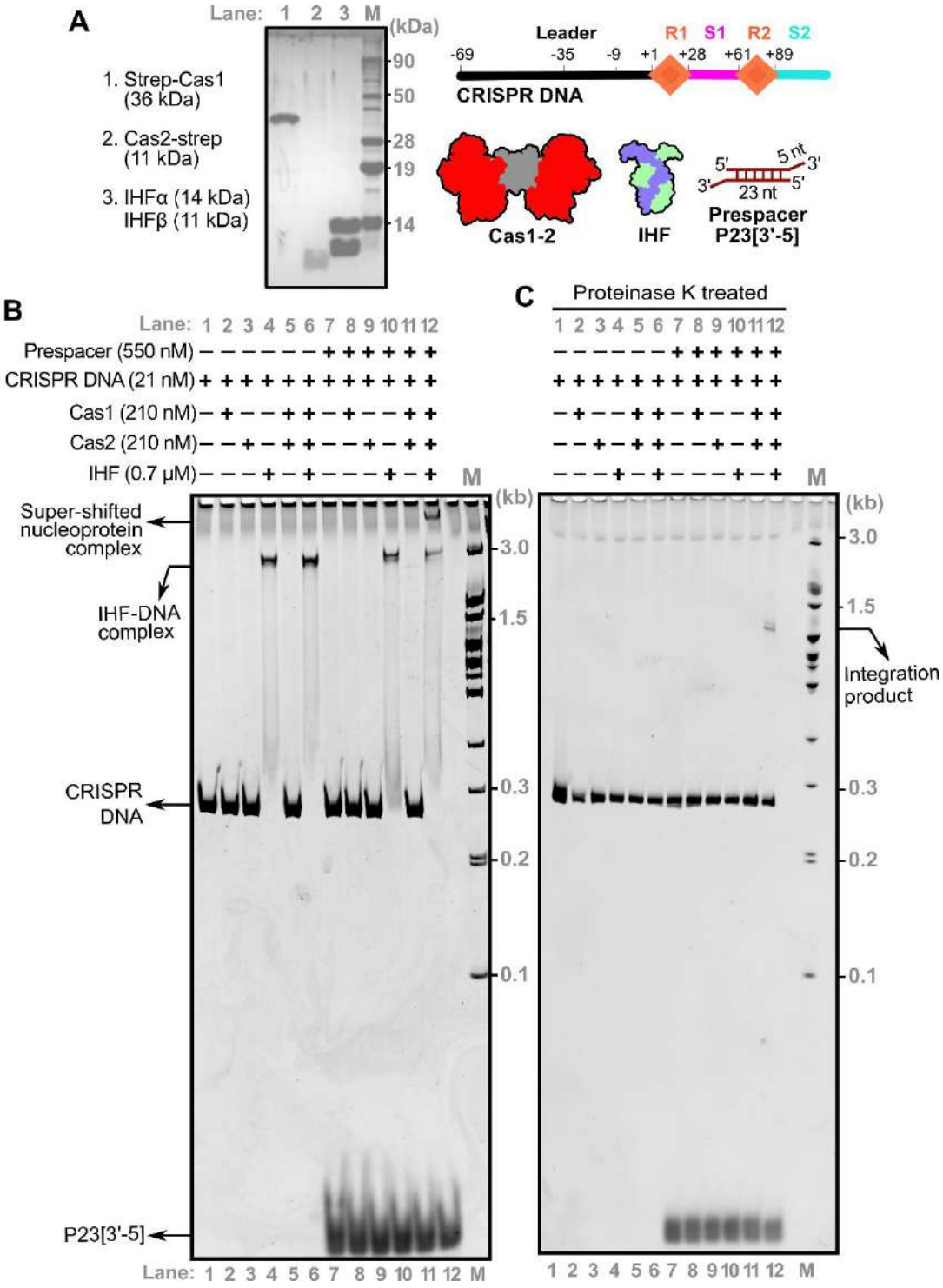


Figure 4.7: IHF necessitates prespacer integration *in vitro*

(A) Gel displaying the SDS-PAGE of purified Cas1 (Lane1), Cas2 (Lane2) and IHF proteins (Lane 3). Molecular weights (kDa) corresponding to proteins in each lane are shown on the left. Cartoon of the components used in integration assay is depicted on the right. Linear CRISPR DNA encompassing leader (Black), repeats (R1-R2 in Orange) and spacers (S1-S2 in Magenta and Cyan, respectively) is shown. Numbering on top of the

schema represents the position of nucleotides with respect to the leader-repeat junction. Prespacer P23[3'-5] contains a 23 bp duplex with 5 nt 3'-overhangs.
(B) Native gel depicting spacer integration assay performed with WT CRISPR DNA. Absence (-) or presence (+) of each reaction component is indicated on top of each lane. Positions of super-shifted nucleoprotein complex, IHF-DNA complex, CRISPR DNA substrate and P23[3'-5] are indicated on the left. DNA marker (M) positions are shown on the right.
(C) Native gel showing Proteinase K treated samples from the assay in (B). Position of integrated product is indicated on the right.

To further probe the requirement of IHF for the formation of super-shifted band, *in vitro* integration assays were performed with CRISPR DNA that encompasses IHF binding site variants (Figure 4.2A). Owing to the deletion (ΔIBS) or mutation of the IHF binding site (IBS), these variants did not interact with IHF and hence did not show IHF-DNA complex (compare Lanes 6 and 9 with 3 in Figure 4.8A) Moreover, the appearance of super-shifted band and the integrated product was also completely abolished in IHF binding site mutants (compare Lanes 6 and 9 with 3 in Figure 4.8A and B). These observations emphasise the importance of IHF binding during CRISPR adaptation.

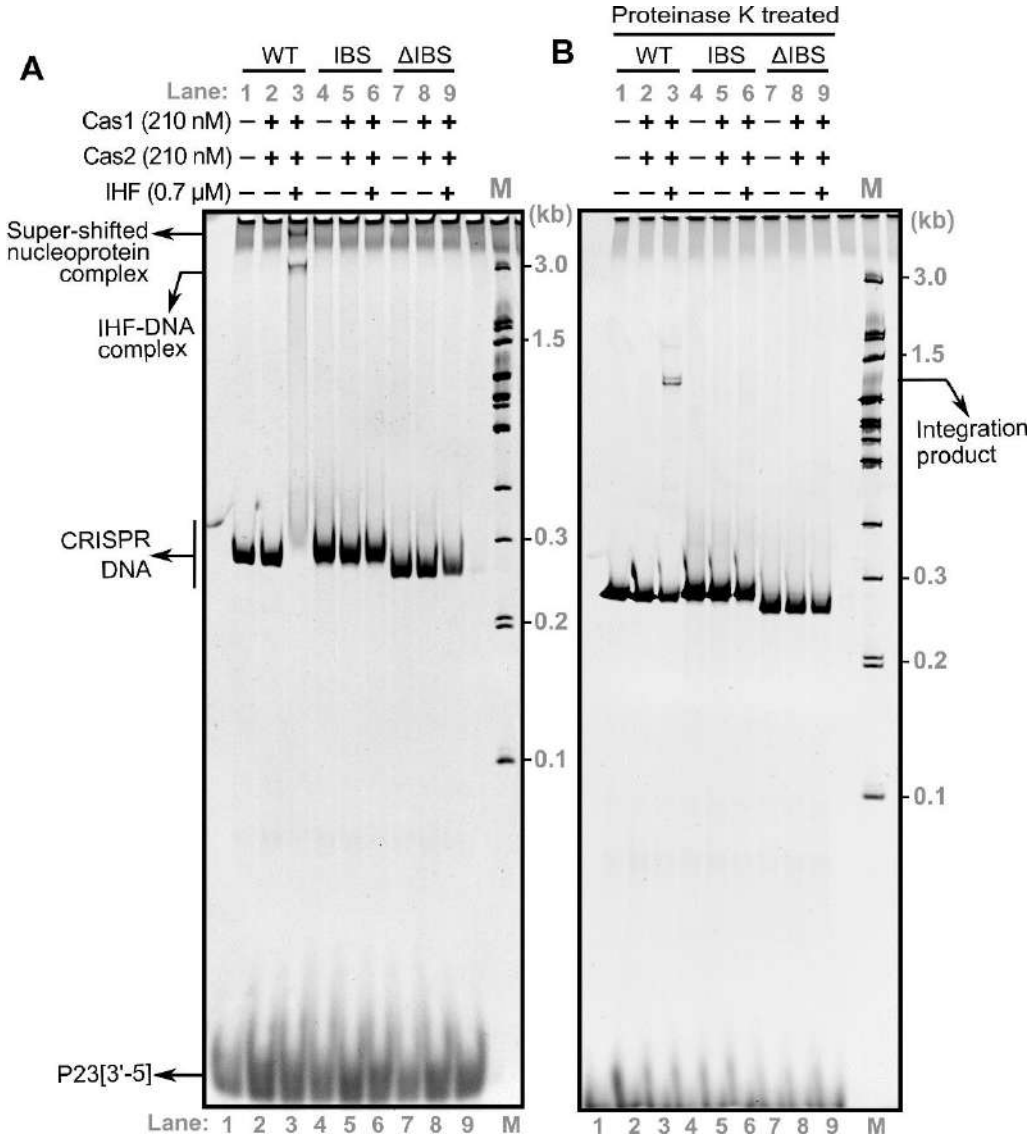


Figure 4.8: Disruption in IHF binding abrogates the prespacer integration

(A) Native gel displaying the integration assay performed with WT (Lanes 1-3) and the substrates harbouring IHF binding site disrupting mutations (IBS, Lanes 4-6) and deletion (ΔIBS, Lanes 7-9). Absence (-) or presence (+) of each reaction component is indicated on top of each lane. Positions of super-shifted nucleoprotein complex, IHF-DNA complex, CRISPR DNA and P23[3'-5] are indicated on the left. Positions corresponding to DNA marker (M) are shown on the right.

(B) Native gel depicting Proteinase K treated samples from the assay in (A). The position corresponding to the integration product is shown on the right.

Cas1-2 is known to mediate prespacer ligation via a nucleophilic attack by the hydroxyl group of the prespacer 3'-end (Nunez et al., 2015b; Rollie et al., 2015). Further, since it was reported that half-site integration intermediate is selectively excised by the Cas1-2 complex (Rollie et al., 2015) (Figure 4.9A), we reasoned that this could serve as a diagnosis for the existence of half-site integration intermediate. Therefore, we purified the reaction mixture containing the half-site integration intermediate and monitored disintegration in the presence of Cas1-2 complex. Indeed, we observed that the presence of Cas1-2 complex led to the drastic reduction of the integrated product and increment in the intensity of the band corresponding to the size of CRISPR DNA (compare Lane 1 with 2 in Figure 4.9B). Interestingly, the disintegration activity of Cas1-2 complex is significantly inhibited in the presence of IHF (Lane 3 in Figure 4.9B).

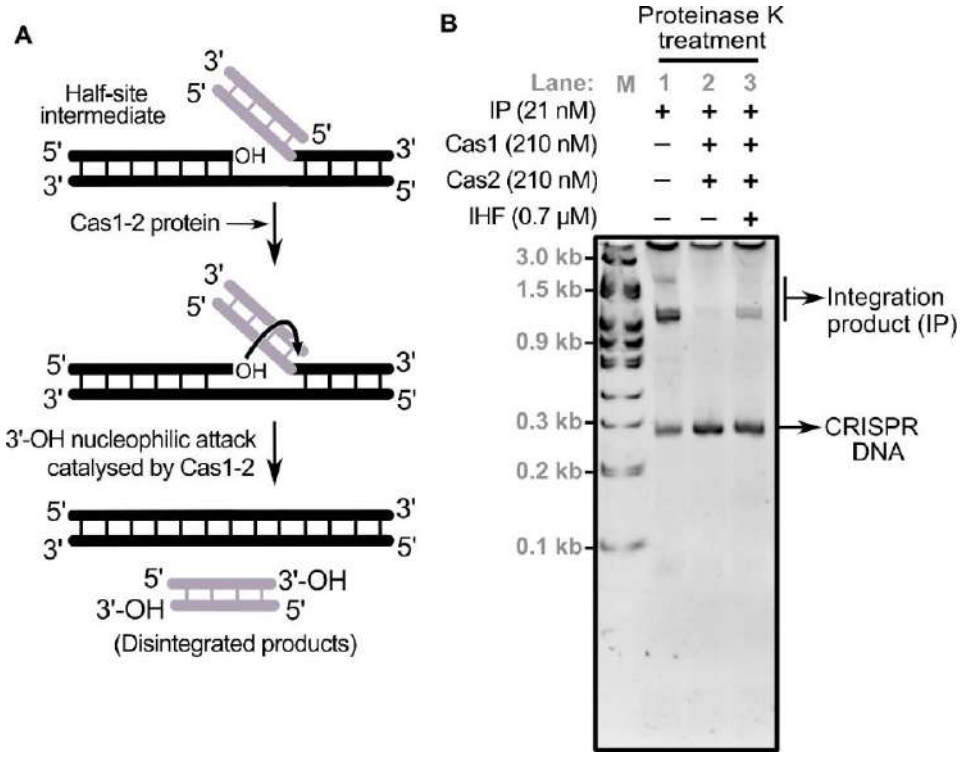


Figure 4.9: Cas1-2 disintegrates half-site intermediate products

(A) Schema highlighting the mechanism of disintegration reaction by Cas1-2. Upon interaction with half-site integration intermediate, Cas1-2 catalyse the nucleophilic attack by free 3'–OH group and results in the disintegration of DNA intermediate.
(B) Native PAGE depicting the disintegration of half-site integration intermediate. Presence (+) or absence (-) of purified WT CRISPR integration product (IP), Cas1, Cas2 and IHF is shown on top of each lane. Positions corresponding to integrated product (IP) and CRISPR DNA are indicated on the right. DNA marker (M) positions are labelled on the left.

4.4. Discussion

The involvement of IHF in the CRISPR-Cas immune response is astounding. This relationship elucidates how host factors can associate with a mobile genetic element (MGE) such as CRISPR-Cas to generate a sophisticated pathway to counter other MGEs such as phages and plasmids. IHF is known to recognise its binding region and induce sharp DNA bends thereby facilitating site-specific recombination and DNA transposition (Chalmers et al., 1998; Leong et al., 1985; Moitoso de Vargas et al., 1989; Pribil and Haniford, 2003). The molecular mechanism of λ phage lysogeny is one such classic example. The λ phage DNA contains a distantly stationed λ integrase attachment site and a low-affinity core site (cleavage site). IHF mediated bending of DNA positions these regions into proximity and facilitates cleavage at the core site, thus ensuing the integration of bacteriophage λ into the *E. coli* genome (Moitoso de Vargas et al., 1989; Segall and Nash, 1996). Similar to the interaction of λ DNA and IHF, our experiments with FRET-based DNA probes revealed that IHF deforms linear CRISPR DNA (Figure 4.4).

A previous study demonstrated that supercoiled plasmids with a CRISPR locus could act as *in vitro* substrates for spacer homing by Cas1-2, whereas no such homing reaction was seen when a linearised CRISPR DNA was employed (Nunez et al., 2015b). In comparison to linear DNA, plasmids are inherently compact and bent. Preference of such supercoiled substrates for spacer integration highlights the importance of the DNA architectural rearrangements during CRISPR adaptation. Our experiments demonstrated that supplementing IHF in an integration reaction mixture prompts prespacer integration by Cas1-2 into linear CRISPR DNA (Figure 4.7). This indicated that IHF might facilitate favourable conformation of CRISPR DNA for spacer integration. The appearance of super-shifted nucleoprotein complex in samples that contained CRISPR DNA, IHF, Cas1-2 and prespacer but not in the ones that lack IHF or IHF binding site, strengthen this proposition (compare Lane 11 with 12 in Figure 4.7B; compare Lane 3 with 6 and 9 in Figure 4.8A; Figure 4.10). In addition, reduced disintegration of half-site intermediate product in the presence of IHF indicates that the DNA bending induced by IHF appears to stabilise the intermediate by modulating the integrase/excisionase activity of Cas1-2 complex (compare Lane 2 with 3 in Figure 4.9). This shows semblance to how IHF along with λ integrase promotes integration over excision (Moitoso de Vargas et al., 1989; Segall and Nash, 1996).

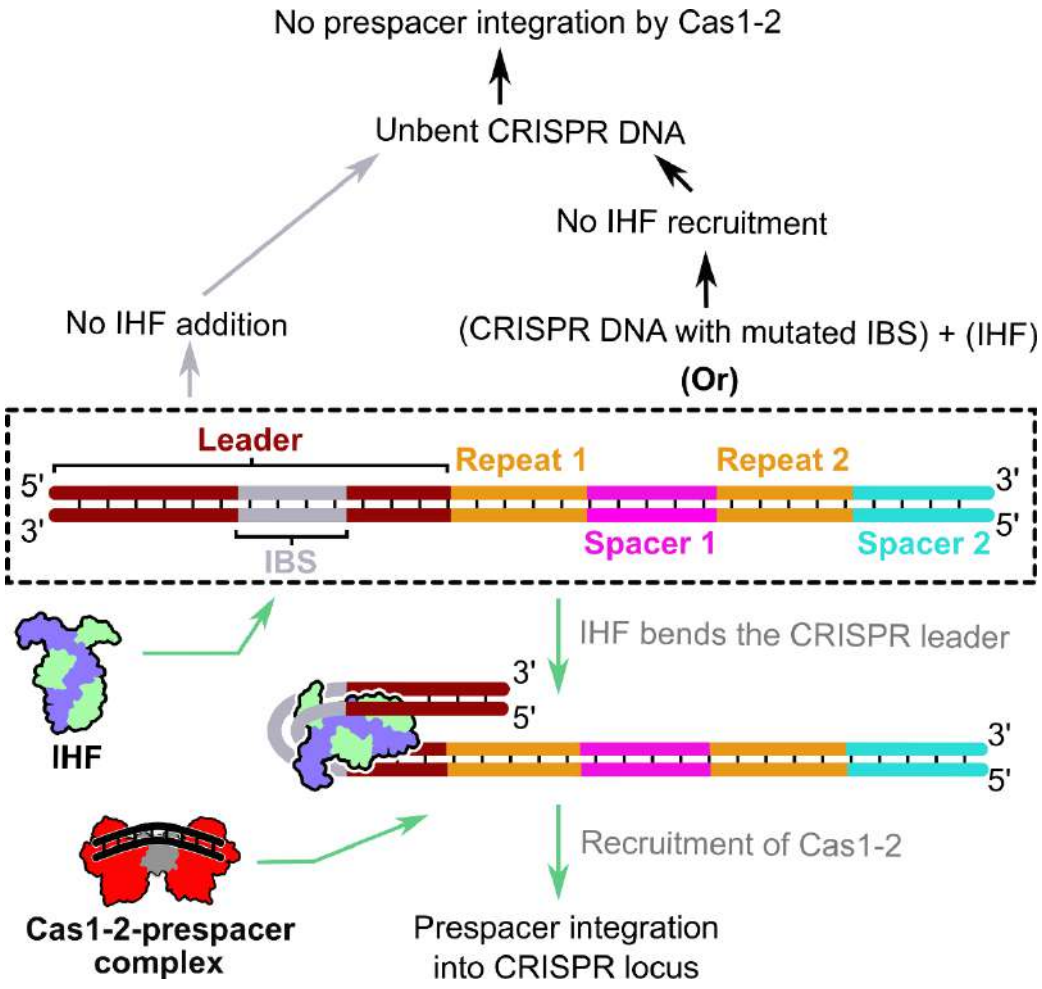


Figure 4.10: IHF interaction at the CRISPR leader prompts prespacer homing by Cas1-2 integrase

During CRISPR adaptation, IHF recognises and binds with CRISPR leader at IBS (IHF binding site). IHF-IBS interactions stimulate bending of CRISPR DNA by 120°, thereby generating a favourable conformation for the integration of prespacers by Cas1-2. Whereas, the absence of IHF or the disruption of IBS do not yield remodelled CRISPR leader and evades the prespacer integration.

Despite unearthing the critical role played by IHF in spacer acquisition, we were yet to comprehend how the 120° DNA bend caused by IHF promoted directional prespacer integration. In the following chapter, we explored the downstream events that led to prespacer integration upon IHF mediated restructuring of CRISPR locus.

4.5. Summary

In this chapter, we identified a region in the CRISPR leader (upstream of the first repeat) that encompasses an IHF binding consensus sequence. We found this region to be critical for CRISPR adaptation by performing spacer acquisition assays in strains that contained mutations in the predicted IHF binding site. Motivated by these observations, we purified IHF and studied its interactions with CRISPR DNA. Through EMSA and FRET-based DNA bending assay, we recognised that the IHF not only binds the CRISPR leader but induces a sharp bend in the DNA. Further, we performed circular permutation gel retardation assay and identified that the IHF induced DNA restructuring bends the CRISPR leader by ~120°. To ascertain the role of this architectural modulation during CRISPR adaptation, we established an assay to monitor spacer integration into linear CRISPR DNA *in vitro*. By this, we understood that the IHF mediated CRISPR restructuring premeditates the prespacer integration by Cas1-2.

Chapter V

Functional insights into the mechanism of directional prespacer integration

5. Chapter V

5.1. Introduction

The interaction of IHF with IBS on the CRISPR leader generates a favourable conformation for prespacer integration (Chapter IV). A close look into this mechanism illustrates that the extreme distortion introduced by IHF within the leader results in the juxtapositioning of the IBS upstream region in proximity to the spacer integration site (i.e., leader-repeat junction) (Figure 4.10). This observation prompted us to question if the DNA region upstream to IBS directs prespacer integration or whether the bending induced by IHF in the CRISPR leader confers physical and structural stability for recruitment of adaptation complex at the integration site. The CRISPR leader encompasses a promoter sequence that initiates transcription of the repeat-spacer array (Hille et al., 2018; Pougach et al., 2010; Pul et al., 2010). In *E. coli*, (-10)-TATA box corresponding to transcription start site is present upstream to IBS (-61 to -66 nt upstream of the first repeat) (Pougach et al., 2010). However, a previous study had demonstrated that the 60 bp region upstream of the first repeat is sufficient for spacer integration in *E. coli* (Yosef et al., 2012). This finding negated the involvement of promoter sequence for CRISPR adaptation. While analysing the CRISPR leader DNA sequence of type I-E candidates, we identified a highly conserved region upstream of the IBS (Figure 4.6). Intrigued by this observation, we went on to characterise the role of this conserved region. In the current chapter, *in vivo* spacer integration assays were utilised to prove that the identified conserved motif is critical for spacer acquisition. Further, with *in vitro* integration assays, we determined that the interaction of Cas1-2-prespacer complex with CRISPR DNA is contingent upon the IHF induced bending and the presence of an IBS upstream conserved motif (termed as "Integrase anchoring site" (IAS)). By employing fluorescence labelled DNA probes in the integration assays, we established that the molecular interplay by CRISPR DNA, IHF and Cas1-2-prespacer complex confers fidelity to CRISPR adaptation by stimulating spacer integration at the cognate site (i.e., leader-repeat junction).

5.2. Materials and Methods

5.2.1. Construction of bacterial strains and plasmids

Descriptions of the strains, plasmids and oligonucleotides are listed in Appendix Table 1, Table 2 and Table 3, respectively. Plasmid pCSIR-T (Diez-Villasenor et al., 2013) was used as a template to amplify WT array and Cas binding site variants of the leader regions (CBS1 (-34 to -45 nt), CBS2 (-46 to -57 nt), CBS3 (-58 to -69 nt), CBS2(L) (-54 to -57 nt), CBS2(C) (-50 to -53 nt) and CBS2(R) (-46 to -49 nt)) – the nucleotide positions are from the leader-repeat junction. WT and mutant array amplicons were individually inserted in between KpnI/PstI sites in plasmid pOSIP-CT (St-Pierre et al., 2013) and subsequently integrated into Phi 21 (P21) locus of *E. coli* IYB5101 strain by clonetegration.

To generate plasmid pBend-CBS2, 81 bp complementary oligos encompassing 69 bp of CBS2 leader sequence was annealed and end filled by PCR. This DNA construct was phosphorylated using T4 polynucleotide kinase and inserted into pBend5 using HpaI site (Zwieb and Adhya, 2009).

Gibson assembly protocol was utilised for generating all the clonetegrated strains recombinant vectors and (Gibson et al., 2009). After this, the resultant constructs were verified by Sanger sequencing (Sanger et al., 1977).

5.2.2. Expression and purification of proteins

Episomally expressed Cas1, Cas2, Cas1-2 and IHF were affinity purified as per the protocol in Section 2.2.2 and Section 4.2.2.

5.2.3. Spacer acquisition assays

In vivo spacer integration assay for the strains that encompass mutation in CRISPR leader (CBS1-3, CBS2(L), CBS2(C) and CBS2(R)) was performed as described in Section

2.2.6. As the mutated CRISPR arrays were integrated into P21 attB site in *E. coli* IYB5101 genome via pOSIP-CT (Section 4.2.1), primers annealing to upstream of KpnI site and downstream of PstI site in pOSIP-CT were utilised to monitor CRISPR array expansion by PCR.

5.2.4. Electrophoretic Mobility Shift Assays

Mutant leader DNA (CBS1-3) was PCR amplified from the strain carrying the respective construct that is integrated into P21 locus. These amplicons were used to monitor the binding of IHF in EMSAs (binding assays were performed as in Section 4.2.4).

5.2.5. Estimation of bending angles by circular permutation gel retardation assay

Plasmid pBend-CBS2 was digested with HindIII and EcoRI to produce a 329 bp DNA fragment. This fragment was digested with BamHI, KpnI, SspI, EcoRV, SpeI, BglII and MluI in separate reactions. The angle of CRISPR DNA bending by IHF was predicted by calculating the mobility differences of these DNA fragments in complex with IHF (Section 4.2.6)

5.2.6. *In vitro* integration assay

WT or mutant leader DNA (CBS1, CBS2, CBS3, CBS2(L), CBS2(C) and CBS2(R)) was PCR amplified from the strain carrying the respective construct that is integrated into P21 locus. *In vitro* integration assays involving these DNA constructs were performed as described in Section 4.2.7.

5.2.7. Identification of prespacer integration site in CRISPR DNA by *in vitro* integration assays

177 bp CRISPR DNA substrate (CD-U) that encompasses 69 bp leader and two repeat-spacer units of CRISPR 2.1 locus of *E. coli* was amplified using pCSIR-T (Diez-Villasenor et al., 2013) as a template. CRISPR DNA substrates labelled with 5'- FAM at the top strand of the leader end (CD-T*) or at the bottom strand of second spacer end (CD-B*) were prepared using PCR. To generate various prespacers (P33, P23[3'-5], P23[5'-5], P23[3'-10], P63, P63mPAM and their 5'-FAM labelled variants), respective oligonucleotides (Appendix Table 3) were mixed in a buffer containing 10 mM Tris-Cl pH 8.5. These mixtures were heated to 95 °C and gradually allowed to cool to room temperature in order to facilitate the formation of duplex and partial-duplex prespacers. In the case of P33ss, a 33 nt long single-stranded oligo was used as a prespacer.

The *in vitro* integration assays were performed as previously described (Section 4.2.7) with minor modifications. Briefly, a mixture containing 210 nM of Cas1 or Cas2 or Cas1-2 and 550 nM of the desired prespacer was incubated at room temperature for 5 mins. To this mixture, 0.5 µM of IHF and 21 nM of CRISPR DNA substrate were supplemented and incubation was continued at 37 °C for 60 mins in integrase buffer (20 mM HEPES–NaOH pH 7.4, 25 mM KCl, 10 mM $MgCl_2$ and 1 mM DTT). Subsequently, the reaction mixtures were supplemented with an equal volume of stopping solution (95 % Formamide, 5 mM EDTA and 0.025 % SDS) followed by heating at 95 °C for 20 mins. These samples were loaded onto pre-heated 12 % denaturing polyacrylamide gels that were maintained at 50 °C and electrophoresed in 1X TBE. Subsequently, gels were stained with EtBr and visualised using a gel documentation system (Bio-Rad). Whereas, in the assays that involve FAM labelled CRISPR DNA or prespacers, gels were imaged without any post-staining step.

5.2.8. *In silico* analysis to tabulate the presence of IHF in type I-E and non-type I-E candidates

The lists comprising of type I-E and other type I (excluding type I-E) organisms were compiled from the previous study (Makarova et al., 2011b). Using IHFα as a query, we initiated blastp (Altschul et al., 1997) search against the genomes harbouring type I-E and other type I (non-type I-E) CRISPR systems. Hits were considered bona fide if the e-value is less than 0.005 and the alignment coverage with respect to the query is at least 60%. Based on these criteria, we estimated the distribution of IHF across the species. Since HU and IHF are structurally similar and also shares similarity at the sequence level (Swinger and Rice, 2004), we relied on the annotation to distinguish between the two. However, for those cases where the annotation for the hits was not available to differentiate between IHF and HU, we presumed the organism in question to harbour IHF if more than two hits were found in the same organism satisfying the aforementioned criteria.

5.3. Results

5.3.1. Cas1-2 complex is localised upstream of IHF binding site

60 bp leader segment adjoining the first CRISPR repeat is essential for spacer acquisition (Yosef et al., 2012). However, the IHF binding region falls within the 35 bp from the first CRISPR repeat (boxed in solid line in Figure 4.6). Given the importance of this region, we wondered what the function of the remaining 25 bp could be in the leader region. Intriguingly, we also noted high conservation of sequence upstream to that of IHF binding site (boxed in dotted line in Figure 4.6). Therefore, we randomly mutated the 36 bp leader region upstream of the IHF binding site, 12 bp at a time (CBS1-3) (Figure 5.1A) and tested whether this modified region could support acquisition *in vivo*. We observed that though CBS1 [-34 to -45] and CBS3 [-58 to -69] did not affect the spacer acquisition, surprisingly, no expansion was seen for the CBS2 [-46 to -57] (compare Lane 6 with 4 and 8 in Figure 5.1B).

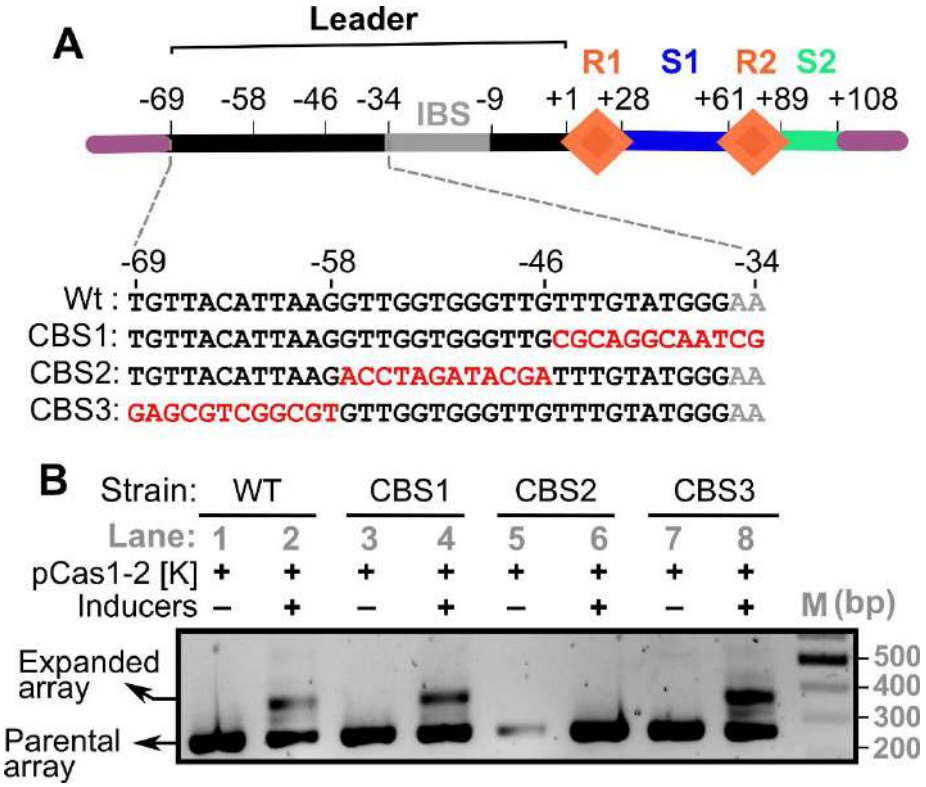


Figure 5.1: Leader region upstream of IBS influences prespacer integration

(A) Linear CRISPR DNA substrate encompassing leader (Black), two repeats (R1-R2 as Orange diamonds), 33 bp spacer1 (S1 in Blue), 19 bp spacer2 (S2 in Cyan) and IHF binding site (in Grey) is shown. The region upstream and downstream to the CRISPR locus is depicted in Purple. Numbering on top of the schema represents the position of the nucleotides with respect to the leader-repeat junction. Mutated residues of CRISPR leader (CBS1-3) are highlighted in Red.

(B) PCR products from spacer acquisition assay performed in *E. coli* harbouring the CRISPR locus integrated into the P21 attB site are shown. WT (Lanes 1-2), CBS1 (Lanes 3-4), CBS2 (Lanes 5-6) and CBS3 (Lanes 7-8) are indicated. Absence (-) or presence (+) of plasmid pCas1-2[K] and inducers is indicated on top of each lane. Positions corresponding to parental and expanded arrays are indicated on the left. DNA marker (M) positions are shown on the right.

The abrogation in CRISPR array expansion led us to assume if any of the mutated nucleotides in CBS2 disrupts the IHF-leader interactions. To further understand this, we tested the binding of IHF to mutant CRISPR DNA by EMSA. In comparison to CBS2 and 3, IHF had shown reduced binding with CBS1 (compare Figure 5.2A with B and C). We owe the impact of CBS1 mutation on IHF binding to its marginal overlap with the IBS (CBS1 in Figure 5.1A).

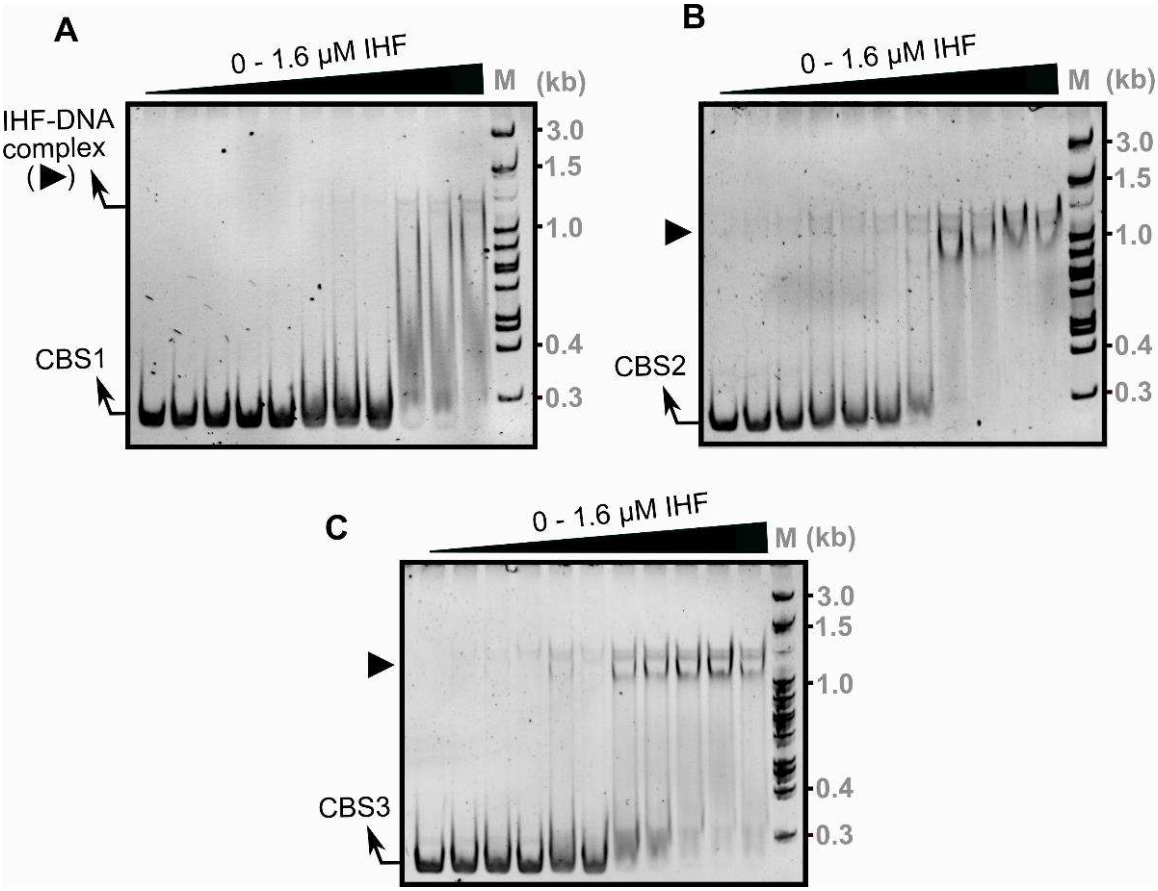


Figure 5.2: IHF interacts with CRISPR variants CBS1, CBS2 and CBS3

(A-C) Native polyacrylamide gel displaying the results of EMSA performed using IHF and CRISPR DNA variants (CBS1 (A), CBS2 (B) and CBS3 (C)). 14 nM of CRISPR DNA was incubated with increasing concentration of IHF (0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.2, 1.4 and 1.6 µM). The corresponding positions of unbound DNA and IHF-DNA complex (▶) are represented on the left. DNA marker (M) positions are shown on the right.

IHF bends the WT CRISPR leader by 122° (Figure 4.5C) and facilitates prespacer integration. Despite the binding of IHF, CBS2 did not support spacer acquisition (Figure 5.1). Hence, we performed circular permutation gel retardation assay to verify the bending of CBS2 by IHF (Figure 4.5A and Figure 5.3). Interestingly, IHF induced bending in CBS2 (118°) is in par with that of WT leader (122°) (Figure 5.3 and Figure 4.5, respectively). This suggests that the impairment of spacer acquisition due to CBS2 is not affected by IHF.

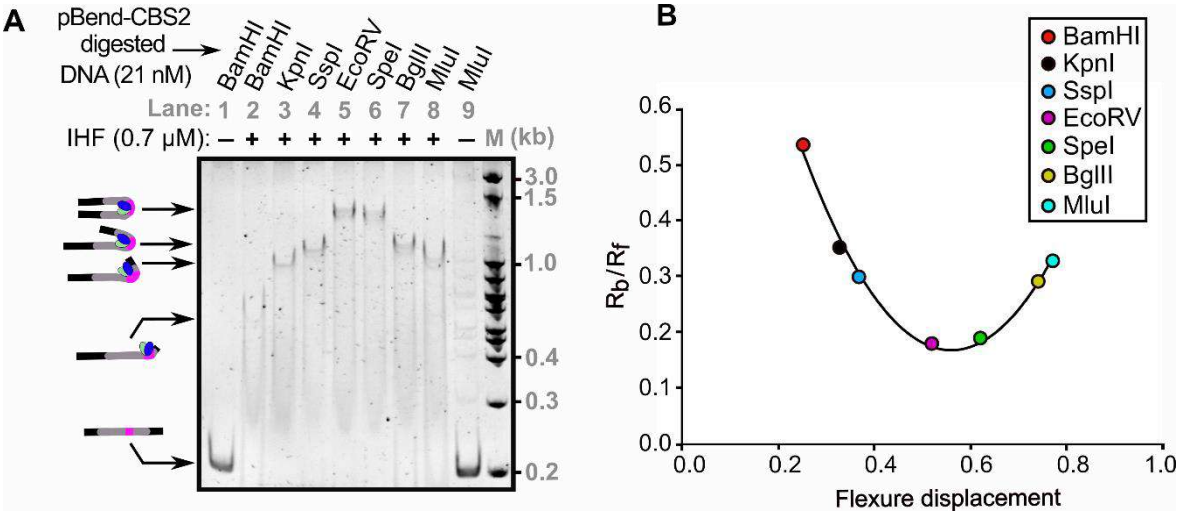


Figure 5.3: IHF bends CBS2 leader

(A) Native polyacrylamide gel displaying the results of circular permutation gel retardation assay. Enzymes labelled on each lane were used to generate the respective restriction DNA fragments from plasmid pBend-CBS2. Absence (-) or presence (+) of IHF is indicated on top of each lane. Positions of bent nucleoprotein complexes are shown on the left. DNA marker (M) positions are labelled on the right.
(B) A plot of the relative mobilities (R_b/R_f) of restricted fragments against the flexure displacement for the assay performed in (A). Positions corresponding to R_b/R_f values of the respective restriction fragments are indicated.

Since CBS2 does not impact the CRISPR leader bending by IHF, we hypothesised that CBS2 could be a binding site for Cas1-2 complex. Therefore, we conducted integration experiments involving CBS1-3. This assay showed that the super-shifted band that was seen in WT (Lane 3 in Figure 5.4A) appeared in CBS1 and CBS3 also (Lanes 6 and 12 in Figure 5.4A). Intriguingly, this band was absent when CBS2 was utilised (Lane 9 in Figure 5.4A). Moreover, IHF dependent mobility shift was prominently seen for WT and CBS2 (Lanes 3 and 9 in Figure 5.4A), albeit it was weak for CBS1 (Lane 6 in Figure 5.4A). For CBS3, the IHF dependent mobility shift was not prominent despite the presence of super-shifted band (Lane 12 in Figure 5.4A). In line with the presence of super-shifted nucleoprotein complex, all except CBS2 showed the presence of integration product (compare Lanes 3, 6 and 12 with 9 in Figure 5.4B). By recapitulating these observations, we propose that the mutations in the CBS2 region disrupt the interaction of Cas1-2-prespacer complex but not the IHF binding.

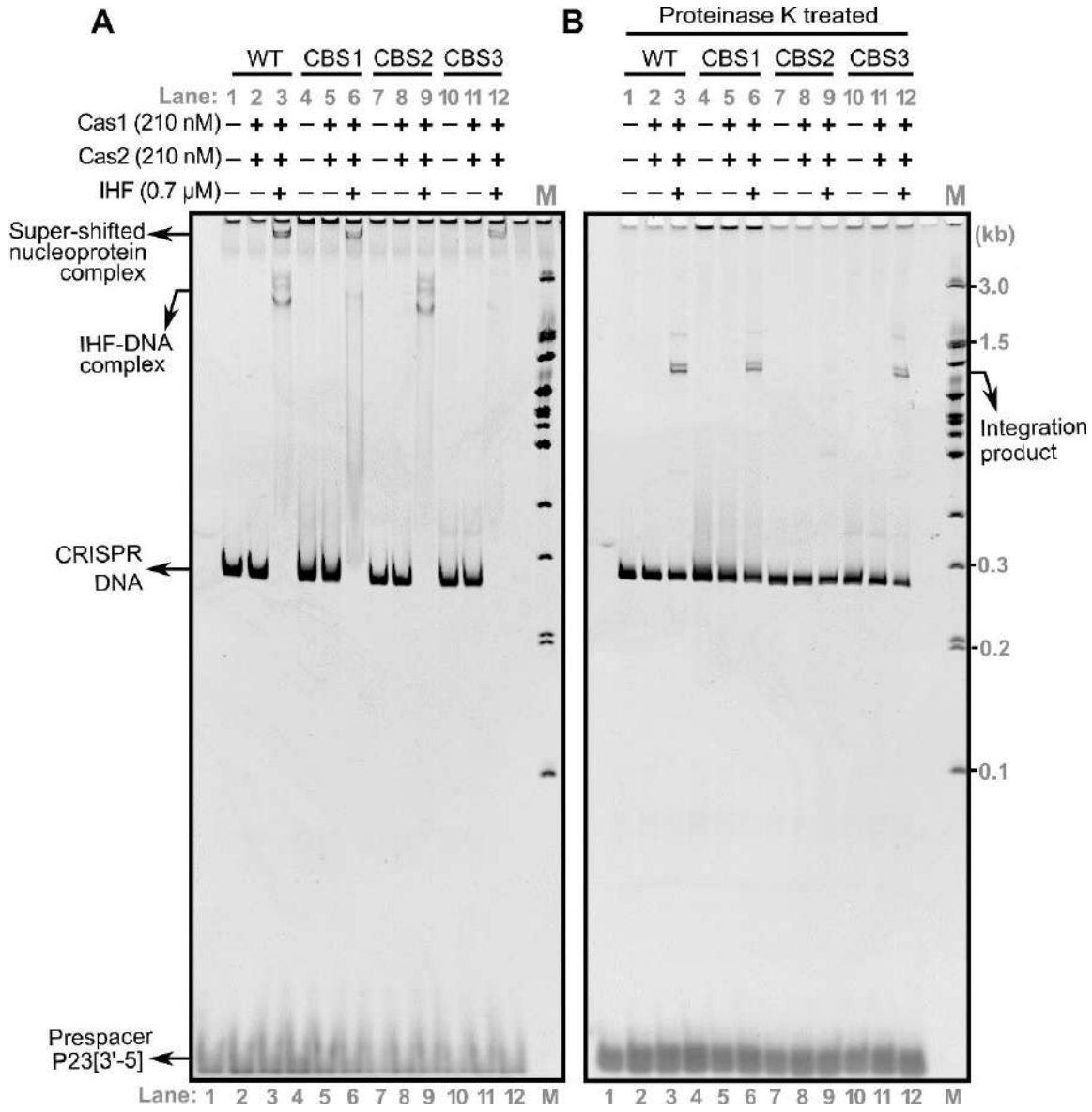


Figure 5.4: Mutations in CBS2 abolish prespacer integration

(A) Native gel displaying the integration assay performed with WT (Lanes 1-3), CBS1 (Lanes 4-6), CBS2 (Lanes 7-9) and CBS3 (Lanes 10-12) CRISPR DNA substrates. Absence (-) or presence (+) of each reaction component is indicated on top of each lane. Positions of super-shifted nucleoprotein complex, IHF-DNA complex, CRISPR DNA and prespacer P23[3'-5] are indicated on the left. Positions corresponding to DNA marker (M) are shown on the right of (B).

(B) Native gel depicting Proteinase K treated samples from the assay in (A). The position corresponding to the integration product is shown on the right.

5.3.2. Highly conserved sub-motif region within the CBS2 is crucial for prespacer integration

To probe the CBS2 further, we made three constructs, *viz.*, CBS2(L), CBS2(C) and CBS2(R). In each of these constructs, 4 bp were mutated with respect to CBS2 (Figure 5.5A). We tested each of these constructs for their ability to support prespacer acquisition *in vivo*. Remarkably, we found that except CBS2(C), other two constructs showed expansion of CRISPR array suggesting that the 4 bp in the middle of CBS2 (-50 to -53 nt) are crucial for prespacer acquisition (Figure 5.5B). To assess how these residues are impacting the prespacer acquisition, we conducted integration assays involving these constructs. In line with the acquisition assay *in vivo*, we observed that both CBS(L) and CBS(R) showed integration products in the presence of IHF and Cas1-2 complex (Lanes 9 and 15 in Figure 5.5D). Noticeably, in the case of CSB2(C), there was no super-shifted complex even in the presence of IHF and Cas1-2 complex (Lane 12 in Figure 5.5C). In line with this, the integration product was absent from the Proteinase K treated CBS2(C) sample (Lane 12 in Figure 5.5D). This observation highlighted the essentiality of CBS2(C) residues for the integration of prespacer fragment into CRISPR DNA. In tune with this, we also noted high conservation of residues corresponding to CBS2(C) in organisms harbouring type I-E system (boxed in dotted line in Figure 4.6). Taken together, the disappearance of the super-shifted band despite the presence of IHF related band in CBS2 and CBS2(C) led us to reason that the CBS2(C) is likely to harbour the binding site for Cas1-2 complex. We refer to residues (-50 to -53 nt) corresponding to CBS2(C) as integrase anchoring site (IAS).

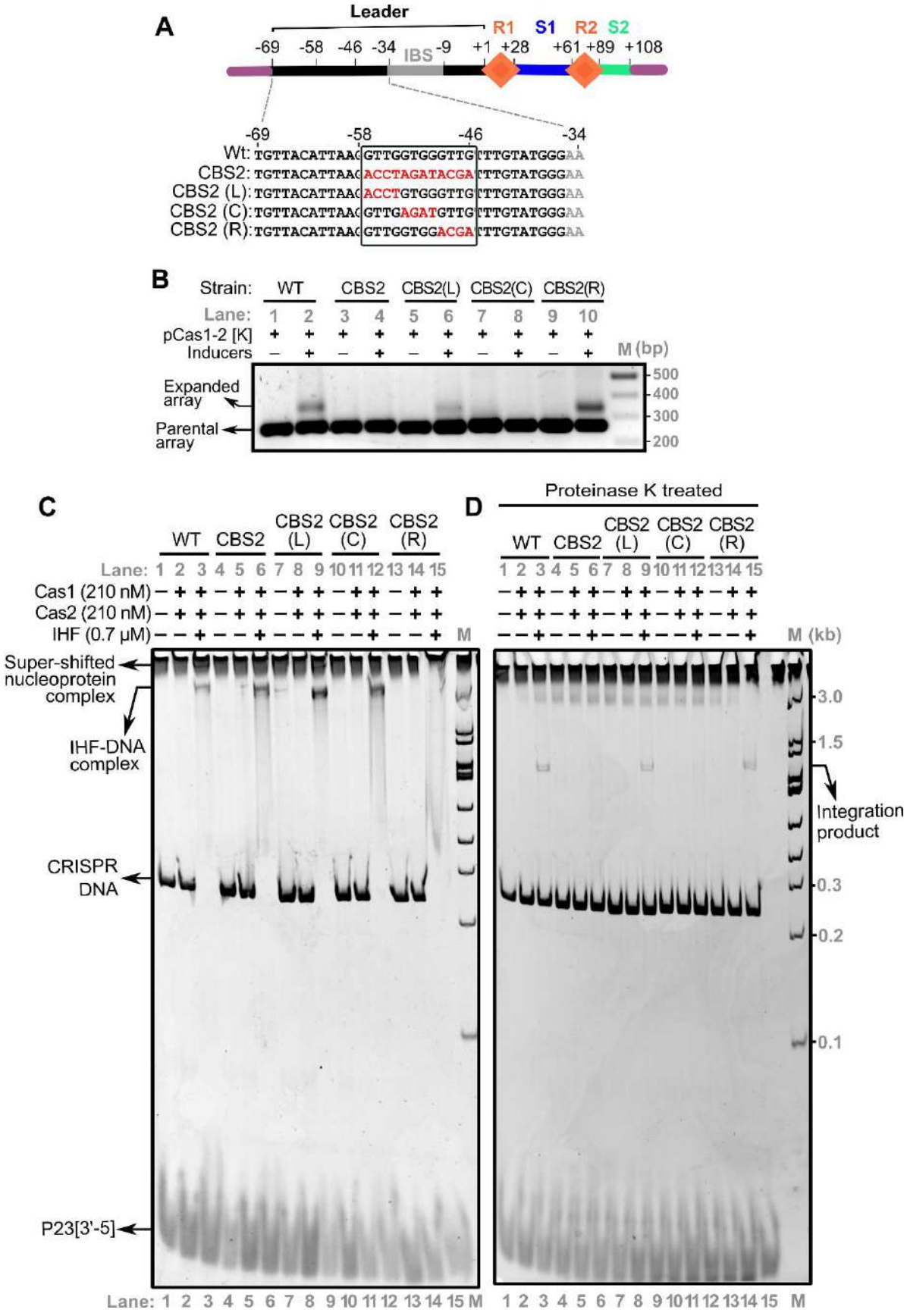


Figure 5.5: Short DNA motif within CBS2 guide prespacer incorporation

(A) Linear CRISPR DNA substrate encompassing leader (Black), two repeats (R1-R2 as Orange diamonds), 33 bp spacer1 (S1 in Blue), 19 bp spacer2 (S2 in Cyan) and IHF binding site (in Grey) is shown. The region upstream and downstream to the CRISPR locus is depicted in Purple. Numbering on top of the schema represents the position of the nucleotides with respect to the leader-repeat junction. Mutated residues of CRISPR leader (CBS2, CBS2(L), CBS2(C) and CBS2(R)) are highlighted in Red.
(B) PCR products from spacer acquisition assay performed in *E. coli* harbouring the CRISPR locus integrated into the P21 attB site are shown. WT (Lanes 1-2), CBS2 (Lanes 3-4), CBS2(L) (Lanes 5-6), CBS2(C) (Lanes 7-8) and CBS2(R) (Lanes 9-10) are indicated. Absence (-) or presence (+) of plasmid pCas1-2[K] and inducers is indicated on top of each lane. Positions corresponding to parental and expanded arrays are indicated on the left. DNA marker (M) positions are shown on the right.
(C) Native gel displaying the integration assay performed with WT (Lanes 1-3), CBS2 (Lanes 4-6), CBS2(L) (Lanes 7-9), CBS2(C) (Lanes 10-12) and CBS2(R) (Lanes 13-15) CRISPR DNA substrates. Absence (-) or presence (+) of each reaction component is indicated on top of each lane. Positions of super-shifted nucleoprotein complex, IHF-DNA complex, CRISPR DNA and prespacer P23[3'-5] are indicated on the left. Positions corresponding to DNA marker (M) are shown on the right of (D).
(D) Native gel depicting Proteinase K treated samples from the assay in (C). The position corresponding to the integration product is shown on the right.

5.3.3. Restructuring of CRISPR leader by IHF ensures polarised incorporation of prespacer into CRISPR locus by Cas1-2

A previous study had demonstrated that the Cas1-2 integrates prespacers efficiently into supercoiled DNA in the absence of IHF. Such integrations were also found to be highly non-specific and had occurred even in the absence of CRISPR sequence (Nunez et al., 2015b). Surprisingly, in linear CRISPR DNA constructs, prespacer integration by Cas1-2 resulted in the generation of half-site integration products only in the presence of IHF (Figure 4.9). Therefore, we questioned if these *in vitro* integration events stimulated by IHF were targeted towards cognate homing site (i.e., leader adjoining repeat). To determine this, we sought to ascertain the site of prespacer homing.

Generally, prespacer integration proceeds via a transesterification reaction, wherein 3'-OH of the Cas1-2 bound prespacer makes two nucleophilic attacks at the target sites to get itself ligated into the CRISPR array (Leader-1st Repeat junction (L-R1) in top strand and 1st

Repeat-1st Spacer junction (R1-S1) in bottom strand) (Figure 5.6A) (Nunez et al., 2015b; Rollie et al., 2015). Initially, we performed integration assay with unlabelled CRISPR DNA (CD-U) and prespacer P23[3'-5]. Upon resolving these samples in denaturation PAGE, we noted the appearance of four smaller ssDNA products in the sample that contained Cas1-2, prespacer, IHF and CRISPR DNA (Lane 8 in Figure 5.6B). These smaller fragments correspond to the size of the products that are resultant of nucleophilic cleavage (L: 69 nt and R': 80 nt in Figure 5.6A) and prespacer ligation (P+R: 136 nt and L'+P: 125 nt in Figure 5.6A) at the top strand of L-R1 and the bottom strand of R1-S1. Coincidentally, these products appeared only in the sample that displayed super-shifted nucleoprotein complex in native PAGE (Lane 12 in Figure 4.7B).

Further, to validate these experiments, we sought to map the site of nucleophilic attack precisely. For this, we employed the CRISPR DNA substrates that were 5'-end labelled with fluorescein (FAM) either on the top strand (CD-T* in Figure 5.6A(i)) or on the bottom strand (CD-B* in Figure 5.6A(ii)). Likewise, to monitor the prespacer ligation, we used an unlabelled CRISPR DNA (CD-U in Figure 5.6A(iii) and (iv)) and prespacers with FAM at their 5'-ends. Here, we observed that prespacer P23[3'-5] makes a nucleophilic attack at the integration sites and result in the generation of the top strand (L) and bottom strand (R') cleavage products from CD-T* and CD-B*, respectively (Lanes 3 and 6 in Figure 5.6C). Further, utilising 5'-FAM labelled P23[3'-5] and CD-U, we observed the ligation of P23[3'-5]* at these nicked sites on the top strand (P+R) and bottom strand (L'+P) (Lane 4 in Figure 5.6D). Overall, these experiments suggested that the IHF mediated structural transitions in CRISPR leader promote the integration of prespacers at the cognate target site (i.e., the top strand of L-R1 and the bottom strand of R1-S1).

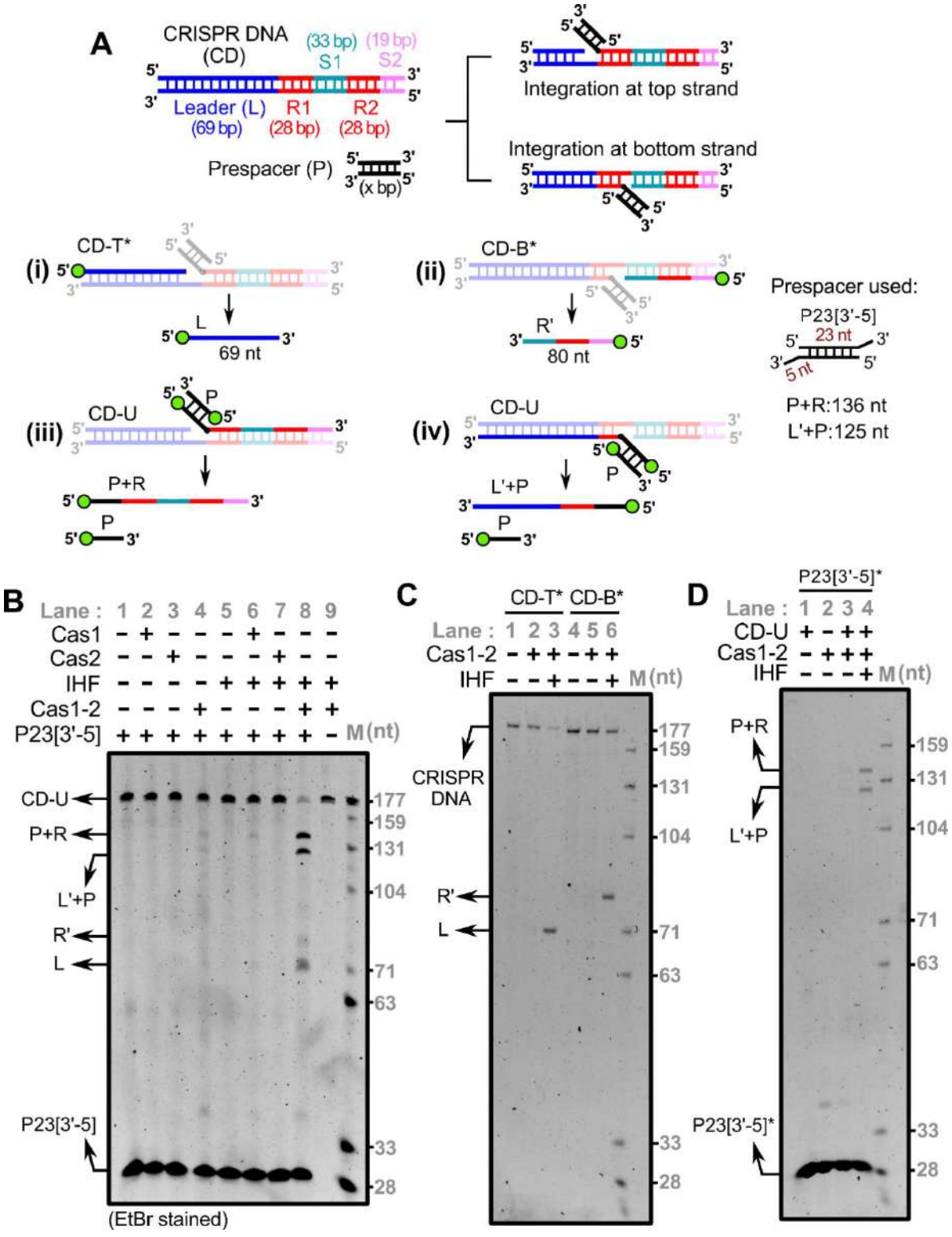


Figure 5.6: IHF interactions with CRISPR leader stimulates directional prespacer integration into CRISPR array

(A) Schema of CRISPR DNA (CD) and prespacer (P) used in Cas1-2 mediated prespacer integration assay is presented. Regions corresponding to 69 bp leader (L in Blue), 28 bp

repeats (R1-R2 in Red), 33 bp spacer 1 (S1 in Cyan) and 19 bp spacer 2 (S2 in Magenta) of the CRISPR DNA are indicated. Two integration events resulted by 3'-OH nucleophilic attack of Cas1-2 bound prespacer at the top strand (L-R1 junction) and bottom strand (R1-S1 junction) and their respective denatured DNA fragments are shown. The design of integration assays to determine the positions of nucleophilic attack (top strand (i) and bottom strand (ii)) and prespacer ligation (top strand (iii) and bottom strand (iv)) are displayed.

(B) Post-stained denaturing gel displaying the results of spacer integration assay is shown. Absence (–) or presence (+) of Cas1, Cas2, IHF, Cas1-2 and prespacer P23[3'-5] is indicated on top of each lane. Positions of bands corresponding to CRISPR DNA (CD-U), prespacer (P23[3'-5]) and the DNA fragments that are generated due to prespacer nucleophilic attack and integration (L, R', L'+P and P+R) are displayed. The DNA molecular weight marker (M) positions are shown on the right.

(C-D) Denaturing gels displaying the prespacer integration at the top strand (CD-T*: Lanes 1-3 in (C)) or at the bottom strand (CD-B*: Lanes 4-6 in (C)) or both (D). Unlabelled P23[3'-5] and 5'-FAM labelled P23[3'-5]* were used as prespacer in (C) and (D), respectively. Absence (–) or presence (+) of each reaction component is indicated on top of the respective lanes. DNA molecular weight marker (M) positions are shown on the right side. The positions of intermediate products of integration (L, R', P+R and L'+P) are displayed at the left side of the respective gels.

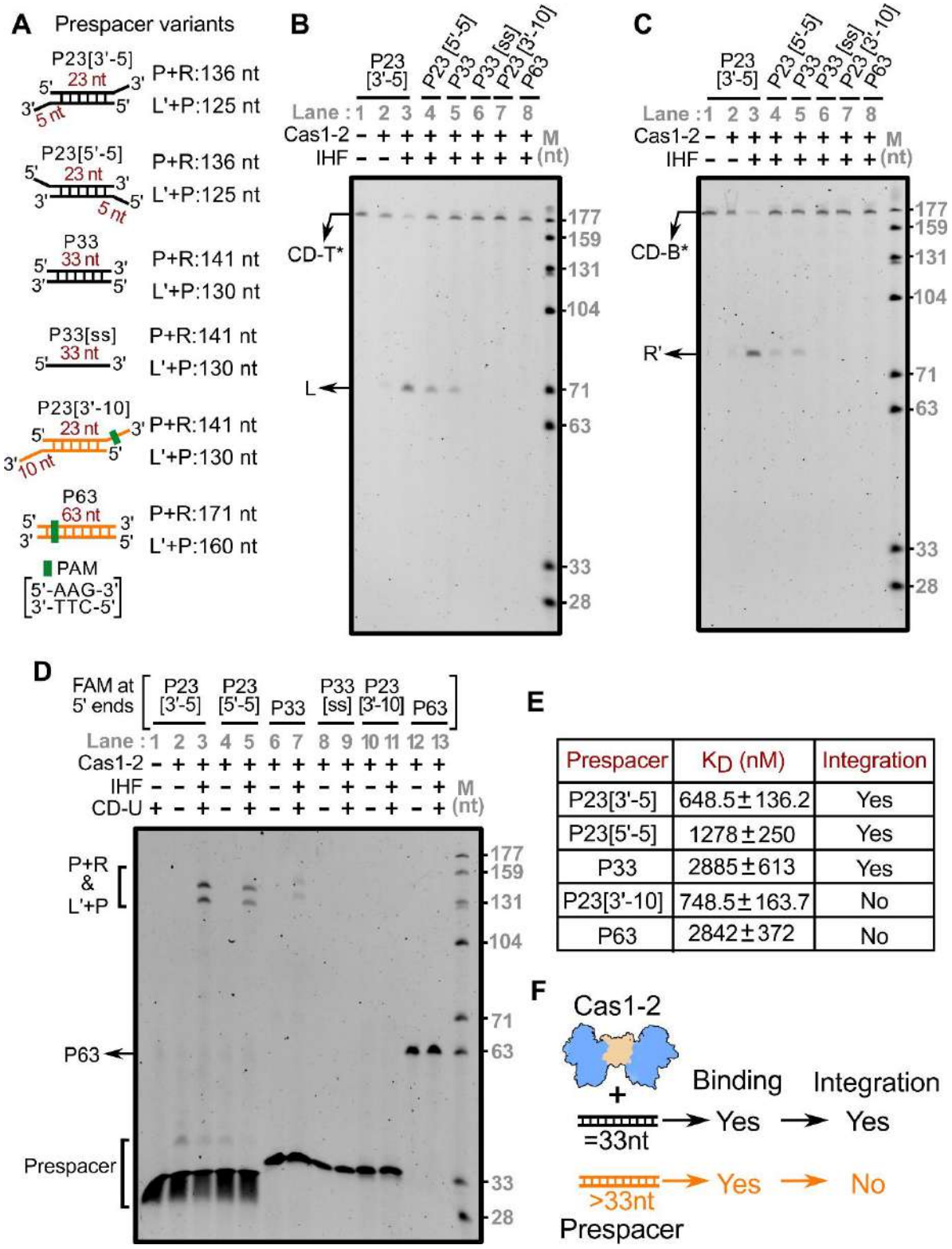
5.3.4. Length of the prespacers dictates their integration at the CRISPR array

Polarised integration of precisely sized prespacers is the characteristic feature of CRISPR adaptation. We now characterised the mechanism by which the Cas adaptation complex specifically integrates the prespacer at leader adjoining repeat and effectuate the directional expansion of CRISPR array. For this, we performed various *in vitro* integration assays, wherein we utilised prespacer substrate (P23[3'-5]) with an effective length of 33 nt (same length as the spacers in *E. coli*). But in Chapter II, we observed that the Cas1-2 binds with variably sized DNA fragments that encompass blunt ends, 3'- or 5'- overhangs (Figure 2.1). More interestingly, we also noted the generation of spacer sized DNA fragments (P63exo+) upon exonuclease digestion of longer DNA bound by Cas1-2 (Figure 2.2B). Now that we established a method to map *in vitro* prespacer integration (Figure 5.6A), we were tempted to test if the various Cas1-2 bound DNA constructs could also be integrated into the CRISPR locus in a polarised fashion. Towards this, we performed integration assays that involved various types of DNA fragments such as P33 (33 bp duplex), P33[ss] (33 nt ssDNA), P23[3'-5] (23 bp duplex with 5 nt 3'-overhangs), P23[5'-5] (23 bp duplex with 5 nt 5'-

overhangs), P23[3'-10] (23 bp duplex with 10 nt 3'-overhangs) and P63 (63 bp blunt duplex) as prespacers (Figure 5.7A).

Here, we observed that P23[3'-5] or P23[5'-5] or P33 could alone make a successful nucleophilic attack at the integration sites and result in the generation of the top strand (L) and bottom strand (R') cleavage products from CD-T* and CD-B*, respectively (Lanes 3, 4 and 5 in Figure 5.7B and C). Further, utilising 5'-FAM labelled prespacers, we observed the ligation of P23[3'-5], P23[5'-5] and P33 at these nicked sites on the top strand (P+R) and bottom strand (L'+P) (Lanes 3, 5 and 7 in Figure 5.7D).

Interestingly, we did not observe bands corresponding to the integration products when we substituted the reaction mixtures with P33[ss], P23[3'-10] and P63 (Lanes 6, 7 and 8 in Figure 5.7B and C; Lanes 9, 11 and 13 in Figure 5.7D). These findings suggest that either duplex (P33) or partial duplex (comprising of 3'-overhang (P23[3'-5]) or 5'-overhang (P23[5'-5])) prespacers with an effective length of 33 nt are strictly required during CRISPR adaptation. This bias in prespacer size preference could possibly arise due to the weakening of Cas1-2 interaction with long substrate precursors (such as P23[3'-10] and P63 with an effective length of 43 nt and 63 nt, respectively) and/or inefficient integration of such DNA fragments at the target site in CRISPR locus. To understand whether longer prespacers (>33 nt) have weak interaction with Cas1-2 thereby leading to inefficient integration, we analysed the disassociation constant values (K_D) measured from Cas1-2 and prespacer binding experiments (Figure 2.1 and Figure 5.7E). Here we observed that the affinity of Cas1-2 towards P23[3'-10] (K_D = 748.5±163.7 nM) is comparable to its affinity to P23[3'-5] (K_D = 648.5±136.2 nM). Similarly, the affinity of P63 (K_D = 2.842±0.372 µM) for Cas1-2 is on par with that of P33 (K_D = 2.885±0.613 µM). These comparisons suggest that Cas1-2 can interact with DNA fragments of varying lengths; however, the integration at the target site could be achieved only in the presence of DNA fragments with an effective length of 33 nt (Figure 5.7F).



Prespacer	K_D (nM)	Integration
P23[3'-5]	648.5±136.2	Yes
P23[5'-5]	1278±250	Yes
P33	2885±613	Yes
P23[3'-10]	748.5±163.7	No
P63	2842±372	No

Figure 5.7: Prespacer length regulates the fate of spacer integration at CRISPR locus

(A) Pictures depicting various prespacers that are employed in the integration assay. Prespacers with an overall length of 33 nt are coloured Black, whereas, those with >33 nt are represented in Orange. Sizes of the respective DNA fragments due to top strand integration (P+R) and bottom strand integration (L'+P) are indicated.

(B-D) Denaturing gels displaying the prespacer integration at the top strand (B) or at the bottom strand (C) or both (D) are shown. Absence (–) or presence (+) of each reaction component and the type of prespacer used in each sample are indicated on top of the respective lanes. DNA molecular weight marker (M) positions are shown on the right side. The positions of intermediate products of integration (L, R', P+R and L'+P) are displayed at the left side of the respective gels.

(E) The equilibrium disassociation constant values (K_D) of Cas1-2 with each type of prespacer substrate are displayed (estimated from Figure 2.1). The success (Yes) or failure (No) of integration for each prespacer is shown.

(F) Cartoon depicting the relationship between prespacer length (33 nt and >33 nt) and Cas1-2 (in Blue and Brown) mediated binding and integration. Possibility of binding and integration of each prespacer is denoted as 'Yes' or 'No'.

Previously, we also observed that the Cas1-2 complex protected DNA fragments (P63exo+) during exonuclease mediated digestion of longer DNA fragments (P63) (Figure 2.2B). As P63exo+ fragments were approximately of *E. coli* spacer size, we wondered whether they could act as potential prespacers for integration. To test this, we purified and utilised P63exo+ DNA fragments as prespacers in spacer integration assay. In line with the previous experiment (Figure 5.7), we could not observe any integration events when we employed longer prespacer P63 (Lanes 3 and 7 in Figure 5.8A). To our surprise integration was observed when P63exo+ was employed (Lane 4 and 8 in Figure 5.8A). In this case, though we monitored efficient nucleophilic attack at the top strand (L in Lane 4 of Figure 5.8A), the integration at the bottom strand seemed to be sparse (R' in Lane 8 of Figure 5.8A). By recapitulating these observations, we suggest that the Cas1-2 mediated binding of large DNA fragments guard the boundaries of integration competent prespacers from the exonucleolytic action of cellular nucleases in *E. coli* (Figure 5.8B).

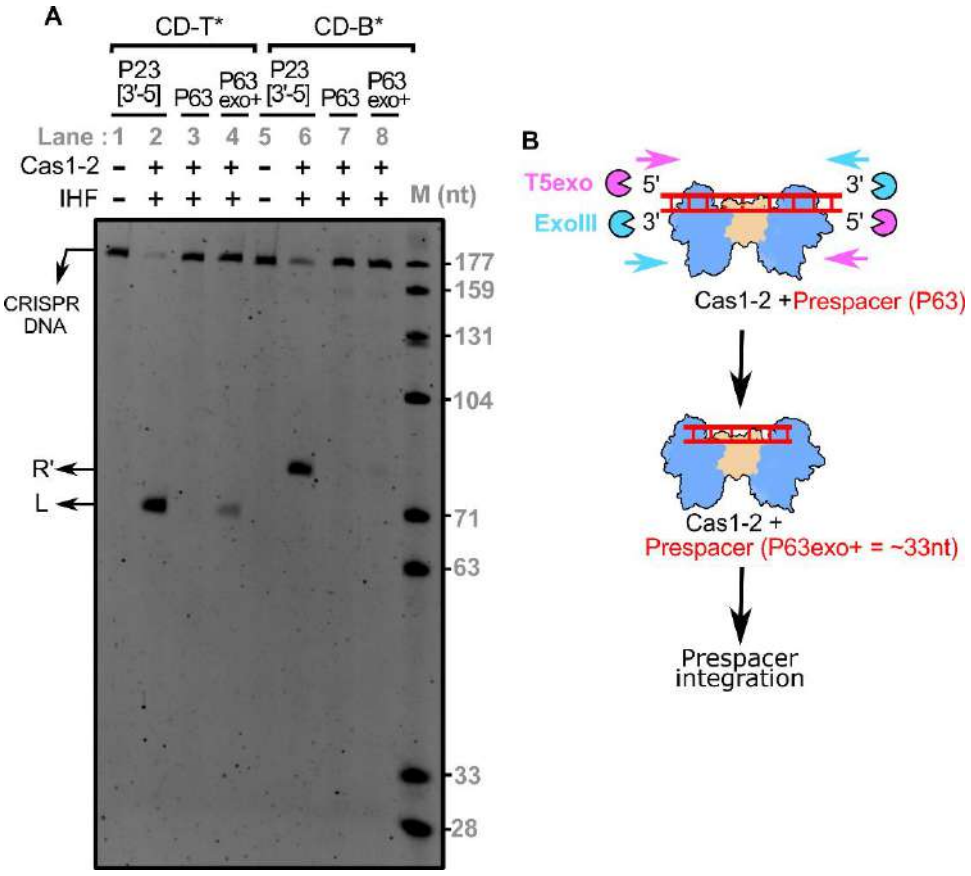


Figure 5.8: Foothold of Cas1-2 protects the boundaries of integration competent prespacers from the action of exonucleases

(A) Gel displaying the denaturing PAGE of the samples from integration reactions that employed various prespacers (P23[3'-5] (Lanes 1-2 and 5-6), P63 (Lanes 3 and 7) and Cas1-2 protected DNA fragments (P63exo+) (Lanes 4 and 8)) and CRISPR DNA substrates (CD-T* (Lanes 1-4); CD-B* (Lanes 5-8)). Presence (+) or absence (-) of each reaction component is labelled on top of each lane. Positions corresponding to labelled DNA products that are resultant of prespacer nucleophilic attack (L and R') are displayed on the left. The DNA molecular weight marker (M) positions are shown on the right.

(B) Schema illustrating the mechanism of Cas1-2 mediated protection of prespacer boundaries is displayed. Cas1-2 (in Blue and Brown), T5 exonuclease (Magenta pie), Exonuclease III (Cyan pie) and prespacer P63 (Red ladder) are portrayed.

5.4. Discussion

In this chapter, we investigated the role of IHF induced DNA bending in directional integration of the prespacer. We previously identified that the IHF bends the DNA by about 120°, which is likely to prompt reversal in the DNA direction (Figure 4.5). One possible consequence of this bending could be to bring the leader region in proximity to the first repeat. While pursuing this hypothesis, we discovered that in addition to the IHF binding site, the leader region also harbours binding site for Cas1-2 complex (referred as IAS) that is located just upstream of IBS (Figure 5.5). We also observed IAS to be highly conserved within the leader region among the type I-E organisms that harbour IHF (boxed in dotted line in Figure 4.6). This presents an attractive proposition that the IHF induced DNA bending is likely to facilitate the proximity between the Cas1-2 complex and the leader-repeat junction. The higher-order nucleoprotein complex (*vide.* Super-shifted band in Figure 4.7B) that appears in the presence of Cas1-2 complex and IHF is also noted in the case of site-specific recombination catalysed by λ integrase and IHF (Kim and Landy, 1992; Segall and Nash, 1996). However, in the absence of IHF, since CRISPR DNA is not bound by Cas1-2 complex, it is likely that IHF induced DNA bending precedes the loading of Cas1-2 complex onto the CRISPR DNA (Lanes 2 and 3 in Figure 5.4A).

Cas1 is reported to have an intrinsic specificity towards the sequences spanning the leader-repeat junction (Rollie et al., 2015). In the vast genome sequence, it is not infrequent for Cas1 to encounter such nucleotide preference and hence this is unlikely to be a principal specificity determinant. Therefore, the role of IHF could be attributed to biasing the preference of Cas1-2 complex towards shape-based recognition as exhibited by homing endonucleases (Lambert et al., 2016). In this context, it is tempting to propose that Cas1-2 complex prefers a bipartite binding site that is complemented by a part of the leader region (IAS) and leader-repeat junction. This is akin to the distantly located low-affinity core site and high-affinity attachment site in the case of λ integrase (Moitoso de Vargas et al., 1989). The proximity of these complementary sites – IAS and leader-repeat junction – mediated by the IHF induced DNA bending is aptly poised to regenerate the cognate binding site for Cas1-2 complex. The following observations appear to bolster this conjecture: First, the formation of higher-order nucleoprotein complex requires IHF induced DNA bending – akin to “intasome” in the case of bacteriophage λ integration – suggesting that the loading of Cas1-2 complex onto the

CRISPR DNA is contingent upon the proximity of the aforementioned complementary sites. Therefore, in the absence of such proximity-induced regeneration of the cognate binding site, Cas1-2 complex is unlikely to facilitate the prespacer integration into the leader proximal end. Second, in line with the above, we could observe IHF binding onto linear CRISPR DNA in the absence of Cas1-2 complex and not *vice versa*. Given this, it is possible to reiterate that Cas1-2 complex loading onto the CRISPR DNA is governed by the IHF mediated regeneration of the distantly located bipartite binding site. A later study by the Jennifer Doudna's research team had revealed the structural features of holo-adaptation complex (Cas1-2, prespacer, IHF and CRISPR DNA) by cryo-electron microscopy (Wright et al., 2017). In line with our findings, this structure had also highlighted the indispensability of IHF mediated CRISPR leader restructuring to facilitate the juxtapositioning the IAS at the leader-repeat junction to recruit Cas1-2-prespacer integrase complex (Figure 5.9).

Upon recruitment of adaptation complex at the CRISPR locus, Cas1-2 catalyse the prespacer integration via 3'-OH nucleophilic attack at the target site (Nunez et al., 2015b; Rollie et al., 2015). Based on integration assays, we deciphered that the homing of prespacer occurs into the top strand and bottom strand of CRISPR DNA at L-R1 junction and R1-S1 junction, respectively (Figure 5.6). Despite the presence of a second repeat, the observed integration events were channelised towards the designated target sites as like in the case of *in vivo* spacer integration. These observations allow us to infer that the IHF guided CRISPR DNA bending prompt polarised prespacer homing by stationing the adaptation complex at the target site.

Site-specific integration of the prespacer and concurrent duplication of the first repeat generates a functional repeat-spacer unit and maintains the integrity of CRISPR array during spacer acquisition (Goren et al., 2012; Yosef et al., 2012). In addition to the sequences bordering the leader-repeat junction, modification of the repeat sequences or structure *in vivo* is also reported to inhibit the prespacer integration (Arslan et al., 2014; Goren et al., 2016; Wang et al., 2016). In particular, the inverted repeat elements present in the CRISPR repeat acts as the molecular rulers and circumscribe the site of nucleophilic attack at a fixed distance (Goren et al., 2012). Moreover, various studies had revealed that the structural framework of Cas adaptation complex acts as another molecular ruler and adjudge the length of prespacer (Nunez et al., 2015a; Wang et al., 2015; Wright et al., 2019; Xiao et al., 2017b). Unlike in the various type I systems (I-A, I-B, I-C and I-U), Cas1-2 alone is sufficient to recognise the PAM

on the longer DNA in type I-E (*E. coli*) (Chapter II). Such PAM directed binding of Cas1-2 was shown to protect ~33 nt sized DNA fragments (P63exo+) from the exonuclease action (Figure 2.2). Utilising *in vitro* integration assays, we observed the integration of P63exo+ fragments into the CRISPR locus at the target site (Figure 5.8). Whereas, no integration was seen in the reactions that contained prespacers of length greater than 33 bp (Figure 5.7). Overall, these results reiterate the importance of prespacer length during spacer acquisition.

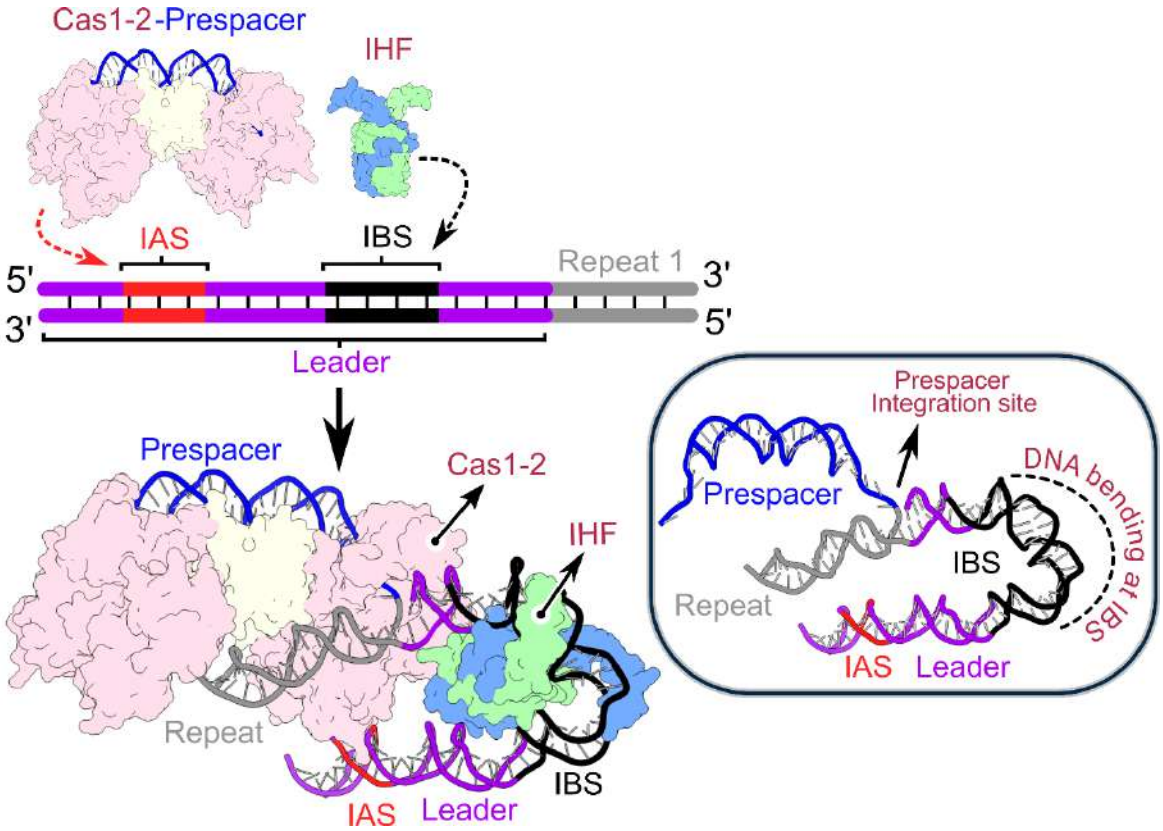


Figure 5.9: Structural features of Cas1-2-Prespacer-IHF-CRISPR DNA holo-complex

Picture displaying Cryo-electron micrograph of the Cas1-2-Prespacer-IHF-CRISPR DNA holo-complex (PDB ID: 5WFE). IHF (in Blue and Green) mediated restructuring of CRISPR leader (in Purple) at the IBS (in Black) induces a sharp bend and juxtaposes the IAS (in Red) at the leader-repeat junction (compare position of IAS in the schema of linear CRISPR DNA on the top and in the inset that displays the bent DNA conformation of the holo-complex on the right). These structural changes allow the recruitment of Cas1-2-prespacer complex (in Pink and Tan) at the leader-repeat junction and effectuate the nucleophilic attack at the target site by the prespacer (inset on the right).

While type I-E system requires accessory factor for prespacer acquisition, it was shown *in vitro* that type II-A system exhibits robust polarised prespacer incorporation into linear CRISPR DNA in the absence of any host factor (Wright and Doudna, 2016). Further, another study showed that substitution or deletion of leader region (-1 to -5 from repeat) bordering leader-repeat junction (termed as leader-anchoring sequence or LAS) in *Streptococcus pyogenes* (type II-A) induces ectopic spacer incorporation at 5th repeat where the sequence derived from 4th spacer acts as LAS (McGinn and Marraffini, 2016a). In *Sulfolobus solfataricus* (type I-A), it was observed that CRISPR locus E alone exhibits ectopic spacer incorporation, whereas polarised acquisition was observed in loci C and D (Erdmann and Garrett, 2012). CRISPR locus E encompasses a deletion of -47 to -70 in the leader region (Erdmann and Garrett, 2012; Garrett et al., 2015), which could possibly disrupt the accessory factor/Cas1-2 binding site. This, in turn, may impair bipartite site formation and since ssoCas1 is shown to have intrinsic sequence specificity (Rollie et al., 2015), it could favour integration at a region that closely resembles that of leader-repeat junction thus tuning it towards ectopic acquisition. These studies lend credence to our hypothesis that the distance between IAS and leader-repeat junction (bipartite site for Cas1-2 binding) governs the requirement of accessory factor(s) for prespacer incorporation.

Based on our data and previous reports (Arslan et al., 2014; Erdmann and Garrett, 2012; McGinn and Marraffini, 2016a; Nunez et al., 2016; Nunez et al., 2015b; Rollie et al., 2015; Wright and Doudna, 2016; Yosef et al., 2012), we present an updated model for CRISPR adaptation (Figure 5.10). This model can be dichotomised based on the proximity between IAS and leader-repeat junction, which allows us to predict the requirement of accessory factor(s). In cases where IAS and leader-repeat junction are segregated, in order to bring them into proximity for Cas1-2 binding, accessory factor(s) may be required. As exemplified by type I-E, this role is adopted by IHF in *E. coli*. IHF binding to the leader region of the CRISPR locus (IBS) leads to DNA bending. This deformed conformation ensue proximity of the distantly located IAS and leader-repeat junction that leads to the regeneration of the cognate binding site for the Cas1-2 integrase complex. Subsequently, this allows the Cas1-2 complex to orient suitably for nucleophilic attacks at the target site and promotes the prespacer homing. We analysed the distribution of IHF in organisms possessing type I CRISPR systems (type I-E and non-type I-E systems). Out of 76 organisms encompassing type I-E CRISPR system, we found that only 39 of them possess IHF (about 51%) and its distribution is predominant among enteric bacteria (Appendix Table 5). In the case of non-

type I-E, 72 out of 242 organisms (about 30%) carry IHF (Appendix Table 6). Similarly, wherever IBS is conserved in type I-E systems, we also noted a strong correlation for the existence of IAS, suggesting that these two sites co-evolve to preserve the CRISPR adaptation active (Figure 4.6). However, since several organisms that harbour type I-E system in our analysis (49%) lack IHF, it is possible to envisage the participation of other DNA architectural proteins such as HU or other Cas proteins to facilitate prespacer integration (Dillon and Dorman, 2010; Wei and Terns, 2016). On the contrary, if the IAS and leader-repeat junction lie juxtaposed as observed in type II-A system (McGinn and Marraffini, 2016a; Wei et al., 2015a; Wright and Doudna, 2016), the requirement of accessory factor(s) may be precluded (Figure 5.10). Nevertheless, co-opting the host proteins during adaptation epitomises just the tip of the iceberg of the functional diversity embodied in the CRISPR-Cas system.

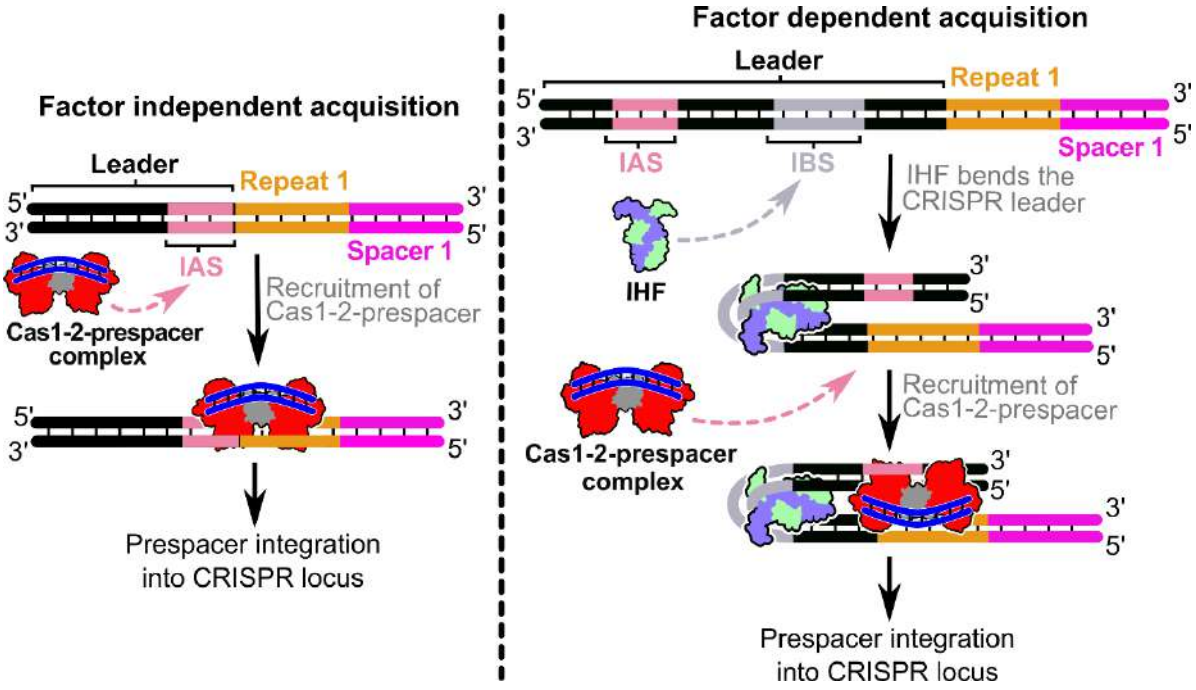


Figure 5.10: Model depicting the factor dependent and independent prespacer integration into the CRISPR locus

Based on the proximity between IAS and leader-repeat junction, the requirement of accessory factor(s) may be predicted. In type I-E, where the IAS and leader-repeat junction are segregated, IHF is required to bring them into proximity (shown in the right panel). In this case, IHF (in Green and Blue) binds to IBS (in Grey) within the CRISPR leader (in Black) that leads to bending of the DNA. This brings IAS (in Pink) and leader-repeat1 junction into close proximity thereby regenerating the cognate binding site for the Cas1-2 integrase complex. This enables the loading of the prespacer-bound Cas1-2 complex in a favourable conformation for the substrate integration at the target sites. On the other hand, as evidenced from type II-A, the requirement of accessory factor(s) may be precluded if IAS and leader-repeat junction lie juxtaposed (shown in the left panel).

5.5. Summary

In the current chapter, the essentiality of IBS upstream sequence during CRISPR adaptation was probed. Utilising integration assays with the CRISPR leader mutants, a key DNA motif termed 'CBS2' was identified to be critical for prespacer homing. By EMSA and circular permutation gel retardation assay, we observed that the mutations in CBS2 region supports the IHF directed DNA restructuring and allowed the bending of CRISPR leader by 118° (in par with WT CRISPR DNA). In integration assays, despite the presence of IHF-DNA complex, no other higher-ordered nucleoprotein complex was observed upon incubation of CBS2 CRISPR DNA with IHF, Cas1-2 and prespacer. This indicates that the CBS2 acts as an anchoring site for Cas1-2-prespacer complex and its disruption enfeebles the interaction of adaptation complex at the target site. Surprisingly, in the absence of IHF, Cas1-2-prespacer complex did not interact at CBS2 on the WT substrate. This infers that the placement of CBS2 in close proximity to the leader-repeat junction by IHF directed DNA bending alone could facilitate the generation of the binding site for the adaptation complex. With further experimentation, we identified that a short motif (-50 to -53 nt upstream of the first repeat) in CBS2 acts as an integrase anchoring site (IAS) and supports the binding of Cas1-2-prespacer complex. By performing integration assays with FAM labelled CRISPR DNA and prespacers, we also identified that the prespacers were integrated at the cognate homing sites on CRISPR DNA.

Utilising the established *in vitro* integration set up, we also tested the importance of prespacer sizing in CRISPR adaptation. In chapter II, we found that the Cas1-2/I-E is sufficient to select the PAM and mark the prespacer boundaries, whereas, the generic exonucleases resect the exposed ends on Cas1-2-DNA complex and result in DNA products that match the cognate spacer size. With integration assays, here we proved that the DNA fragments protected by Cas1-2 were indeed competent for integration into the CRISPR locus.

Chapter VI

Conclusion, future directions and applications

6. Conclusion, future directions and applications

The evolutionary arms race had resulted in a plethora of defence and counter-defence mechanisms in prokaryotes and MGE, respectively. The persistent tuning of these defence pathways by upgrading the critical nucleic acid and protein components via mutations results in the competitive coevolution of these prokaryotic hosts and pathogenic phages. In the course of evolution, the prokaryotic hosts acquired sophisticated defence pathways such as CRISPR-Cas to counter the ever-mutating pathogenic phages. The discovery of CRISPR-Cas to be an adaptive immunity akin to the humoral immune response in higher eukaryotes has surprised the scientific community and resulted in a colossal urge to understand the molecular mechanism of this defence pathway. A close look into the Cas gene clusters and repeat motifs of CRISPR-Cas subtypes displays a high degree of variability (Makarova et al., 2020b). This symbolises a rapid rate of host adaptation against the parasitic competitive challenges by MGE. Despite a vast disparity in the architecture of CRISPR-Cas locus, the basic stages involved in this adaptive immune response remain unchanged (*viz*., adaptation, maturation and interference) (Hille et al., 2018). The CRISPR adaptation engenders the host immunisation by acquiring the prespacers derived from the invaders. This crucial stage entails two sub-steps: I) prespacer capture and processing and II) integration of prespacers at the target site on the CRISPR locus (Figure 6.1). Upon infection by MGE, the nuclease action of DNA repair pathways (such as RecBCD) (Levy et al., 2015) or the CRISPR-Cas interference machinery (during primed acquisition) (Datsenko et al., 2012; Künne et al., 2016; Semenova et al., 2016; Shiriaeva et al., 2019) generates the raw material for the prespacers from the invader's DNA (Figure 6.1). An additional trimming step by PAM-directed nuclease activity of Cas4 orchestrates the generation of integration-competent prespacers in the majority of the type I CRISPR-Cas encompassing prokaryotes (Almendros et al., 2019; Kieper et al., 2018; Lee et al., 2019; Lee et al., 2018; Liu et al., 2017d; Rollie et al., 2018; Shiimori et al., 2018; Zhang et al., 2019b). Strikingly, in *E. coli* (type I-E), despite the lack of Cas4 (Makarova et al., 2020b), the prespacers are precisely sized and also encompass the PAM at their respective site on MGE (Shipman et al., 2016; Yosef et al., 2012). Intrigued by this, we concerted our efforts and went on to unravel the molecular mechanism by which the Cas adaptation complex of *E. coli* brings out the prespacer capture. In the present work, we revealed that the specificity of PAM resides with Cas1-2 in *E. coli*, whereas the prespacer processing is co-opted by

cellular non-Cas exonucleases, thereby offsetting the need for Cas4 (Yoganand et al., 2019) (Figure 6.1). Lending credence to our observations, contemporary studies also demonstrated that the nonspecific nucleases such as DnaQ or Exonuclease T could also efficiently process the DNA fragments captured by Cas1-2 of *E. coli* (Kim et al., 2019b; Ramachandran et al., 2019). Surprisingly, a comprehensive comparison of Cas1 sequences had revealed that the presence of a longer C-terminal tail is a lineage-specific feature of type I-E candidates (Yoganand et al., 2019). Such extended C-terminal tail of Cas1 confers the PAM specificity in *E. coli* and the absence of Cas4 coincides with the occurrence of extended Cas1 C-terminal tail in type I-E individuals. Owing to these facts, we are tempted to propose that the prespacer processing mechanism in all the type I-E candidates could be similar to that of *E. coli*.

In CRISPR-Cas, PAM selection ability during prespacer acquisition is not limited to Cas1-2 or Cas4 alone. In the type II-A system, the Cas9 effector nuclease is shown to be indispensable for spacer uptake (Heler et al., 2015; Nussenzweig et al., 2019; Wei et al., 2015b). Though Cas9 mandates the PAM specificity during prespacer selection, its nuclease activity was found to be dispensable for CRISPR adaptation. This raises intriguing questions about the molecular players involved in prespacer trimming. Interestingly, the type II-A system also lacks Cas4 (Makarova et al., 2020b). Therefore, based on our understanding of prespacer capture in type I-E systems, we present a proposition that the cellular nucleases could process the prespacers that are selected by adaptation complex in type II-A. Moreover, type II-B and II-C2 CRISPR variants contain both Cas4 and Cas9 (Makarova et al., 2020b). As these are prime players for prespacer selection in type I (Cas4) and type II-A (Cas9), it would be fascinating to understand the distribution of roles for PAM selection and prespacer trimming among Cas4 and Cas9 in type II-B and II-C2.

Upon generation of appropriately sized prespacers, Cas1-2 integrates them into the CRISPR locus. Homing of prespacers is site-specific and is always targeted towards the leader-repeat junction (Barrangou et al., 2007; McGinn and Marraffini, 2019; Yosef et al., 2012). In contrast to this, the Cas1-2 purified from *E. coli* was shown to integrate the prespacers non-specifically at all the repeats and non-CRISPR sites on the plasmid substrates (Nunez et al., 2015b). Such a discrepancy in the preference of homing sites during *in vivo* and *in vitro* spacer integration events by Cas1-2 had motivated us to envisage the involvement of host factors in guiding the site-specific integration of prespacers in *E. coli*. To address this lacuna, we set out to identify the host factors that direct CRISPR adaptation among the

thousands of other cellular proteins. As an initial step in this quest, we sought to capture and detect the proteins interacting at the CRISPR locus. Therefore, we employed CRISPR-dCas9 based immunoprecipitation as a molecular tool (Fujita and Fujii, 2013) and identified the proteins interacting at the CRISPR locus by mass spectrometry (Yoganand et al., 2017). Based on the cues from previous studies, we followed a reductionist approach and predicted and validated the IHF to be an essential accessory protein to assist the Cas1-2 integrase during spacer acquisition (Yoganand et al., 2017). This finding demonstrates the robustness of CRISPR/dCas9 based immunoprecipitation to map the interactions at the defined regions on DNA (i.e., *E. coli* CRISPR locus). Therefore, we strongly contemplate that this tool could greatly assist the scientific community in understanding the biomolecular interactions at any site-of-interest on the genomic DNA. Further, we noted an IHF binding site (IBS) in the CRISPR leader (Figure 6.1). Along with the contemporary work by Jennifer Doudna's group, we were able to establish the importance of IHF interactions at IBS (Nunez et al., 2016; Yoganand et al., 2017).

The discovery of a host protein (IHF) participation in CRISPR-Cas immunity is astounding. The interaction of IHF at IBS induces 120° bends in CRISPR leader (Figure 6.1). While probing the essentiality of this molecular phenomenon, we identified a sequence motif that is critical for Cas1-2-CRISPR interaction (integrase anchoring site (IAS)) in upstream of IBS. IHF directed deformation of CRISPR leader brings IAS close to the leader-repeat junction to generate a recruitment site for the Cas1-2-prespacer complex (Figure 6.1). Upon binding of adaptation complex, the 3'-OH end of the prespacer makes a nucleophilic attack at the leader-repeat1 junction on the top strand or the repeat1-spacer1 junction on the bottom strand to generate a half-site integration intermediate (Yoganand et al., 2017). The second nucleophilic attack by the free 3'-OH end of the prespacer produces a full-site integration product. These integration events followed by the DNA repair results in expanded CRISPR array that contains newly acquired spacer and concomitantly duplicated repeat (Figure 6.1).

Unlike in type I-E, the full-site integration in type II-A does not involve any host factor (Wright and Doudna, 2016; Xiao et al., 2017b). Surprisingly, the positioning of IAS near the leader-repeat region seems to generate a readymade recruitment site for the adaptation complex (McGinn and Marraffini, 2016b). This observation reiterates the importance of IAS placement in dictating the requirement of host factors for spacer acquisition. The non-specific integration events by Cas1-2/I-E *in vitro* (Nunez et al., 2015b) could have resulted due to the

intrinsic sequence specificity displayed by Cas1 (Rollie et al., 2015). Such events could potentially generate lethal mutations *in vivo*. In this scenario, IHF mediated regulation seems to bring in an additional shape-based target site recognition for the Cas1-2 integrase and ensures the vitality by conferring the fidelity for CRISPR adaptation. Removal of vulnerable traits and the acquisition of beneficial traits are the key to gain fitness against evolutionary challenges. CRISPR-Cas systems also face competitive pressure from anti-CRISPR systems (Davidson et al., 2020). Could the substitution of Cas4 guided PAM selection by Cas1 C-terminal tail during prespacer capture and employment of host protein IHF as a fidelity defining factor during prespacer integration be a countermeasure to tackle evolutionary pressures such as anti-CRISPR systems?

In the absence of interfering nucleases, CRISPR adaptation led to the incorporation of spacers that are even derived from the host genome (Shipman et al., 2016; Wei et al., 2015b). Such a phenomenon could lead to rampant self-targeting in the presence of effector nuclease. Owing to these lethal effects, evolution seems to play a critical role in enacting the CRISPR-Cas immune response by risk-benefit analysis. For example, the CRISPR-Cas locus in *E. coli* is repressed by nucleoid protein H-NS and can be activated by LeuO (Pul et al., 2010; Westra et al., 2010). But the molecular events that control the expression of these regulatory proteins are yet to be understood. Moreover, CRISPR-Cas loci in numerous prokaryotes are present in mobile genetic defence islands and are clustered with the genes that encode various other defence pathways such as restriction-modification systems, toxin-antitoxin systems and others (Makarova et al., 2011c). Though these observations drove the scientific community to discover new defence pathways through the "guilt-by-association" approach (Doron et al., 2018; Goldfarb et al., 2015; Ofir et al., 2018), the functional advantages provided by such confluence of various defence systems is poorly understood. Hence, we believe that the studies focussed on the understanding of CRISPR-Cas immune response elicitation and the molecular interplay between multiple defence pathways can unveil valuable yet overlooked mechanistic details of stress adaptation.

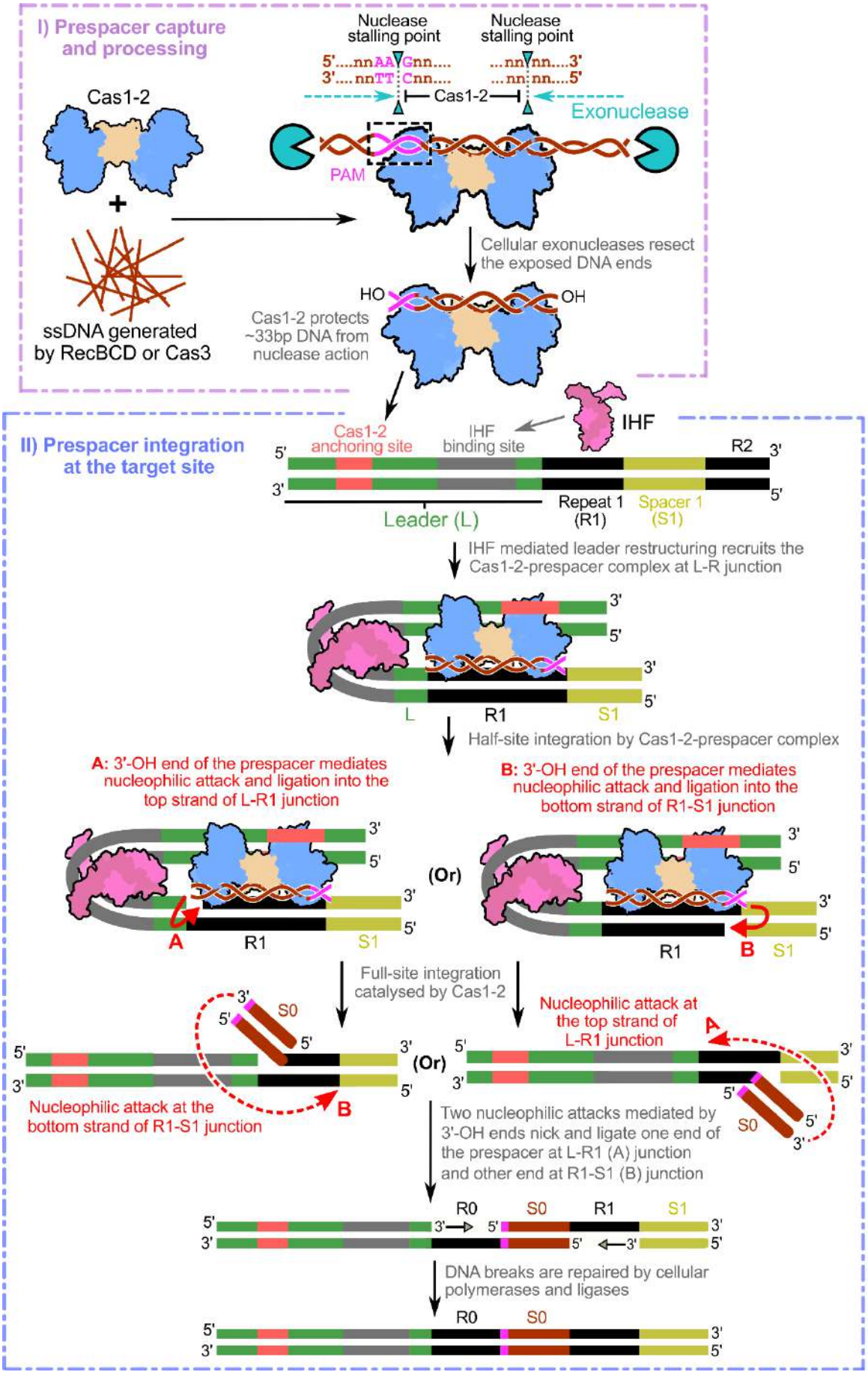


Figure 6.1: Model depicting the prespacer processing and integration in *E. coli*

Cas1-2 complex captures the DNA fragments (in Brown) that arise from the nuclease activity of RecBCD or Cas3. Owing to its intrinsic specificity, Cas1-2 is recruited at PAM (5'-AAG-3' in Magenta) on the fragmented DNA. Upon this, cellular exonucleases (in Cyan) degrade the exposed DNA ends, whereas Cas1-2 hold acts as a roadblock and stalls the exonuclease. This action protects 33 bp prespacer region starting with 'G' residue of the PAM sequence (5'-AAG-3' in Magenta). At the CRISPR locus, IHF (in Pink) mediated deformation of IHF binding site (in Grey) generates a cognate binding site for the Cas1-2-prespacer complex by juxtaposing Cas1-2 anchoring site (in Coral red) with the leader (L)-repeat1 (R1) junction. Consequent to localisation of the Cas1-2-prespacer complex at the L-R1 junction, nucleophilic attack by 3'-OH ends prompt homing of prespacer by transesterification. Here, the nucleophilic attack by the non-PAM end of the prespacer at L-R1 junction in the top strand (A) or the nucleophilic attack by the 3'-OH end of the residue derived from the PAM (in Magenta) at R1-spacer1 (S1) junction in the bottom strand (B) results in the formation of half-site integration product. Subsequent to this, the second nucleophilic attack by the free 3'-OH end of the prespacer at the R1-S1 junction (B) or L-R1 junction (A) results in the full integration of the prespacer. Following this, cellular polymerases and ligases repair the DNA lesions to generate a CRISPR locus expanded with new spacer (S0) and duplicated repeat (R0).

The current research exploration in the area of CRISPR-Cas immunity opens up the possibility of developing novel applications in a plethora of fields ranging from therapeutics to digital memory storing devices. An upcoming area of therapeutics, called phage therapy, employs virulent phages to specifically target and kill disease-causing bacteria in humans (Gordillo Altamirano and Barr, 2019). Many of these bacteria possess CRISPR-Cas immune response and can evade phage infections, thus leading to the failure of medical treatments. As in the current study, understanding the molecular events leading to CRISPR-Cas immunity paves the way for designing drug inhibitors to silence CRISPR-Cas response and helps to promote the efficacy of phage therapy.

In recent times, researchers harnessed the spacer integration ability of the CRISPR-Cas system to transform *E. coli* cells into data storage devices (Shipman et al., 2016, 2017). As the CRISPR-Cas system can collect and store short spacer DNA information, the researchers had repurposed this mechanism to store synthetic spacers that are encoded with the desired data module in a sequential fashion. Using advanced sequencing techniques, the spacer information encoded within the CRISPR locus was read in serial order and the output was obtained in the form of images and videos. In this context, the research performed here

helps to shed light on the molecular mechanism by which spacer information can be stored in sequential order within a CRISPR locus. Empowered with these mechanistic details of CRISPR memory generation, the scientific community could potentially fine-tune the DNA storage devices to achieve the utmost precision in data capture and storage.

References

Aakre, C.D., Phung, T.N., Huang, D., and Laub, M.T. (2013). A bacterial toxin inhibits DNA replication elongation through a direct interaction with the beta sliding clamp. Mol Cell *52*, P617-628.

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.

Acharya, S., Mishra, A., Paul, D., Ansari, A.H., Azhar, M., Kumar, M., Rauthan, R., Sharma, N., Aich, M., Sinha, D., *et al.* (2019). *Francisella novicida* Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. Proc Natl Acad Sci U S A *116*, 20959-20968.

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.

Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F., and Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. Bioinformatics *32*, i576-i585.

Almendros, C., Mojica, F.J., Diez-Villasenor, C., Guzman, N.M., and Garcia-Martinez, J. (2014). CRISPR-Cas functional module exchange in *Escherichia coli*. MBio *5*, e00767-13.

Almendros, C., Nobrega, F.L., McKenzie, R.E., and Brouns, S.J.J. (2019). Cas4–Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. Nucleic Acids Res *47*, 5223–5230.

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

Anantharaman, V., Makarova, K.S., Burroughs, A.M., Koonin, E.V., and Aravind, L. (2013). Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biol Direct *8*, 15.

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.

Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nat Biotechnol *38*, 824-844.

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.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol *2*, 1-11.

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.

Baker, J.R., Dong, S., and Pritchard, D.G. (2002). The hyaluronan lyase of *Streptococcus pyogenes* bacteriophage H4489A. Biochem J *365*, 317-322.

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.

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. FEMS Microbiol Lett *363*, fnw002.

Bingham, R., Ekunwe, S.I., Falk, S., Snyder, L., and Kleanthous, C. (2000). The major head protein of bacteriophage T4 binds specifically to elongation factor Tu. J Biol Chem *275*, 23219-23226.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. Science *277*, 1453-1462.

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.

Branda, S.S., Vik, S., Friedman, L., and Kolter, R. (2005). Biofilms: the matrix revisited. Trends Microbiol *13*, 20-26.

Braun-Breton, C., and Hofnung, M. (1981). *In vivo* and *in vitro* functional alterations of the bacteriophage lambda receptor in lamB missense mutants of *Escherichia coli* K-12. J Bacteriol *148*, 845-852.

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

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

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.

Castillo, F.J., and Bartell, P.F. (1974). Studies on the bacteriophage 2 receptors of *Pseudomonas aeruginosa*. J Virol *14*, 904-909.

Chalmers, R., Guhathakurta, A., Benjamin, H., and Kleckner, N. (1998). IHF modulation of Tn10 transposition: sensory transduction of supercoiling status via a proposed protein/DNA molecular spring. Cell *93*, 897-908.

Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., and Doudna, J.A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science *360*, 436-439.

Cheng, H.Y., Soo, V.W., Islam, S., McAnulty, M.J., Benedik, M.J., and Wood, T.K. (2014). Toxin GhoT of the GhoT/GhoS toxin/antitoxin system damages the cell membrane to reduce adenosine triphosphate and to reduce growth under stress. Environ Microbiol *16*, 1741-1754.

Chinenova, T.A., Mkrtumian, N.M., and Lomovskaia, N.D. (1982). Genetic characteristics of a new phage resistance trait in *Streptomyces coelicolor* A3(2). Genetika *18*, 1945-1952.

Chowdhury, S., Carter, J., Rollins, M.F., Golden, S.M., Jackson, R.N., Hoffmann, C., Nosaka, L., Bondy-Denomy, J., Maxwell, K.L., Davidson, A.R., *et al.* (2017). Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell *169*, 47-57 e11.

Chung, I.Y., Jang, H.J., Bae, H.W., and Cho, Y.H. (2014). A phage protein that inhibits the bacterial ATPase required for type IV pilus assembly. Proc Natl Acad Sci U S A *111*, 11503-11508.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

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.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A *97*, 6640-6645.

Davidson, A.R., Lu, W.T., Stanley, S.Y., Wang, J., Mejdani, M., Trost, C.N., Hicks, B.T., Lee, J., and Sontheimer, E.J. (2020). Anti-CRISPRs: Protein Inhibitors of CRISPR-Cas Systems. Annu Rev Biochem *89*, 309-332.

de Lorenzo, V., Herrero, M., Metzke, M., and Timmis, K.N. (1991). An upstream XylR- and IHF-induced nucleoprotein complex regulates the sigma 54-dependent Pu promoter of TOL plasmid. EMBO J *10*, 1159-1167.

Del Val, E., Nasser, W., Abaibou, H., and Reverchon, S. (2019). RecA and DNA recombination: a review of molecular mechanisms. Biochem Soc Trans *47*, 1511-1531.

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.

Deng, L., Garrett, R.A., Shah, S.A., Peng, X., and She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. Mol Microbiol *87*, 1088-1099.

Depardieu, F., Didier, J.P., Bernheim, A., Sherlock, A., Molina, H., Duclos, B., and Bikard, D. (2016). A Eukaryotic-like Serine/Threonine Kinase Protects *Staphylococci* against Phages. Cell Host Microbe *20*, 471-481.

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.

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.

Dillard, K.E., Brown, M.W., Johnson, N.V., Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S.D., Kim, Y., Myler, L.R., Anslyn, E.V., *et al.* (2018). Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. Cell *175*, 934-946 e15.

Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev *72*, 642-671.

Dillon, S.C., and Dorman, C.J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol *8*, 185-195.

Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D., *et al.* (2016). The crystal structure of Cpf1 in complex with CRISPR RNA. Nature *532*, 522-526.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. Science *359*, eaar4120.

Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R.W., Zimmerly, S., and Miller, J.F. (2004). Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476-481.

Drabavicius, G., Sinkunas, T., Silanskas, A., Gasiunas, G., Venclovas, Č., and Siksnys, V. (2018). DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. EMBO Rep *19*, e45543.

Dryden, D.T., Cooper, L.P., and Murray, N.E. (1993). Purification and characterization of the methyltransferase from the type 1 restriction and modification system of *Escherichia coli* K12. J Biol Chem *268*, 13228-13236.

Dryden, D.T., Murray, N.E., and Rao, D.N. (2001). Nucleoside triphosphate-dependent restriction enzymes. Nucleic Acids Res *29*, 3728-3741.

Durmaz, E., and Klaenhammer, T.R. (2007). Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. J Bacteriol *189*, 1417-1425.

Dy, R.L., Przybilski, R., Semeijn, K., Salmond, G.P., and Fineran, P.C. (2014a). A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. Nucleic Acids Res *42*, 4590-4605.

Dy, R.L., Richter, C., Salmond, G.P., and Fineran, P.C. (2014b). Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol *1*, 307-331.

East-Seletsky, A., O'Connell, M.R., Burstein, D., Knott, G.J., and Doudna, J.A. (2017). RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. Mol Cell *66*, 373-383 e3.

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.

Elmore, J.R., Sheppard, N.F., Ramia, N., Deighan, T., Li, H., Terns, R.M., and Terns, M.P. (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev *30*, 447-459.

Emond, E., Holler, B.J., Boucher, I., Vandenbergh, P.A., Vedamuthu, E.R., Kondo, J.K., and Moineau, S. (1997). Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from *Lactococcus lactis*. Appl Environ Microbiol *63*, 1274-1283.

Erdmann, S., and Garrett, R.A. (2012). Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol *85*, 1044-1056.

Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., *et al.* (2017). Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proc Natl Acad Sci U S A *114*, E5122-E5128.

Fang, X., Liu, Q., Bohrer, C., Hensel, Z., Han, W., Wang, J., and Xiao, J. (2018). Cell fate potentials and switching kinetics uncovered in a classic bistable genetic switch. Nat Commun *9*, 2787.

Faure, G., Makarova, K.S., and Koonin, E.V. (2019a). CRISPR-Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. J Mol Biol *431*, 3-20.

Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E.V. (2019b). CRISPR-Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol *17*, 513-525.

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., Daly, C., and Fitzgerald, G.F. (1999). Identification of four phage resistance plasmids from *Lactococcus lactis* subsp. cremoris HO2. Appl Environ Microbiol *65*, 1540-1547.

Forde, A., and Fitzgerald, G.F. (2003). Molecular organization of exopolysaccharide (EPS) encoding genes on the lactococcal bacteriophage adsorption blocking plasmid, pCI658. Plasmid *49*, 130-142.

Friedman, D.I. (1988). Integration host factor: a protein for all reasons. Cell *55*, 545-554.

Fujita, T., and Fujii, H. (2013). Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. Biochem Biophys Res Commun *439*, 132-136.

Gao, P., Yang, H., Rajashankar, K.R., Huang, Z., and Patel, D.J. (2016). Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. Cell Res *26*, 901-913.

Garcia-Doval, C., Schwede, F., Berk, C., Rostol, J.T., Niewoehner, O., Tejero, O., Hall, J., Marraffini, L.A., and Jinek, M. (2020). Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6. Nat Commun *11*, 1596.

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.

Garrett, R.A., Shah, S.A., Erdmann, S., Liu, G., Mousaei, M., Leon-Sobrino, C., Peng, W., Gudbergsdottir, S., Deng, L., Vestergaard, G., *et al.* (2015). CRISPR-Cas Adaptive Immune Systems of the Sulfolobales: Unravelling Their Complexity and Diversity. Life (Basel) *5*, 783-817.

Garrett, R.A., Vestergaard, G., and Shah, S.A. (2011). Archaeal CRISPR-based immune systems: exchangeable functional modules. Trends Microbiol *19*, P549-556.

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-E2586.

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.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-345.

Gilbert, Luke A., Larson, Matthew H., Morsut, L., Liu, Z., Brar, Gloria A., Torres, Sandra E., Stern-Ginossar, N., Brandman, O., Whitehead, Evan H., Doudna, Jennifer A., *et al.* (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell *154*, 442-451.

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019). PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biol *16*, 504-517.

Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., and Ferrin, T.E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci *27*, 14-25.

Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633-637.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., and Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. EMBO J *34*, 169-183.

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.

Gordeeva, J., Morozova, N., Sierro, N., Isaev, A., Sinkunas, T., Tsvetkova, K., Matlashov, M., Truncaite, L., Morgan, R.D., Ivanov, N.V., *et al.* (2019). BREX system of *Escherichia coli* distinguishes self from non-self by methylation of a specific DNA site. Nucleic Acids Res *47*, 253-265.

Gordillo Altamirano, F.L., and Barr, J.J. (2019). Phage Therapy in the Postantibiotic Era. Clin Microbiol Rev *32*, e00066-18.

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep *16*, 2811-2818.

Goren, M.G., Yosef, I., Auster, O., and Qimron, U. (2012). Experimental Definition of a Clustered Regularly Interspaced Short Palindromic Duplicon in *Escherichia coli*. J Mol Biol *423*, 14-16.

Grainy, J., Garrett, S., Graveley, B.R., and M, P.T. (2019). CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. Nucleic Acids Res *47*, 7518-7531.

Guerrero, F., Ciragan, A., and Iwaï, H. (2015). Tandem SUMO fusion vectors for improving soluble protein expression and purification. Protein Expr Purif *116*, 42-49.

Guo, M., Zhang, K., Zhu, Y., Pintilie, G.D., Guan, X., Li, S., Schmid, M.F., Ma, Z., Chiu, W., and Huang, Z. (2019). Coupling of ssRNA cleavage with DNase activity in type III-A CRISPR-Csm revealed by cryo-EM and biochemistry. Cell Res *29*, 305-312.

Guo, T.W., Bartesaghi, A., Yang, H., Falconieri, V., Rao, P., Merk, A., Eng, E.T., Raczkowski, A.M., Fox, T., Earl, L.A., *et al.* (2017). Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. Cell *171*, 414-426 e12.

Hadi, S.M., Bachi, B., Shepherd, J.C., Yuan, R., Ineichen, K., and Bickle, T.A. (1979). DNA recognition and cleavage by the EcoP15 restriction endonuclease. J Mol Biol *134*, 655-666.

Halpin-Healy, T.S., Klompe, S.E., Sternberg, S.H., and Fernandez, I.S. (2020). Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. Nature *577*, 271-274.

Hammad, A.M.M. (1998). Evaluation of alginate-encapsulated *Azotobacter chroococcum* as a phage-resistant and an effective inoculum. J Basic Microbiol *38*, 9-16.

Hanlon, G.W., Denyer, S.P., Olliff, C.J., and Ibrahim, L.J. (2001). Reduction in exopolysaccharide viscosity as an aid to bacteriophage penetration through *Pseudomonas aeruginosa* biofilms. Appl Environ Microbiol *67*, 2746-2753.

Harms, A., Brodersen, D.E., Mitarai, N., and Gerdes, K. (2018). Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. Mol Cell *70*, 768-784.

Harrington, L.B., Burstein, D., Chen, J.S., Paez-Espino, D., Ma, E., Witte, I.P., Cofsky, J.C., Kyrpides, N.C., Banfield, J.F., and Doudna, J.A. (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science *362*, 839-842.

Harrington, L.B., Ma, E., Chen, J.S., Witte, I.P., Gertz, D., Paez-Espino, D., Al-Shayeb, B., Kyrpides, N.C., Burstein, D., Banfield, J.F., *et al.* (2020). A scoutRNA Is Required for Some Type V CRISPR-Cas Systems. Mol Cell *79*, 416-424.e5.

Harvey, H., Bondy-Denomy, J., Marquis, H., Sztanko, K.M., Davidson, A.R., and Burrows, L.L. (2018). *Pseudomonas aeruginosa* defends against phages through type IV pilus glycosylation. Nat Microbiol *3*, 47-52.

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.

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., 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.

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.

Hille, F., Richter, H., Wong, S.P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. Cell *172*, 1239-1259.

Hochstrasser, M.L., and Doudna, J.A. (2015). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem Sci *40*, 58-66.

Hochstrasser, M.L., Taylor, D.W., Kornfeld, J.E., Nogales, E., and Doudna, J.A. (2016). DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. Mol Cell *63*, 840-851.

Hoikkala, V., Ravantti, J., Díez-Villaseñor, C., Tiirola, M., Conrad, R.A., McBride, M.J., and Sundberg, L.-R. (2020). Cooperation between CRISPR-Cas types enables adaptation in an RNA-targeting system. bioRxiv, 2020.02.20.957498.

Holbrook, J.A., Tsodikov, O.V., Saecker, R.M., and Record, M.T. (2001). Specific and non-specific interactions of integration host factor with DNA: thermodynamic evidence for disruption of multiple IHF surface salt-bridges coupled to DNA binding. J Mol Biol *310*, 379-401.

Hoskisson, P.A., Sumby, P., and Smith, M.C.M. (2015). The phage growth limitation system in *Streptomyces coelicolor* A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. Virology *477*, 100-109.

Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S., and Koonin, E.V. (2017). Phylogenomics of Cas4 family nucleases. BMC Evol Biol *17*, 232.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M.D., 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.

Hynes, W.L., Hancock, L., and Ferretti, J.J. (1995). Analysis of a second bacteriophage hyaluronidase gene from *Streptococcus pyogenes*: evidence for a third hyaluronidase involved in extracellular enzymatic activity. Infect Immun *63*, 3015-3020.

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). Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. Science *345*, 1473-1479.

Jackson, S.A., Birkholz, N., Malone, L.M., and Fineran, P.C. (2019). Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. Cell Host Microbe *25*, 250-260.e4.

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J. (2017). CRISPR-Cas: Adapting to change. Science *356*, eaal5056.

Janscak, P., MacWilliams, M.P., Sandmeier, U., Nagaraja, V., and Bickle, T.A. (1999). DNA translocation blockage, a general mechanism of cleavage site selection by type I restriction enzymes. EMBO J *18*, 2638-2647.

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

Jia, N., Jones, R., Yang, G., Ouerfelli, O., and Patel, D.J. (2019). CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity. Mol Cell *75*, 944-956 e6.

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

Jiang, W., Samai, P., and Marraffini, L.A. (2016b). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell *164*, 710-721.

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., 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.

Johnson, A.D., Poteete, A.R., Lauer, G., Sauer, R.T., Ackers, G.K., and Ptashne, M. (1981). λ Repressor and cro--components of an efficient molecular switch. Nature *294*, 217-223.

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.

Ka, D., Jang, D.M., Han, B.W., and Bae, E. (2018). Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. Nucleic Acids Res *46*, 9805-9815.

Kaya, E., Doxzen, K.W., Knoll, K.R., Wilson, R.C., Strutt, S.C., Kranzusch, P.J., and Doudna, J.A. (2016). A bacterial Argonaute with noncanonical guide RNA specificity. Proc Natl Acad Sci U S A *113*, 4057-4062.

Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science *357*, 605-609.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep *22*, 3377-3384.

Kieper, S.N., Almendros, C., and Brouns, S.J.J. (2019). Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. FEMS Microbiol Lett *366*, fnz129.

Kim, E., Koo, T., Park, S.W., Kim, D., Kim, K., Cho, H.Y., Song, D.W., Lee, K.J., Jung, M.H., Kim, S., *et al.* (2017). *In vivo* genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. Nat Commun *8*, 14500.

Kim, J.G., Garrett, S., Wei, Y., Graveley, B.R., and Terns, M.P. (2019a). CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. Nucleic Acids Res *47*, 8632–8648.

Kim, S., and Landy, A. (1992). Lambda Int protein bridges between higher order complexes at two distant chromosomal loci attL and attR. Science *256*, 198-203.

Kim, S., Loeff, L., Colombo, S., Brouns, S.J.J., and Joo, C. (2019b). Selective Prespacer Processing Ensures Precise CRISPR-Cas Adaptation. bioRxiv, 10.1101/608976.

Kim, T.Y., Shin, M., Huynh Thi Yen, L., and Kim, J.S. (2013). Crystal structure of Cas1 from *Archaeoglobus fulgidus* and characterization of its nucleolytic activity. Biochem Biophys Res Commun *441*, 720-725.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. Nature *571*, 219-225.

Knott, G.J., East-Seletsky, A., Cofsky, J.C., Holton, J.M., Charles, E., O'Connell, M.R., and Doudna, J.A. (2017). Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. Nat Struct Mol Biol *24*, 825-833.

Ko, C.C., and Hatfull, G.F. (2018). Mycobacteriophage Fruitloop gp52 inactivates Wag31 (DivIVA) to prevent heterotypic superinfection. Mol Microbiol *108*, 443-460.

Kondrateva, E., Demchenko, A., Lavrov, A., and Smirnikhina, S. (2020). An overview of currently available molecular Cas-tools for precise genome modification. Gene, DOI: https://doi.org/10.1016/j.gene.2020.145225.

Kornberg, S.R., Zimmerman, S.B., and Kornberg, A. (1961). Glucosylation of deoxyribonucleic acid by enzymes from bacteriophage-infected *Escherichia coli*. J Biol Chem *236*, 1487-1493.

Kulp, A., and Kuehn, M.J. (2010). Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu Rev Microbiol *64*, 163-184.

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

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell *63*, 852-864.

Kuzmenko, A., Oguienko, A., Esyunina, D., Yudin, D., Petrova, M., Kudinova, A., Maslova, O., Ninova, M., Ryazansky, S., Leach, D., *et al.* (2020). DNA targeting and interference by a bacterial Argonaute nuclease. Nature, DOI: https://doi.org/10.1038/s41586-020-2605-1.

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

Lambert, A.R., Hallinan, J.P., Shen, B.W., Chik, J.K., Bolduc, J.M., Kulshina, N., Robins, L.I., Kaiser, B.K., Jarjour, J., Havens, K., *et al.* (2016). Indirect DNA Sequence Recognition and Its Impact on Nuclease Cleavage Activity. Structure *24*, 862-873.

Lange, S.J., Alkhnbashi, O.S., Rose, D., Will, S., and Backofen, R. (2013). CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. Nucleic Acids Res *41*, 8034-8044.

Lee, H., Dhingra, Y., and Sashital, D.G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. Elife *8*, e44248.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Mol Cell *70*, 48-59.e5.

Leong, J.M., Nunes-Duby, S., Lesser, C.F., Youderian, P., Susskind, M.M., and Landy, A. (1985). The phi 80 and P22 attachment sites. Primary structure and interaction with *Escherichia coli* integration host factor. J Biol Chem *260*, 4468-4477.

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., and Craigie, R. (2005). Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. J Biol Chem *280*, 29334-29339.

Li, M., Gong, L., Zhao, D., Zhou, J., and Xiang, H. (2017). The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. Nucleic Acids Res *45*, 4642-4654.

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.

Lin, J., Chen, H., Dröge, P., and Yan, J. (2012). Physical Organization of DNA by Multiple Non-Specific DNA-Binding Modes of Integration Host Factor (IHF). PLoS One *7*, e49885.

Lindsay, J.A., Ruzin, A., Ross, H.F., Kurepina, N., and Novick, R.P. (1998). The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. Mol Microbiol *29*, 527-543.

Lisitskaya, L., Aravin, A.A., and Kulbachinskiy, A. (2018). DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. Nat Commun *9*, 5165.

Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M., and Wang, Y. (2017a). C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. Mol Cell *65*, 310-322.

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., Wang, M., Zhang, X., and Wang, Y. (2017b). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. Cell *170*, 714-726 e10.

Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G., and Wang, Y. (2017c). Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell *168*, 121-134.

Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q., and Peng, N. (2017d). Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in *Sulfolobus islandicus*. Nucleic Acids Res *45*, 8978-8992.

Liu, T.Y., Liu, J.-J., Aditham, A.J., Nogales, E., and Doudna, J.A. (2019). Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. Nat Commun *10*, 3001.

Loenen, W.A., Dryden, D.T., Raleigh, E.A., Wilson, G.G., and Murray, N.E. (2014). Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res *42*, 3-19.

Loenen, W.A., and Raleigh, E.A. (2014). The other face of restriction: modification-dependent enzymes. Nucleic Acids Res *42*, 56-69.

Lu, M.J., Stierhof, Y.D., and Henning, U. (1993). Location and unusual membrane topology of the immunity protein of the *Escherichia coli* phage T4. J Virol *67*, 4905-4913.

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. Front Genet *5*, 102.

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., Timinskas, A., Wolf, Y.I., Gussow, A.B., Siksnys, V., Venclovas, C., and Koonin, E.V. (2020a). Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res *48*, 8828-8847.

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., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., *et al.* (2020b). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol *18*, 67-83.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2018). Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? CRISPR J *1*, 325-336.

Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011c). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol *193*, 6039-6056.

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

Makroczyova, J., Resetarova, S., Florek, P., and Barak, I. (2014). Topology of the *Bacillus subtilis* SpoIISA protein and its role in toxin-antitoxin function. FEMS Microbiol Lett *358*, 180-187.

Manghwar, H., Li, B., Ding, X., Hussain, A., Lindsey, K., Zhang, X., and Jin, S. (2020). CRISPR/Cas Systems in Genome Editing: Methodologies and Tools for sgRNA Design, Off-Target Evaluation, and Strategies to Mitigate Off-Target Effects. Adv Sci (Weinh) *7*, 1902312.

Manning, A.J., and Kuehn, M.J. (2011). Contribution of bacterial outer membrane vesicles to innate bacterial defense. BMC Microbiol *11*, 258.

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

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J *17*, 10-12.

Martinez-Santos, V.I., Medrano-Lopez, A., Saldana, Z., Giron, J.A., and Puente, J.L. (2012). Transcriptional regulation of the ecp operon by EcpR, IHF, and H-NS in attaching and effacing *Escherichia coli*. J Bacteriol *194*, 5020-5033.

McGinn, J., and Marraffini, L.A. (2016a). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol Cell.

McGinn, J., and Marraffini, L.A. (2016b). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol Cell *64*, 616-623.

McGinn, J., and Marraffini, L.A. (2019). Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat Rev Microbiol *17*, 7-12.

Medina-Aparicio, L., Rebollar-Flores, J.E., Gallego-Hernandez, A.L., Vazquez, A., Olvera, L., Gutierrez-Rios, R.M., Calva, E., and Hernandez-Lucas, I. (2011). The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. J Bacteriol *193*, 2396-2407.

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.

Meisel, A., Mackeldanz, P., Bickle, T.A., Kruger, D.H., and Schroeder, C. (1995). Type III restriction endonucleases translocate DNA in a reaction driven by recognition site-specific ATP hydrolysis. EMBO J *14*, 2958-2966.

Mekler, V., Minakhin, L., and Severinov, K. (2017). Mechanism of duplex DNA destabilization by RNA-guided Cas9 nuclease during target interrogation. Proc Natl Acad Sci U S A *114*, 5443-5448.

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.

Miller, H.I., Mozola, M.A., and Friedman, D.I. (1980). int-h: an int mutation of phage λ that enhances site-specific recombination. Cell *20*, 721-729.

Moch, C., Fromant, M., Blanquet, S., and Plateau, P. (2017). DNA binding specificities of Escherichia coli Cas1-Cas2 integrase drive its recruitment at the CRISPR locus. Nucleic Acids Res *45*, 2714-2723.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101-104.

Mohr, G., Silas, S., Stamos, J.L., Makarova, K.S., Markham, L.M., Yao, J., Lucas-Elío, P., Sanchez-Amat, A., Fire, A.Z., Koonin, E.V., *et al.* (2018). A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition. Mol Cell *72*, 700-714.e8.

Moitoso de Vargas, L., Kim, S., and Landy, A. (1989). DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase. Science *244*, 1457-1461.

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.

Molina, R., Stella, S., Feng, M., Sofos, N., Jauniskis, V., Pozdnyakova, I., Lopez-Mendez, B., She, Q., and Montoya, G. (2019). Structure of Csx1-cOA4 complex reveals the basis of RNA decay in Type III-B CRISPR-Cas. Nat Commun *10*, 4302.

Morad, I., Chapman-Shimshoni, D., Amitsur, M., and Kaufmann, G. (1993). Functional expression and properties of the tRNA(Lys)-specific core anticodon nuclease encoded by *Escherichia coli* prrC. J Biol Chem *268*, 26842-26849.

Morona, R., Klose, M., and Henning, U. (1984). *Escherichia coli* K-12 outer membrane protein (OmpA) as a bacteriophage receptor: analysis of mutant genes expressing altered proteins. J Bacteriol *159*, 570-578.

Mulepati, S., and Bailey, S. (2013). *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J Biol Chem *288*, 22184-22192.

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

Muller, M., Fazi, F., and Ciaudo, C. (2020). Argonaute Proteins: From Structure to Function in Development and Pathological Cell Fate Determination. Front Cell Dev Biol *7*, 360.

Musharova, O., Vyhovskyi, D., Medvedeva, S., Guzina, J., Zhitnyuk, Y., Djordjevic, M., Severinov, K., and Savitskaya, E. (2018). Avoidance of Trinucleotide Corresponding to Consensus Protospacer Adjacent Motif Controls the Efficiency of Prespacer Selection during Primed Adaptation. MBio *9*, e02169-18.

Muskavitch, K.M., and Linn, S. (1982a). A unified mechanism for the nuclease and unwinding activities of the recBC enzyme of Escherichia coli. J Biol Chem *257*, 2641-2648.

Muskavitch, K.M., and Linn, S. (1982b). A unified mechanism for the nuclease and unwinding activities of the recBC enzyme of Escherichia coli. J Biol Chem *257*, 2641-8.

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.

Naorem, S.S., Han, J., Wang, S., Lee, W.R., Heng, X., Miller, J.F., and Guo, H. (2017). DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. Proc Natl Acad Sci U S A *114*, E10187-E10195.

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (2017). Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature *548*, 543-548.

Niewoehner, O., and Jinek, M. (2016). Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. RNA *22*, 318-329.

Nimkar, S., and Anand, B. (2020). Cas3/I-C mediated target DNA recognition and cleavage during CRISPR interference are independent of the composition and architecture of Cascade surveillance complex. Nucleic Acids Res *48*, 2486-2501.

Nirwan, N., Singh, P., Mishra, G.G., Johnson, C.M., Szczelkun, M.D., Inoue, K., Vinothkumar, K.R., and Saikrishnan, K. (2019). Hexameric assembly of the AAA+ protein McrB is necessary for GTPase activity. Nucleic Acids Res *47*, 868-882.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of *Staphylococcus aureus* Cas9. Cell *162*, 1113-1126.

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.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205-217.

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.

Nuñez, 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.

Nussenzweig, P.M., McGinn, J., and Marraffini, L.A. (2019). Cas9 Cleavage of Viral Genomes Primes the Acquisition of New Immunological Memories. Cell Host Microbe *26*, 515-526.e6.

Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., Doron, S., and Sorek, R. (2018). DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat Microbiol *3*, 90-98.

Ofir, G., and Sorek, R. (2018). Contemporary Phage Biology: From Classic Models to New Insights. Cell *172*, 1260-1270.

Ogura, T., and Hiraga, S. (1983). Mini-F plasmid genes that couple host cell division to plasmid proliferation. Proc Natl Acad Sci U S A *80*, 4784-4788.

Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D.K., and Aravin, A.A. (2013). Bacterial argonaute samples the transcriptome to identify foreign DNA. Mol Cell *51*, 594-605.

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.

Özcan, A., Pausch, P., Linden, A., Wulf, A., Schuhle, K., Heider, J., Urlaub, H., Heimerl, T., Bange, G., and Randau, L. (2019). Type IV CRISPR RNA processing and effector complex formation in *Aromatoleum aromaticum*. Nat Microbiol *4*, 89-96.

Panne, D., Raleigh, E.A., and Bickle, T.A. (1999). The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. J Mol Biol *290*, 49-60.

Papapanagiotou, I., Streeter, S.D., Cary, P.D., and Kneale, G.G. (2007). DNA structural deformations in the interaction of the controller protein C.AhdI with its operator sequence. Nucleic Acids Res *35*, 2643-2650.

Parreira, R., Ehrlich, S.D., and Chopin, M.C. (1996). Dramatic decay of phage transcripts in lactococcal cells carrying the abortive infection determinant AbiB. Mol Microbiol *19*, 221-230.

Pedruzzi, I., Rosenbusch, J.P., and Locher, K.P. (1998). Inactivation *in vitro* of the *Escherichia coli* outer membrane protein FhuA by a phage T5-encoded lipoprotein. FEMS Microbiol Lett *168*, 119-125.

Penades, J.R., and Christie, G.E. (2015). The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. Annu Rev Virol *2*, 181-201.

Peters, J.E., Makarova, K.S., Shmakov, S., and Koonin, E.V. (2017). Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc Natl Acad Sci U S A *114*, E7358-E7366.

Pimentel, B., Nair, R., Bermejo-Rodriguez, C., Preston, M.A., Agu, C.A., Wang, X., Bernal, J.A., Sherratt, D.J., and de la Cueva-Mendez, G. (2014). Toxin Kid uncouples DNA replication and cell division to enforce retention of plasmid R1 in *Escherichia coli* cells. Proc Natl Acad Sci U S A *111*, 2734-2739.

Pingoud, A., and Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. Nucleic Acids Res *29*, 3705-3727.

Pinilla-Redondo, R., Mayo-Munoz, D., Russel, J., Garrett, R.A., Randau, L., Sorensen, S.J., and Shah, S.A. (2020). Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. Nucleic Acids Res *48*, 2000-2012.

Polisky, B., Greene, P., Garfin, D.E., McCarthy, B.J., Goodman, H.M., and Boyer, H.W. (1975). Specificity of substrate recognition by the EcoRI restriction endonuclease. Proc Natl Acad Sci U S A *72*, 3310-3314.

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.

Pribil, P.A., and Haniford, D.B. (2003). Target DNA bending is an important specificity determinant in target site selection in Tn10 transposition. J Mol Biol *330*, 247-259.

Puigbo, P., Makarova, K.S., Kristensen, D.M., Wolf, Y.I., and Koonin, E.V. (2017). Reconstruction of the evolution of microbial defense systems. BMC Evol Biol *17*, 94.

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.

Punetha, A., Yoganand, K.N.R., Nimkar, S., and Anand, B. (2018). Cutting it right: Plasticity and strategy of CRISPR RNA specific nucleases. Proc Indian Natl Sci Acad B Biol Sci *84*, 455-477.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173-1183.

Radovcic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L., and Ivancic-Bace, I. (2018). CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. Nucleic Acids Res *46*, 10173-10183.

Ramachandran, A., Summerville, L., Learn, B., DeBell, L., and Bailey, S. (2019). Processing and integration of functionally oriented prespacers in the *E. coli* CRISPR system depends on bacterial host exonucleases. J Biol Chem, DOI: 10.1074/jbc.RA119.012196.

Ramisetty, B.C.M., and Sudhakari, P.A. (2019). Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. Front Genet *10*, 65.

Rao, C., Chin, D., and Ensminger, A.W. (2017). Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA *23*, 1525-1538.

Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. Cell *163*, 854-865.

Reeks, J., Sokolowski, R.D., Graham, S., Liu, H., Naismith, J.H., and White, M.F. (2013). Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. Biochem J *452*, 223-230.

Reich, S., Gossl, I., Reuter, M., Rabe, J.P., and Kruger, D.H. (2004). Scanning force microscopy of DNA translocation by the Type III restriction enzyme EcoP15I. J Mol Biol *341*, 337-343.

Reyes-Robles, T., Dillard, R.S., Cairns, L.S., Silva-Valenzuela, C.A., Housman, M., Ali, A., Wright, E.R., and Camilli, A. (2018). *Vibrio cholerae* Outer Membrane Vesicles Inhibit Bacteriophage Infection. J Bacteriol *200*, e00792-17.

Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell *87*, 1295-1306.

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.

Richter, H., Lange, S.J., Backofen, R., and Randau, L. (2013). Comparative analysis of Cas6b processing and CRISPR RNA stability. RNA Biol *10*, 700-707.

Riede, I., and Eschbach, M.L. (1986). Evidence that TraT interacts with OmpA of *Escherichia coli*. FEBS Lett *205*, 241-245.

Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res *42*, W320-W324.

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-D299.

Rollie, C., Graham, S., Rouillon, C., and White, M.F. (2018). Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Res *46*, 1007-1020.

Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L., and White, M.F. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. Elife *4*, e08716.

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.

Rossmann, M.G., Mesyanzhinov, V.V., Arisaka, F., and Leiman, P.G. (2004). The bacteriophage T4 DNA injection machine. Curr Opin Struct Biol *14*, 171-180.

Rostol, J.T., and Marraffini, L. (2019). (Ph)ighting Phages: How Bacteria Resist Their Parasites. Cell Host Microbe *25*, 184-194.

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.

Ryan, V.T., Grimwade, J.E., Nievera, C.J., and Leonard, A.C. (2002). IHF and HU stimulate assembly of pre-replication complexes at *Escherichia coli* oriC by two different mechanisms. Mol Microbiol *46*, 113-124.

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., Belanger, M., and Moineau, S. (2013a). Effect of the abortive infection mechanism and type III toxin/antitoxin system AbiQ on the lytic cycle of *Lactococcus lactis* phages. J Bacteriol *195*, 3947-3956.

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

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463.

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

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol Cell *46*, 606-615.

Sasnauskas, G., and Siksnys, V. (2020). CRISPR adaptation from a structural perspective. Curr Opin Struct Biol *65*, 17-25.

Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A., and Severinov, K. (2013). High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. RNA Biol *10*, 716-725.

Schifano, J.M., Cruz, J.W., Vvedenskaya, I.O., Edifor, R., Ouyang, M., Husson, R.N., Nickels, B.E., and Woychik, N.A. (2016). tRNA is a new target for cleavage by a MazF toxin. Nucleic Acids Res *44*, 1256-1270.

Scholl, D., Adhya, S., and Merril, C. (2005). *Escherichia coli* K1's capsule is a barrier to bacteriophage T7. Appl Environ Microbiol *71*, 4872-4874.

Scholl, D., Rogers, S., Adhya, S., and Merril, C.R. (2001). Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of *Escherichia coli*. J Virol *75*, 2509-2515.

Sefcikova, J., Roth, M., Yu, G., and Li, H. (2017). Cas6 processes tight and relaxed repeat RNA via multiple mechanisms: A hypothesis. Bioessays *39*.

Segall, A.M., and Nash, H.A. (1996). Architectural flexibility in lambda site-specific recombination: three alternate conformations channel the attL site into three distinct pathways. Genes Cells *1*, 453-463.

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., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A., Logacheva, M.D., and Severinov, K. (2016). Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. Proc Natl Acad Sci U S A *113*, 7626-7631.

Shah, S.A., Alkhnbashi, O.S., Behler, J., Han, W., She, Q., Hess, W.R., Garrett, R.A., and Backofen, R. (2019). Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biol *16*, 530-542.

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.

Sheppard, N.F., Glover, C.V., 3rd, Terns, R.M., and Terns, M.P. (2016). The CRISPR-associated Csx1 protein of *Pyrococcus furiosus* is an adenosine-specific endoribonuclease. RNA *22*, 216-224.

Shiimori, M., Garrett, S.C., Graveley, B.R., and Terns, M.P. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol Cell *70*, 814-824.e6.

Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175.

Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature *547*, 345-349.

Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., *et al.* (2019). Detection of spacer precursors formed in vivo during primed CRISPR adaptation. Nat Commun *10*, 4603.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., *et al.* (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell *60*, 385-397.

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M.D., Datsenko, K.A., and Severinov, K. (2014). Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res *42*, 5907-5916.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., *et al.* (2017). Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol *15*, 169-182.

Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Severinov, K.V., and Koonin, E.V. (2018). Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc Natl Acad Sci U S A *115*, E5307-E5316.

Silas, S., Makarova, K.S., Shmakov, S., Paez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S., Krishnamurthy, S.R., Fu, B.X.H., *et al.* (2017). On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. MBio *8*, e00897-17.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science *351*, aad4234.

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.

Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. EMBO J *32*, 385-394.

Sistla, S., and Rao, D.N. (2004). S-Adenosyl-L-methionine-dependent restriction enzymes. Crit Rev Biochem Mol Biol *39*, 1-19.

Smargon, A.A., Cox, D.B., Pyzocha, N.K., Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S., *et al.* (2017). Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol Cell *65*, 618-630.e7.

Smith, G.R. (2001). Chi Sequences. In Encyclopedia of Genetics, S. Brenner, and J.H. Miller, eds. (New York: Academic Press), pp. 325-328.

Smith, G.R. (2012). How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev *76*, 217-228.

Smith, H.O., and Wilcox, K.W. (1970). A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. J Mol Biol *51*, 379-391.

St-Pierre, F., Cui, L., Priest, D.G., Endy, D., Dodd, I.B., and Shearwin, K.E. (2013). One-step cloning and chromosomal integration of DNA. ACS Synth Biol *2*, 537-541.

Staals, R.H., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M., and Fineran, P.C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. Nat Commun *7*, 12853.

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.

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.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110-113.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62-67.

Steven, A.C., Trus, B.L., Maizel, J.V., Unser, M., Parry, D.A., Wall, J.S., Hainfeld, J.F., and Studier, F.W. (1988). Molecular substructure of a viral receptor-recognition protein. The gp17 tail-fiber of bacteriophage T7. J Mol Biol *200*, 351-365.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E.V., and Zhang, F. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. Science *365*, 48-53.

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. (2003). Phase variation in the phage growth limitation system of *Streptomyces coelicolor* A3(2). J Bacteriol *185*, 4558-4563.

Sun, X., Gohler, A., Heller, K.J., and Neve, H. (2006). The *ltp* gene of temperate *Streptococcus thermophilus* phage TP-J34 confers superinfection exclusion to *Streptococcus thermophilus* and *Lactococcus lactis*. Virology *350*, 146-157.

Sutherland, E., Coe, L., and Raleigh, E.A. (1992). McrBC: a multisubunit GTP-dependent restriction endonuclease. J Mol Biol *225*, 327-348.

Swarts, D.C., Hegge, J.W., Hinojo, I., Shiimori, M., Ellis, M.A., Dumrongkulraksa, J., Terns, R.M., Terns, M.P., and van der Oost, J. (2015). Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. Nucleic Acids Res *43*, 5120-5129.

Swarts, D.C., and Jinek, M. (2019). Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. Mol Cell *73*, 589-600 e4.

Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., Brouns, S.J.J., *et al.* (2014). DNA-guided DNA interference by a prokaryotic Argonaute. Nature *507*, 258-261.

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.

Swinger, K.K., and Rice, P.A. (2004). IHF and HU: flexible architects of bent DNA. Curr Opin Struct Biol *14*, 28-35.

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.

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.

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

Tormo-Mas, M.A., Mir, I., Shrestha, A., Tallent, S.M., Campoy, S., Lasa, I., Barbe, J., Novick, R.P., Christie, G.E., and Penades, J.R. (2010). Moonlighting bacteriophage proteins derepress staphylococcal pathogenicity islands. Nature *465*, 779-782.

Toro, N., Mestre, M.R., Martinez-Abarca, F., and Gonzalez-Delgado, A. (2019). Recruitment of Reverse Transcriptase-Cas1 Fusion Proteins by Type VI-A CRISPR-Cas Systems. Front Microbiol *10*, 2160.

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

Vidakovic, L., Singh, P.K., Hartmann, R., Nadell, C.D., and Drescher, K. (2018). Dynamic biofilm architecture confers individual and collective mechanisms of viral protection. Nat Microbiol *3*, 26-31.

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.

Waldminghaus, T., and Skarstad, K. (2010). ChIP on Chip: surprising results are often artifacts. BMC Genomics *11*, 414.

Wang, C., Villion, M., Semper, C., Coros, C., Moineau, S., and Zimmerly, S. (2011). A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA *in vitro*. Nucleic Acids Res *39*, 7620-7629.

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., and Li, H. (2012). The mysterious RAMP proteins and their roles in small RNA-based immunity. Protein Sci *21*, 463-470.

Wang, R., Li, M., Gong, L., Hu, S., and Xiang, H. (2016). DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in *Haloarcula hispanica*. Nucleic Acids Res *44*, 4266-4277.

Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M., and Wood, T.K. (2010). Cryptic prophages help bacteria cope with adverse environments. Nat Commun *1*, 147.

Wang, X., and Wood, T.K. (2016). Cryptic prophages as targets for drug development. Drug Resist Updat *27*, 30-38.

Warren, R.A. (1980). Modified bases in bacteriophage DNAs. Annu Rev Microbiol *34*, 137-158.

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*, 1749-1758.

Wei, Y., and Terns, M.P. (2016). CRISPR Outsourcing: Commissioning IHF for Site-Specific Integration of Foreign DNA at the CRISPR Array. Mol Cell *62*, 803-804.

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.

Werts, C., Michel, V., Hofnung, M., and Charbit, A. (1994). Adsorption of bacteriophage lambda on the LamB protein of *Escherichia coli* K-12: point mutations in gene *J* of lambda responsible for extended host range. J Bacteriol *176*, 941-947.

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., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., and Brouns, S.J. (2013). Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. PLoS Genet *9*, e1003742.

Westra, E.R., van Erp, P.B., Kunne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., *et al.* (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell *46*, 595-605.

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

Wigley, D.B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nat Rev Microbiol *11*, 9-13.

Wilkinson, M., Drabavicius, G., Silanskas, A., Gasiunas, G., Siksnys, V., and Wigley, D.B. (2019). Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. Mol Cell *75*, 90-101.e5.

Wilson, R.C., and Doudna, J.A. (2013). Molecular mechanisms of RNA interference. Annu Rev Biophys *42*, 217-239.

Winther, K., Tree, J.J., Tollervey, D., and Gerdes, K. (2016). VapCs of *Mycobacterium tuberculosis* cleave RNAs essential for translation. Nucleic Acids Res *44*, 9860-9871.

Wright, A.V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat Struct Mol Biol *23*, 876-883.

Wright, A.V., Liu, J.J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. Science *357*, 1113-1118.

Wright, A.V., Wang, J.Y., Burstein, D., Harrington, L.B., Paez-Espino, D., Kyrpides, N.C., Iavarone, A.T., Banfield, J.F., and Doudna, J.A. (2019). A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. Mol Cell *73*, 727-737.e3.

Wu, D., Guan, X., Zhu, Y., Ren, K., and Huang, Z. (2017). Structural basis of stringent PAM recognition by CRISPR-C2c1 in complex with sgRNA. Cell Res *27*, 705-708.

Xiao, Y., Luo, M., Dolan, A.E., Liao, M., and Ke, A. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *361*, eaat0839.

Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017a). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48-60 e11.

Xiao, Y., Ng, S., Nam, K.H., and Ke, A. (2017b). How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature *550*, 137-141.

Xue, C., Whitis, N.R., and Sashital, D.G. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol Cell *64*, 826-834.

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., *et al.* (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell *165*, 949-962.

Yan, W.X., Hunnewell, P., Alfonse, L.E., Carte, J.M., Keston-Smith, E., Sothiselvam, S., Garrity, A.J., Chong, S., Makarova, K.S., Koonin, E.V., *et al.* (2019). Functionally diverse type V CRISPR-Cas systems. Science *363*, 88-91.

Yang, J.Y., Jayaram, M., and Harshey, R.M. (1996). Positional information within the Mu transposase tetramer: catalytic contributions of individual monomers. Cell *85*, P447-455.

Yoganand, K.N., Muralidharan, M., Nimkar, S., and Anand, B. (2019). Fidelity of prespacer capture and processing is governed by the PAM-mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system. J Biol Chem *294*, 20039-20053.

Yoganand, K.N.R., Sivathanu, R., Nimkar, S., and Anand, B. (2017). Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Res *45*, 367-381.

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.

Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T., and Qimron, U. (2013). DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. Proc Natl Acad Sci U S A *110*, 14396-14401.

Young, R. (2014). Phage lysis: three steps, three choices, one outcome. J Microbiol *52*, 243-258.

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, B., Ye, Y., Ye, W., Perculija, V., Jiang, H., Chen, Y., Li, Y., Chen, J., Lin, J., Wang, S., *et al.* (2019a). Two HEPN domains dictate CRISPR RNA maturation and target cleavage in Cas13d. Nat Commun *10*, 2544.

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.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics *9*, 40.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. Mol Cell *50*, 488-503.

Zhang, Z., Pan, S., Liu, T., Li, Y., and Peng, N. (2019b). Cas4 nucleases can effect specific integration of CRISPR spacers. J Bacteriol *201*, e00747-18.

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.

Zwieb, C., and Adhya, S. (2009). Plasmid vectors for the analysis of protein-induced DNA bending. Methods Mol Biol *543*, 547-562.

Appendix

Table 1: List of strains used in the study

S/N	*E. coli* strain	Genotype	Source
1	IYB5101 (WT)	F^-Δ(araD-araB)567 ΔlacZ4787 (::rrnB-3) λ^- rph-1 Δ(rhaD-rhaB)568 hsdR514 araB::T7-RNAp-tetA, Tet^R	(Yosef et al., 2012)
2	JW1702	F^-Δ(araD-araB)567 ΔlacZ4787 (::rrnB-3) λ^-, ΔihfA786::kan , rph-1 Δ(rhaD-rhaB)568 hsdR514, Kan^R	CGSC#: 9441 (Baba et al., 2006)
3	JW0895	F^-Δ(araD-araB)567 ΔlacZ4787 (::rrnB-3) λ^-, ΔihfB735::kan , rph-1 Δ(rhaD-rhaB)568 hsdR514, Kan^R	CGSC#: 8917 (Baba et al., 2006)
4	BL21-AI	F^- ompT hsdsB(rB-, mB-) gal dcm araB:: T7-RNAp-tetA, Tet^R	Invitrogen
5	BL21(DE3)	F^- ompT hsdsB(rB-, mB-) gal dcm λ(DE3)	
6	TOP10	F^- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 Δ lacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL endA1 nupG, Str^R	Invitrogen
7	DH5α	F^- Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17(r_k^-, m_k^+) phoA supE44 thi-1 gyrA96 relA1 λ^-	Invitrogen
8	ΔIHFα	IYB5101 ΔihfA::kan, Kan^R, Tet^R	This study
9	ΔIHFβ	IYB5101 ΔihfB::kan, Kan^R, Tet^R	This study
10	WT	IYB5101 P21::CRISPR 2.1 leader with 2 repeats, Cam^R, Tet^R	This study
11	IBS	IYB5101 P21:: CRISPR 2.1 leader with mutated IHF binding site and two repeats, Cam^R, Tet^R	This study
12	ΔIBS	IYB5101 P21:: CRISPR 2.1 leader with IHF binding site deletion and two repeats, Cam^R, Tet^R	This study
13	CBS1	IYB5101 P21:: CRISPR 2.1 leader with mutated -34-45 region and two repeats, Cam^R, Tet^R	This study
14	CBS2	IYB5101 P21:: CRISPR 2.1 leader with mutated -46-57 region and two repeats, Cam^R, Tet^R	This study
15	CBS3	IYB5101 P21:: CRISPR 2.1 leader with mutated -58-69 region and two repeats, Cam^R, Tet^R	This study
16	CBS2(L)	IYB5101 P21:: CRISPR 2.1 leader with mutated -54-57 region and two repeats, Cam^R, Tet^R	This study
17	CBS2(C)	IYB5101 P21:: CRISPR 2.1 leader with mutated -50-53 region and two repeats, Cam^R, Tet^R	This study

Table 2: List of plasmids used in the study

S/N	Plasmid name	Description	Source
1	pdCas9-bacteria	ori p15A, Cam^R, tetR, expresses S. pyogenes dCas9 (D10A, H840A under anhydrotetracycline inducible promoter ($P_{LtetO-1}$).	Addgene #44249 (Qi et al., 2013)
2	p3XFLAG-dCas9	pdCas9-bacteria modified with the construct encoding 3XFLAG-dCas9-StrepII.	This study
3	pgRNA-bacteria	ori pMB1, Amp^R, expresses gRNA under the control of constitutive promoter (P_{J23119}).	Addgene #44251 (Qi et al., 2013)
4	pgRNA-leader	pgRNA-bacteria modified with a construct expressing gRNA targeted to 86 bp upstream of BL21-AI CRISPR leader-repeat junction.	This study
5	pKD46	ori R101, repA101ts, Amp^R, araC, expresses λ Red genes (gam-bet-exo) under the control of arabinose inducible promoter (P_{araBAD}).	CGSC #7739 (Datsenko and Wanner, 2000)
6	pOSIP-CT	ori Rγ, ori pUC, Cam^R, attP P21, ccdB, λ (cI857) encodes P21 integrase under the control of λ promoter (λ pR).	Addgene #45981 (St-Pierre et al., 2013)
7	pCas1-2[K]	ori CloDF13, Spc^R, lacI, expresses Cas1 and Cas2 protein coding genes under the control of IPTG inducible promoter (P_{T7lac}).	(Diez-Villasenor et al., 2013)
8	pCSIR-T	ori pMB1, Amp^R, Kan^R, lacI, encompass leader and two repeat spacer units derived from CRISPR 2.1 of *E. coli* K-12 MG1655.	(Diez-Villasenor et al., 2013)
9	p8R	ori pMB1, Amp^R, araC, expresses the gene of interest to synthesise N-terminal StrepII tagged proteins under the control of arabinose inducible promoter (P_{araBAD}).	Addgene #37506 (Scott Gradia)
10	p8R-IHFαβ	p8R modified with a bicistronic cassette encoding IHFα and IHFβ.	This study
11	p1R	ori pMB1, Kan^R, lacI, expresses the gene of interest to synthesise N-terminal StrepII tagged protein under the control IPTG inducible promoter (P_{T7lac}).	Addgene #29664 (Scott Gradia)
12	p1R-IHFαβ	p1R modified with a bicistronic cassette encoding IHFα and IHFβ.	This study
13	p13SR	ori CloDF13, Spc^R, lacI, expresses the gene of interest to synthesize N-terminal StrepII tagged protein under the control IPTG inducible promoter (P_{T7lac}).	Addgene #48328 (Scott Gradia)
14	p13SR-Cas1	p13SR modified with a construct encoding N-terminally StrepII tagged Cas1.	This study
15	p1S	ori pMB1, Kan^R, lacI, expresses the gene of interest to synthesise N-terminally His-SUMO tagged protein under the control of IPTG inducible promoter (P_{T7lac}).	Addgene #29659 (Scott Gradia)
16	p1S-Cas2	p1S modified with a construct encoding His-SUMO-Cas2-StrepII.	This study
17	pBend5	ori pMB1, Amp^R, encompass circularly permuted restriction sites bordered around HpaI cutting site.	(Zwieb and Adhya, 2009)
18	pBend-WT	pBEND5 modified with a WT CRISPR 2.1 leader DNA.	This study
19	pBend-CBS2	pBEND5 modified with a CBS2 CRISPR 2.1 leader DNA.	This study
20	pMS	T7-lac inducible, RSF origin plasmid for expressing N-terminally 6X His-MBP-SUMO tagged genes.	Addgene #64693 (Guerrero et al., 2015)
21	pFGET19_Ulp1	Lac inducible plasmid for expression of 6X His tagged SUMO protease catalytic domain ($Ulp1_{403-621}$).	Addgene #64697 (Guerrero et al., 2015)
22	pMS-Cas2	The plasmid expressing N-terminally 6X His-MBP-SUMO tagged, C-terminally Strep-II tagged Cas2 under T7-lac promoter.	This study
23	pCas1-2H	The plasmid expressing Cas1 (WT), C-terminally 6X His tagged Cas2 under T7-lac promoter.	This study

24	p5M	The plasmid expressing Cas1 (5M: Q24H, P202Q, G241D, E276D, L297Q), C-terminally 6X His tagged Cas2 under T7-lac promoter.	This study
25	pY22A	The plasmid expressing Cas1 (Y22A), C-terminally 6X His tagged Cas2 under T7-lac promoter.	This study
26	pΔC	The plasmid expressing Cas1 (ΔC: ΔP279-S305), C-terminally 6X His tagged Cas2 under T7-lac promoter.	This study

Table 3: List of oligonucleotides used in the study

S/N	Oligo name	Sequence (5'-3')	Description
1	ΔIHFα FP	GCCTCGCCATAAGCCTGATCC	Amplification of IHFα deletion cassette from CGSC strain # 9441 (Baba et al., 2006).
2	ΔIHFα RP	GGCGGGCAAGTGTCAGCATG	
3	ΔIHFβ FP	CCGAAGTTTGTTGAGTTTACTTGACAG	Amplification of IHFβ deletion cassette from CGSC strain # 8917 (Baba et al., 2006).
4	ΔIHFβ RP	GGTGAAAACTGTTCAAAACTCTGCG	
5	3XFLAG-dCas9 FP1	GTGATAGAGAAAAGAATTCA AAAGATCTAAAGAGGAGAAA GGATCTATGGATTACAAGGA	Creation of 3XFLAG fragment for amplification of 3XFLAG-dCas9-StrepII construct.
6	3XFLAG-dCAS9 RP1	AGTCGATGTCGTGATCTTTGTA GTCGCCGTCATGGTCCTTGTAA TCCATAGATCCTTTCT	
7	3XFLAG-dCAS9 FP2	GACTACAAAGATCACGACATCG ACTACAAAGACGATGACGATAA AATGGATAAGAAATAC	
8	3XFLAG-dCAS9 RP2	GCTAAGCCTATTGAGTATTTCTT ATCCATTTTATCGTCATCGT	
9	Strep-dCas9 FP1	GATTTGAGTCAGCTAGGAGGTGA CTCCGCCTGGAGCCATCCTCAATT CGAGAAATAGTAA	
10	Strep-dCas9 RP1	ATTTGATGCCTGGAGATCCTTACT CGAGTTACTATTTCTCGAATTGAG GATGGCTC	
11	gRNA FP1	GACAGCTAGCTCAGTCCTAGGTAT AATACTAGTTTAAGTACTC	Amplification of gRNA-Cas9 handle-terminator construct.
12	gRNA RP1	GCTATTTCTAGCTCTAAAACTTATGTTAAA GAGTACTTAAACTAGTATTATACCTAGGAC	
13	gRNA FP2	AACATAAGTTTTAGAGCTAGAAATAGCAAG TTAAAATAAGGCTAGTCCGTTATCAACTTG	
14	gRNA RP2	GAGATGAGTTTTTGTTCGGGCCCAAGCTTC AAAAAAAGCACCGACTCGGTGCCACTTTTT CAAGTTGATAACGGACTAGCCTTATTTTAA	
15	BG3474	AAATGTTACATTAAGGTTGGTG	To check expansion of CRISPR 2.1 locus (Swarts et al., 2012) of IYB5101 (WT).
16	BG3475	GAAATTCCAGACCCGATCC	
17	2.1 array F	GGAAATGTTACATTAAGGTTGGTGGGTTG	Monitoring the spacer incorporation into the 2.1 CRISPR array during in vivo integration assay.
18	2.1 array R	CGCTCAGAAATTCCAGACCCGATCC	
19	WT FP	GGAATTGGGGATCGGAATTCGAGCTCGGTA CCCACCTTTGGCTTCGGCTGCGGT	Amplification of CRISPR 2.1 array from plasmid pCSIR-T (Diez-Villasenor et al., 2013), for clonetegration into P21 attB site.
20	WT RP	TAGGCGCCATGCATCTCGAGGCATGCCTGC AGCGGCCGCCAGTGTGATGGATATCTG	
21	IBS FP	CATTTGTGCCCGGCGCATCGCTTCCC ATACAAACAACCCACCA	Along with WT FP and RP, these primers are used to amplify IBS (IHF site mutant), for clonetegration into P21 attB site.
22	IBS RP	GCGATGCGCCGGGCACAAATGTATA CTTTTAGAGAGTTCCCC	
23	Δ FP1	TTCCCATACAAACAACCCACCA	Along with WT FP and RP, these primers are used to amplify ΔIBS (IHF site deletion), for clonetegration into P21 attB site.
24	Δ RP1	ACAAATGTATACTTTTAGAGAGTTCCCC	
25	Δ FP2	GGGGAACTCTCTAAAAGTATACATTT	

		GTTTCCCATACAAACAACCCACCA	
26	CBS1 FP	CTTAAAGCATTTTCGATTGCCTGCGCA ACCCACCAACCTTAATGTAACA	Along with WT FP and RP, these primers are used to amplify CBS1, for clonetegration into P21 attB site.
27	CBS1 RP	GGTTGCGCAGGCAATCGAAAATGCTT TAAGAACAAATGTATACTTTTAGAGAG	
28	CBS2 FP	TTTCCCATACAAATCGTATCTAGGTCT TAATGTAACAAAAGCCGAATTCTGC	Along with WT FP and RP, these primers are used to amplify CBS2, for clonetegration into P21 attB site.
29	CBS2 RP	TTAAGACCTAGATACGATTTGTATGGG AAAAAATGCTTTAAGAACAAA	
30	CBS3 FP	ACAACCCACCAACACGCCGACGCTCAA AGCCGAATTCTGCAGATATCCA	Along with WT FP and RP, these primers are used to amplify CBS3, for clonetegration into P21 attB site.
31	CBS3 RP	GCTTTGAGCGTCGGCGTGTTGGTGGGTT GTTTGTATGGG	
32	CBS2(L) FP	TTAAGACCTGTGGGTTGTTTGTATGGGA AAAAATGCTTTAAGAACAAA	Along with WT FP and RP, these primers are used to amplify CBS2(L), for clonetegration into P21 attB site.
33	CBS2(L) RP	TTTCCCATACAAACAACCCACAGGTCTT AATGTAACAAAAGCCGAATTCTGC	
34	CBS2(C) FP	TTAAGGTTGAGATGTTGTTTGTATGGGA AAAAATGCTTTAAGAACAAA	Along with WT FP and RP, these primers are used to amplify CBS2(C), for clonetegration into P21 attB site.
35	CBS2(C) RP	TTTCCCATACAAACAACATCTCAACC TTAATGTAACAAAAGCCGAATTCTGC	
36	CBS2(R) FP	TTAAGGTTGGTGGACGATTTGTATGG GAAAAAATGCTTTAAGAACAAA	Along with WT FP and RP, these primers are used to amplify CBS2(R), for clonetegration into P21 attB site.
37	CBS2(R) RP	TTTCCCATACAAATCGTCCACCAACCTT AATGTAACAAAAGCCGAATTCTGC	
42	P21 FP	GGATCGGAATTCGAGCTCGGT	Amplification of CRISPR 2.1 array from plasmid pCSIR-T (Diez-Villasenor et al., 2013), for clonetegration into P21 attB site.
43	P21 RP	TGCATCTCGAGGCATGCC	
44	IHF WT strand 1	GTTCTTAAAGCATTTTTTCC-3'-6-FAM	Preparation of FRET bending assay substrate encompassing -4 to -38 nt region upstream to WT leader-repeat junction.
45	IHF WT strand 2	GGGAAAAAATGCTTTAAGAACAAAT GTATACTTTT	
46	IHF strand 3	IOWA Black FQ-5'-TAAAAGTATACATTT	
47	IHF strand 3 quencher-	TAAAAGTATACATTT	
48	IHF Δ strand1	GTTTCCCATACAAACAACCC-3'-6-FAM	Along with IHF strand 3, these oligos were annealed to prepare substrate with -4 to -38 nt region upstream to ΔIBS leader-repeat junction.
49	IHF Δ strand2	TGGGTTGTTTGTATGGGAAACAAATGT ATACTTTT	
50	Cas1 LIC FP	TACTTCCAATCCAATGCAATGACCTGGC TTCCCCTTAATCC	Amplification of gene encoding Cas1 with flanking regions of p13S-R SspI site.
51	Cas1 LIC RP	TTATCCACTTCCAATGTTATTATCAGCTA CTCCGATGGCCTGC	
52	Cas2 LIC FP	TACTTCCAATCCAATGCAATGAGTATGT TGGTCGTGGTCACTG	Amplification gene encoding Cas2-StrepII with flanking regions of p1S SspI site.
53	Cas2 LIC Strep RP	TTATCCACTTCCAATGTTATTATTTTTC GAACTGCGGGTGGCTCCAAGCGCTAAC AGGTAAAAAAGACACCAACCTTAAAC	
54	Ihfα LIC FP	TACTTCCAATCCAATGCAATGGCGCT TACAAAAGCTGAAATGT	Amplification of bicistronic cassette encoding IHFα and IHFβ with flanking sites of p8R SspI site.
55	Ihfα-RBS RP	TTGGTCATGGTATATCTCCTTCTTAAAG TTAATTACTCGTCTTTGGGCGAAGC	
56	RBS-Ihfβ FP	ATTAACTTTAAGAAGGAGATATACCATG ACCAAGTCAGAATTGATAGAAAGACT	
57	Ihfβ LIC RP	TTATCCACTTCCAATGTTATTATTAACCG TAAATATTGGCGCGATCGC	

58	Cas2 MS F	TCACAGAGAACAGATTGGTGGATCCGGAG GTATGAGTATGTTGGTCGTGGTCACTG	Generation of DNA encoding Cas2 with flanking sequences of BamHI and HindIII digested pMS.
59	Cas2 strep R	TTATTATTTTTCGAACTGCGGGTGGCTC CAAGCGCTAACAGGTAAAAAAGACACCA ACCTTAAAC	
60	MS strep R	CTTTACCAGACTCGAGTGCGGCCGCAAGC TTTTATTATTTTTCGAACTGCGGGTGGC	
61	CDF wt Cas1 F	AACTTTAATAAGGAGATATACCATGGCCT GGCTTCCCCTTAATCCC	Generation of DNA constructs encoding Cas1-2 variants (Wt, 5M, ΔC, Y22A) with flanking sequences of NcoI and NotI digested pCas1-2[K].
62	CDF Cas2 His R	TTATTAGTGATGGTGATGGTGATGAGCG CTAACAGGTAAAAAAGACACCAACCTTA AACC	
63	CDF His R	TTTCTTTACCAGACTCGAGTGCGGCCGCT TATTAGTGATGGTGATGGTGATGAGC	
64	ΔC Cas1-2 R	TTCAGGTGGGGCCGGCGGCTATTATTGTA TTTCTCCAGCGGCAAGC	
65	ΔC Cas1-2 F	TAATAGCCGCCGGCCCCACCTGAA	
66	Cas1 Y22A F	TCGCGTCTCCATGATCTTTCTGCAAGCTG GGCAGATCGAT	
67	Bend WT FP	TAGAGTTCTCTAAAAGTATACATTTGTTCT TAAAGCATTTTTTCCCATACAAACAACCCA	Creation of WT leader DNA for blunt-end ligation into plasmid pBEND5
68	Bend WT RP	GTTCTGTTACATTAAGGTTGGTGGGTTGTT TGTATGGGAAAAAATGCTTTAAGAACAAAT	
69	Bend CBS2 FP	TAGAGTTCTCTAAAAGTATACATTTGTTCT TAAAGCATTTTTTCCCATACAAATCGTATC	Creation of CBS2 leader DNA for blunt end ligation into plasmid pBEND5.
70	Bend CBS2 RP	GTTCTGTTACATTAAGACCTAGATACGATT TGTATGGGAAAAAATGCTTTAAGAACAAAT	
71	Leader F	TGCATCTCGAGGCATGCCTGCAGCGGCCG CCAGTGTGATGGATATCTG	Generation of CRISPR DNA substrates (CD-U, CD-T* and CD-B*) for in vitro integration assay.
72	Repeat2 R	GGATCGGAATTCGAGCTCGGTACCCACCT TTGGCTTCGGCTGC	
73	P23[3'-5] F	ATTTACTACTCGTTCTGGTGTTTCTCGT	These oligos were annealed to prepare P23[3'-5] prespacer.
74	P23[3'-5] R	AAACACCAGAACGAGTAGTAAATTGGGC	
75	P23[3'-10] F	ATTTACTACTCGTTCTGGTGTTTCTCGTC AGGG	These oligos were annealed to prepare P23[3'-10] prespacer.
76	P23[3'-10] R	AAACACCAGAACGAGTAGTAAATTGGGCT TGAG	

77	P33 F	GCCCAATTTACTACTCGTTCTGGTGTTTC TCGT	These oligos were annealed to prepare P33 prespacer. Whereas, P33 F oligo alone is used as P33[ss] prespacer.
78	P33 R	ACGAGAAACACCAGAACGAGTAGTAAAT TGGGC	
79	P23[5'-5] F	GCCCAATTTACTACTCGTTCTGGTGTTT	These oligos were annealed to prepare P23[5'-5] prespacer.
80	P23[5'-5] R	ACGAGAAACACCAGAACGAGTAGTAAAT	
81	P63 F	CTCCGCGCTGTAG**AAG**TCACCATTGTTGT GCACGACGACATCATTCCGTGGCGTTAT CCAGCT	These oligos were annealed to prepare P63 prespacer (Shipman et al., 2016). Residues corresponding to the PAM are underlined.
82	P63 R	AGCTGGATAACGCCACGGAATGATGTCGT CGTGCACAACAATGGTGA**CTT**CTACAGCG CGGAG	
83	P63mPAM F	CTCCGCGCTGTAGCCCTCACCACTGTTGT GCACGACGACACCAGTCCGTGGCGTTAT CCAGCT	These oligos were annealed to prepare P63mPAM prespacer. Mutated residues in the oligo are shaded.
84	P63mPAM R	AGCTGGATAACGCCACGGACTGGTGTCGT CGTGCACAACAGTGGTGAGGGCTACAG CGCGGAG	
85	P63 3'FAM-F	CTCCGCGCTGTAG**AAG**TCACCATTGTTGT GCACGACGACATCATTCCGTGGCGT TATCCAGCT - FAM(3')	This oligo was annealed with P63 R to generate P63 T*.
86	P63 3'FAM-R	AGCTGGATAACGCCACGGAATGATGTCGT CGTGCACAACAATGGTGA**CTT**CTACAGC GCGGAG - FAM(3')	This oligo was annealed with P63 F to generate P63 B*.
87	P63mPAM 3'FAM F	CTCCGCGCTGTAGCCCTCACCACTGTTGT GCACGACGACACCAGTCCGTGGCGTTA TCCAGCT - FAM(3')	This oligo was annealed with P63mPAM R to generate P63mPAM T*.
88	P63mPAM 3'FAM R	AGCTGGATAACGCCACGGACTGGTGTCGT CGTGCACAACAGTGGTGAGGGCTACAGC GCGGAG - FAM(3')	This oligo was annealed with P63mPAM F to generate P63mPAM B*.

Table 4: List of *E. coli* proteins identified by mass spectrometry of CRISPR/dCas9 based immunoprecipitated mixture

S/N	Accession	Description	Score	Coverage	#Unique Peptides	#Peptides	#PSM
1	388476136	Chaperone Hsp70	801.08	45.77	0	21	27
2	388480091	Cpn60 chaperonin GroEL	750.79	42.88	17	21	25
3	388479899	Translation elongation factor EF-Tu	516.62	54.06	0	16	23
4	388479898	Translation elongation factor EF-G	397.79	17.47	9	9	9
5	388476237	Aconitate hydratase B	311.65	9.02	6	6	6
6	388476542	Peptidyl-prolyl cis/trans isomerase	306.60	29.63	0	12	13
7	388477317	Transcriptional regulator H-NS	283.75	35.04	5	7	8
8	388479943	RNA polymerase, subunit α	269.85	42.25	7	11	15
9	388476578	High temperature protein G	257.55	23.08	0	12	12
10	388478624	Protein disaggregation chaperone	253.84	11.32	0	8	8
11	388479281	RNA polymerase, subunit β'	249.24	8.03	0	11	11
12	388479282	RNA polymerase, subunit β	231.83	7.45	8	9	10
13	388477852	Glyceraldehyde-3-phosphate dehydrogenase A	226.82	20.54	6	6	7
14	388476815	Succinate dehydrogenase	203.89	12.59	4	6	6
15	388476819	Succinyl-CoA synthetase, subunit β	203.79	14.95	4	5	5
16	16130662	CRISPR adaptation protein, Cas1	197.58	40.66	7	9	15
17	388480090	Cpn10 chaperonin GroES	180.97	55.67	4	5	6
18	388476453	Transcriptional repressor LacI	178.06	17.78	6	6	6
19	387615242	Streptomycin 3"-adenylyltransferase	156.78	19.77	0	4	4
20	388479464	Transcription termination factor Rho	148.01	17.42	0	6	6
21	388476234	Dihydrolipoyltransacetylase	139.15	8.25	0	4	4
22	388479505	F1 sector of ATP synthase subunit β	127.00	8.70	3	3	3
23	388476818	Dihydrolipoyltranssuccinase	126.16	8.89	2	2	2
24	388478373	3-oxoacyl-[acyl-carrier-protein] synthase I	125.75	5.42	2	2	2
25	388476995	Ribosomal protein S1	118.03	13.64	4	6	6
26	388479220	ClpXP protease-specificity-enhancing factor	114.36	20.61	0	3	3
27	388479228	Malate dehydrogenase	110.76	9.62	2	2	3
28	388476235	Lipoamide dehydrogenase	85.25	7.17	2	3	3
29	16128325	Cyanate aminohydrolase	84.41	13.46	2	2	3
30	386610010	Heat shock protein GrpE	78.15	7.11	0	1	1
31	388476977	Seryl-tRNA synthetase	77.58	4.65	2	2	2
32	388479285	Ribosomal protein L1	74.43	21.37	3	4	4
33	388479897	Ribosomal protein S7	73.40	15.64	0	2	2
34	387615228	Chloramphenicol acetyltransferase	71.81	10.50	1	2	3
35	388479924	Ribosomal protein S3	65.13	22.32	3	5	5
36	388478716	Recombinase RecA	64.71	6.52	2	2	2
37	388477173	3-oxoacyl-[acyl-carrier-protein] reductase	64.37	11.89	0	3	3
38	388479917	Ribosomal protein S10	63.64	40.78	2	4	4
39	388478347	Phosphate acetyltransferase	63.04	3.22	0	2	2

40	388476137	Chaperone Hsp40	61.97	4.52	0	2	2
41	388476763	Hypothetical protein Y75_p0650	59.49	3.90	0	1	1
42	388477216	Isocitrate dehydrogenase	59.28	5.29	0	2	2
43	388479283	Ribosomal protein L7/L12	59.21	19.83	1	2	2
44	388479622	Glycine C-acetyltransferase	58.66	3.27	0	1	1
45	388479889	FKBP-type peptidyl prolyl cis-trans isomerase	58.37	4.59	1	1	1
46	388479942	Ribosomal protein S4	58.36	12.14	2	3	3
47	388476709	Alkyl hydroperoxide reductase, C22 subunit	57.93	21.39	1	2	3
48	388477793	Threonyl-tRNA synthetase	57.42	1.87	0	1	1
49	386610718	cAMP-activated transcriptional regulator CRP	56.44	4.76	1	1	1
50	388476148	Isoleucyl-tRNA synthetase	54.84	1.28	0	1	1
51	388476544	ATPase subunit of ClpX-ClpP serine protease	52.95	7.78	1	2	2
52	388476973	Leucine-responsive transcriptional regulator Lrp	52.09	9.76	2	2	2
53	388476579	Adenylate kinase	50.45	5.61	0	1	1
54	388476628	Peptidyl-prolyl cis-trans isomerase B	49.49	6.71	0	1	1
55	90111482	CRISPR adaptation ssRNA endonuclease, Cas2	49.48	21.28	2	2	2
56	388476816	Succinate dehydrogenase, FeS subunit	48.50	10.50	0	2	2
57	388476820	Succinyl-CoA synthetase subunit α	47.70	6.92	2	2	2
58	388479164	Transcription termination/ antitermination L factor	47.44	2.42	1	1	1
59	388479503	F1 sector of ATP synthase subunit α	46.66	2.92	2	2	2
60	388477776	Phosphoenolpyruvate synthase	44.77	3.54	0	3	3
61	388480185	PTS trehalose transporter subunit IIBC	44.21	6.55	0	2	2
62	388477786	Integration host factor (IHF) subunit α	44.16	10.10	1	1	1
63	388479918	Ribosomal protein L3	43.63	10.05	1	2	2
64	388479067	RNA polymerase, σ 70 factor	43.34	3.43	0	2	2
65	388479896	Ribosomal protein S12	40.97	6.45	1	1	1
66	388479160	Ribosomal protein S15	40.14	7.87	1	1	1
67	544578490	Hypothetical protein ECOPMV1_04296	39.98	15.38	1	1	1
68	388479930	Ribosomal protein L5	39.92	6.15	1	1	1
69	388478458	Cysteine synthase A	39.69	12.07	1	3	3
70	388479929	Ribosomal protein L24	38.82	9.62	0	1	1
71	388477174	Acyl carrier protein	38.73	11.54	1	1	2
72	388479286	Ribosomal protein L11	37.78	7.04	1	1	1
73	388477368	Enoyl-[acyl-carrier-protein] reductase	37.52	3.82	0	1	1
74	388476287	Ribosomal protein S2	36.54	10.79	1	2	2
75	388479406	Fatty acid oxidation complex subunit α	35.39	1.10	0	1	2
76	388476817	2-oxoglutarate decarboxylase E1 component	34.45	0.96	0	1	1
77	388479222	Ribosomal protein S9	34.36	6.15	1	1	1
78	388479933	Ribosomal protein L6	33.85	18.08	1	3	3
79	388478544	IMP dehydrogenase	33.48	2.87	0	1	1

80	388478637	Ribosomal protein L19	32.95	13.04	1	1	1
81	387619372	Dihydrolipoamide succinyl transferase E2 component	32.68	4.17	0	2	2
82	388477175	3-oxoacyl-[acyl-carrier-protein] synthase II	31.09	3.39	1	1	1
83	388479802	HTH-type transcriptional regulator GntR	31.03	2.42	0	1	6
84	388479333	Glycerol kinase	30.99	2.19	0	1	1
85	544574870	Inosine-guanosine kinase	30.53	1.84	1	1	2
86	388476812	Citrate synthase	30.46	7.49	2	3	3
87	388478949	Methionine adenosyltransferase 1	29.82	2.60	1	1	1
88	388479328	HslU--HslV peptidase ATPase subunit	29.76	2.93	0	1	1
89	388479163	Translation initiation factor 2, IF2	28.80	4.04	0	4	5
90	387617138	Putative transcriptional regulator (LysR family)	28.31	7.27	1	1	1
91	388476996	Integration host factor (IHF) subunit β	28.27	11.70	1	1	1
92	388479171	ATP-dependent zinc metalloprotease FtsH	28.08	1.24	0	1	1
93	388478587	Serine hydroxymethyltransferase	27.57	5.28	1	2	2
94	487585326	L-Ala-D/L-Glu epimerase	27.05	4.05	0	1	1
95	388479928	Ribosomal protein L14	24.09	13.01	1	2	2
96	388478278	DNA gyrase, subunit A	23.63	1.03	0	1	1
97	84060915	Putative transposase	23.28	1.13	1	1	1
98	754638329	DNA topoisomerase IV subunit B	23.01	2.22	1	1	4
99	16130434	Exonuclease VII, large subunit	22.78	2.85	0	1	1
100	388477014	Asparaginyl tRNA synthetase	22.64	1.50	0	1	1
101	16128212	Antitoxin of YafQ-DinJ toxin-antitoxin system	22.10	12.79	1	1	1
102	388476288	Translation elongation factor EF-Ts	21.47	13.78	0	3	3
103	544578675	HTH-type transcriptional repressor CytR	20.70	2.62	0	1	1
104	387615345	Hypothetical protein NRG857_00010	20.29	11.54	1	1	1
105	387617413	GDP-mannose mannosyl hydrolase	20.25	5.00	0	1	1
106	544576814	Hypothetical protein ECOPMV1_02516	20.14	3.62	0	2	2

Table 5: IHF distribution among type I-E organisms

S/N	Organism Name	IHF presence
1	Acetobacter_pasteurianus_IFO_3283_01_uid59279	Yes
2	Alkalilimnicola_ehrlichii_MLHE_1_uid58467	Yes
3	Allochromatium_vinosum_DSM_180_uid46083	Yes
4	Amycolatopsis_mediterranei_S699_uid158689	Yes
5	Anaeromyxobacter_dehalogenans_2CP_1_uid58989	Yes
6	Arcanobacterium_haemolyticum_DSM_20595_uid49489	Yes
7	Azotobacter_vinelandii_DJ_uid57597	Yes
8	Candidatus_Accumulibacter_phosphatis_clade_IIA_UW_1_uid59207	Yes
9	Candidatus_Nitrospira_defluvii_uid51175	Yes
10	Catenulispora_acidiphila_DSM_44928_uid59077	Yes
11	Cellulomonas_fimi_ATCC_484_uid66779	Yes
12	Chlorobium_tepidum_TLS_uid57897	Yes
13	Chromohalobacter_salexigens_DSM_3043_uid62921	Yes
14	Cronobacter_turicensis_z3032_uid40821	Yes
15	Cycloclasticus_zancles_7_ME_uid214092	Yes
16	Desulfatibacillum_alkenivorans_AK_01_uid58913	Yes
17	Desulfococcus_oleovorans_Hxd3_uid58777	Yes
18	Desulfomonile_tiedjei_DSM_6799_uid168320	Yes
19	Dinoroseobacter_shibae_DFL_12_uid58707	Yes
20	Erwinia_amylovora_ATCC_49946_uid46943	Yes
21	Escherichia_coli_K_12_substr__MG1655_uid57779	Yes
22	Escherichia_coli_O157_H7_Sakai_uid57781	Yes
23	Gardnerella_vaginalis_ATCC_14019_uid55487	Yes
24	Gluconacetobacter_diazotrophicus_PAl_5_uid61587	Yes
25	Gluconobacter_oxydans_H24_uid179202	Yes
26	Granulibacter_bethesdensis_CGDNIH1_uid58661	Yes
27	Heliobacterium_modesticaldum_Ice1_uid58279	Yes
28	Kitasatospora_setae_KM_6054_uid77027	Yes
29	Marinithermus_hydrothermalis_DSM_14884_uid65783	Yes
30	Marinomonas_MWYL1_uid58715	Yes
31	Meiothermus_silvanus_DSM_9946_uid49485	Yes
32	Methylobacterium_nodulans_ORS_2060_uid59023	Yes
33	Methylococcus_capsulatus_Bath_uid57607	Yes
34	Methylomicrobium_alcaliphilum_uid77119	Yes
35	Nocardia_brasiliensis_ATCC_700358_uid86913	Yes
36	Oceanithermus_profundus_DSM_14977_uid60855	Yes
37	Pectobacterium_SCC3193_uid193707	Yes
38	Pelobacter_propionicus_DSM_2379_uid58255	Yes
39	Photobacterium_profundum_SS9_uid62923	Yes
40	Photorhabdus_luminescens_laumondii_TTO1_uid61593	Yes
41	Polymorphum_gilvum_SL003B_26A1_uid65447	Yes
42	Propionibacterium_acidipropionici_ATCC_4875_uid179069	Yes
43	Prosthecochloris_aestuarii_DSM_271_uid58151	Yes
44	Pseudogulbenkiania_NH8B_uid73423	Yes
45	Psychromonas_ingrahamii_37_uid58521	Yes
46	Rhodospirillum_centenum_SW_uid58805	Yes
47	Rhodothermus_marinus_DSM_4252_uid41729	Yes
48	Rubrivivax_gelatinosus_IL144_uid158163	Yes
49	Salmonella_enterica_serovar_Typhimurium_LT2_uid57799	Yes

50	Syntrophobacter_fumaroxidans_MPOB_uid58177	Yes
51	Thauera_MZ1T_uid58987	Yes
52	Thermus_thermophilus_HB8_uid58223	Yes
53	Thiocystis_violascens_DSM_198_uid74025	Yes
54	Truepera_radiovictrix_DSM_17093_uid49533	Yes
55	Verrucosispora_maris_AB_18_032_uid66297	Yes
56	Xenorhabdus_nematophila_ATCC_19061_uid49133	Yes
57	Acidimicrobium_ferrooxidans_DSM_10331_uid59215	No
58	Coriobacterium_glomerans_PW2_uid65787	No
59	Finegoldia_magna_ATCC_29328_uid58867	No
60	Methanocella_arvoryzae_MRE50_uid61623	No
61	Methanococcoides_burtonii_DSM_6242_uid58023	No
62	Methanosalsum_zhilinae_DSM_4017_uid68249	No
63	Methanosphaerula_palustris_E1_9c_uid59193	No
64	Methanospirillum_hungatei_JF_1_uid58181	No
65	Nakamurella_multipartita_DSM_44233_uid59221	No
66	Nocardiopsis_alba_ATCC_BAA_2165_uid174334	No
67	Roseiflexus_RS_1_uid58523	No
68	Saccharomonospora_viridis_DSM_43017_uid59055	No
69	Saccharothrix_espanaensis_DSM_44229_uid184826	No
70	Salinispora_arenicola_CNS_205_uid58659	No
71	Sphaerobacter_thermophilus_DSM_20745_uid41997	No
72	Symbiobacterium_thermophilum_IAM_14863_uid58165	No
73	Thermobaculum_terrenum_ATCC_BAA_798_uid42011	No
74	Thermobifida_fusca_YX_uid57703	No
75	Thermobispora_bispora_DSM_43833_uid48999	No
76	Thermomonospora_curvata_DSM_43183_uid41885	No

Table 6: IHF distribution among non-type I-E organisms

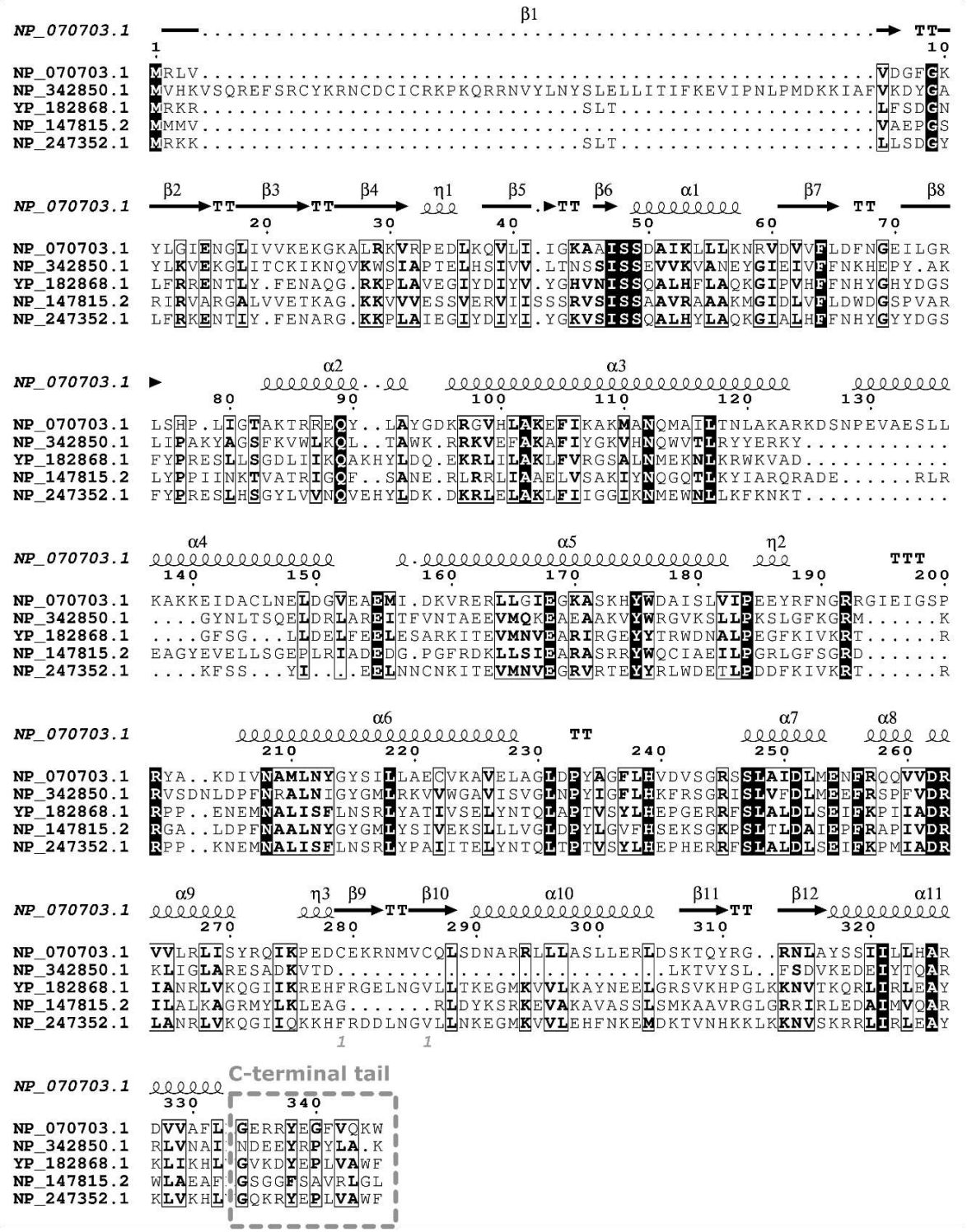
S/N	Organism Name	IHF presence
1	Acetobacterium_woodii_DSM_1030_uid88073	Yes
2	Acidovorax_avenae_ATCC_19860_uid42497	Yes
3	Aggregatibacter_actinomycetemcomitans_D7S_1_uid46989	Yes
4	Alicyclobacillus_acidocaldarius_Tc_4_1_uid158681	Yes
5	Aquifex_aeolicus_VF5_uid57765	Yes
6	Aromatoleum_aromaticum_EbN1_uid58231	Yes
7	Bibersteinia_trehalosi_192_uid193709	Yes
8	Caldisericum_exile_AZM16c01_uid158173	Yes
9	Calditerrivibrio_nitroreducens_DSM_19672_uid60821	Yes
10	Candidatus_Chloracidobacterium_thermophilum_B_uid73587	Yes
11	Candidatus_Methylomirabilis_oxyfera_uid161981	Yes
12	Cellvibrio_japonicus_Ueda107_uid59139	Yes
13	Chloroherpeton_thalassium_ATCC_35110_uid59187	Yes
14	Chromobacterium_violaceum_ATCC_12472_uid58001	Yes
15	Comamonadaceae_bacterium_CR_uid223378	Yes
16	Coprothermobacter_proteolyticus_DSM_5265_uid59253	Yes
17	Coxiella_burnetii_Dugway_5J108_111_uid58629	Yes
18	Deferribacter_desulfuricans_SSM1_uid46653	Yes
19	Delftia_acidovorans_SPH_1_uid58703	Yes
20	Desulfarculus_baarsii_DSM_2075_uid51371	Yes
21	Desulfobacca_acetoxidans_DSM_11109_uid65785	Yes
22	Desulfobacterium_autotrophicum_HRM2_uid59061	Yes
23	Desulfobacula_toluolica_Tol2_uid175777	Yes
24	Desulfobulbus_propionicus_DSM_2032_uid62265	Yes
25	Desulfocapsa_sulfexigens_DSM_10523_uid189952	Yes
26	Desulfohalobium_retbaense_DSM_5692_uid59183	Yes
27	Desulfotalea_psychrophila_LSv54_uid58153	Yes
28	Desulfovibrio_vulgaris_Hildenborough_uid57645	Yes
29	Desulfurispirillum_indicum_S5_uid45897	Yes
30	Desulfurivibrio_alkaliphilus_AHT2_uid49487	Yes
31	Dichelobacter_nodosus_VCS1703A_uid57643	Yes
32	Dickeya_dadantii_3937_uid52537	Yes
33	Ethanoligenens_harbinense_YUAN_3_uid46255	Yes
34	Flexistipes_sinusarabici_DSM_4947_uid68147	Yes
35	Frankia_EAN1pec_uid58367	Yes
36	Gallibacterium_anatis_UMN179_uid66567	Yes
37	Geobacter_metallireducens_GS_15_uid57731	Yes
38	Hahella_chejuensis_KCTC_2396_uid58483	Yes
39	Haliangium_ochraceum_DSM_14365_uid41425	Yes
40	Haliscomenobacter_hydrossis_DSM_1100_uid66777	Yes
41	Halomonas_elongata_DSM_2581_uid52781	Yes
42	Halothiobacillus_neapolitanus_c2_uid41317	Yes
43	Hydrogenobacter_thermophilus_TK_6_uid45927	Yes
44	Ignavibacterium_album_JCM_16511_uid162097	Yes
45	Isosphaera_pallida_ATCC_43644_uid62207	Yes
46	Ketogulonicigenium_vulgare_Y25_uid59581	Yes
47	Kosmotoga_olearia_TBF_19_5_1_uid59205	Yes
48	Laribacter_hongkongensis_HLHK9_uid59265	Yes
49	Leptothrix_cholodnii_SP_6_uid58971	Yes
50	Magnetococcus_MC_1_uid57833	Yes
51	Mannheimia_haemolytica_M42548_uid198769	Yes

52	Marinitoga_piezophila_KA3_uid81629	Yes
53	Megasphaera_elsdenii_DSM_20460_uid71135	Yes
54	Melioribacter_roseus_P3M_uid170941	Yes
55	Mesotoga_prima_MesG1_Ag_4_2_uid52599	Yes
56	Methylobacillus_flagellatus_KT_uid58049	Yes
57	Methylomonas_methanica_MC09_uid67363	Yes
58	Methylophaga_JAM1_uid162947	Yes
59	Moraxella_catarrhalis_BBH18_uid48809	Yes
60	Morganella_morganii_KT_uid180867	Yes
61	Myxococcus_xanthus_DK_1622_uid58003	Yes
62	Niastella_koreensis_GR20_10_uid83125	Yes
63	Nitrosococcus_halophilus_Nc4_uid46803	Yes
64	Odoribacter_splanchnicus_DSM_20712_uid63397	Yes
65	Parabacteroides_distasonis_ATCC_8503_uid58301	Yes
66	Parvularcula_bermudensis_HTCC2503_uid51641	Yes
67	Pasteurella_multocida_Pm70_uid57627	Yes
68	Pelodictyon_phaeoclathratiforme_BU_1_uid58173	Yes
69	Persephonella_marina_EX_H1_uid58119	Yes
70	Petrotoga_mobilis_SJ95_uid58747	Yes
71	Planctomyces_brasiliensis_DSM_5305_uid60583	Yes
72	Providencia_stuartii_MRSN_2154_uid162193	Yes
73	Pseudoxanthomonas_spadix_BD_a59_uid75113	Yes
74	Ramlibacter_tataouinensis_TTB310_uid68279	Yes
75	Rhodobacter_sphaeroides_KD131_uid59277	Yes
76	Rhodoferax_ferrireducens_T118_uid58353	Yes
77	Rhodomicrobium_vannielii_ATCC_17100_uid43247	Yes
78	Rothia_dentocariosa_ATCC_17931_uid49331	Yes
79	Ruminococcus_albus_7_uid51721	Yes
80	Runella_slithyformis_DSM_19594_uid68317	Yes
81	Singulisphaera_acidiphila_DSM_18658_uid81777	Yes
82	Sorangium_cellulosum_So0157_2_uid210741	Yes
83	Sphaerochaeta_pleomorpha_Grapes_uid82365	Yes
84	Spirochaeta_smaragdinae_DSM_11293_uid51369	Yes
85	Sulfurihydrogenibium_azorense_Az_Fu1_uid58121	Yes
86	Syntrophus_aciditrophicus_SB_uid58539	Yes
87	Tannerella_forsythia_ATCC_43037_uid83157	Yes
88	Teredinibacter_turnerae_T7901_uid59267	Yes
89	Thalassolituus_oleivorans_MIL_1_uid195604	Yes
90	Thermodesulfatator_indicus_DSM_15286_uid68285	Yes
91	Thermodesulfovibrio_yellowstonii_DSM_11347_uid59257	Yes
92	Thermosipho_africanus_TCF52B_uid59095	Yes
93	Thermotoga_maritima_MSB8_uid202924	Yes
94	Thermovibrio_ammonificans_HB_1_uid62095	Yes
95	Thioalkalivibrio_nitratireducens_DSM_14787_uid184011	Yes
96	Thioflavicoccus_mobilis_8321_uid184343	Yes
97	Tistrella_mobilis_KA081020_065_uid167486	Yes
98	Tolumonas_auensis_DSM_9187_uid59395	Yes
99	Verminephrobacter_eiseniae_EF01_2_uid58675	Yes
100	Xanthomonas_oryzae_PXO99A_uid59131	Yes
101	Xylanimonas_cellulosilytica_DSM_15894_uid41935	Yes
102	Yersinia_pestis_CO92_uid57621	Yes
103	Zunongwangia_profunda_SM_A87_uid48073	Yes
104	Zymomonas_mobilis_NCIMB_11163_uid41019	Yes
105	Acetohalobium_arabaticum_DSM_5501_uid51423	No
106	Acidianus_hospitalis_W1_uid66875	No

107	Aeropyrum_pernix_K1_uid57757	No
108	Alkaliphilus_metalliredigens_QYMF_uid58171	No
109	Aminobacterium_colombiense_DSM_12261_uid47083	No
110	Ammonifex_degensii_KC4_uid41053	No
111	Anabaena_cylindrica_PCC_7122_uid183339	No
112	Anaerobaculum_mobile_DSM_13181_uid168323	No
113	Anaerolinea_thermophila_UNI_1_uid62245	No
114	Anoxybacillus_flavithermus_WK1_uid59135	No
115	Archaeoglobus_fulgidus_DSM_4304_uid57717	No
116	Arcobacter_nitrofigilis_DSM_7299_uid49001	No
117	Bacillus_halodurans_C_125_uid57791	No
118	Caldicellulosiruptor_saccharolyticus_DSM_8903_uid58289	No
119	Caldilinea_aerophila_DSM_14535___NBRC_104270_uid158165	No
120	Caldivirga_maquilingensis_IC_167_uid58711	No
121	Calothrix_PCC_7507_uid182930	No
122	Candidatus_Arthromitus_SFB_mouse_Japan_uid71379	No
123	Candidatus_Desulforudis_audaxviator_MP104C_uid59067	No
124	Candidatus_Korarchaeum_cryptofilum_OPF8_uid58601	No
125	Candidatus_Nitrososphaera_gargensis_Ga9_2_uid176707	No
126	Carboxydothermus_hydrogenoformans_Z_2901_uid57821	No
127	Cenarchaeum_symbiosum_A_uid61411	No
128	Chamaesiphon_PCC_6605_uid183005	No
129	Chloroflexus_aggregans_DSM_9485_uid58621	No
130	Chroococcidiopsis_thermalis_PCC_7203_uid183002	No
131	Clostridium_difficile_630_uid57679	No
132	Crinalium_epipsammum_PCC_9333_uid183113	No
133	Cyanobacterium_PCC_10605_uid183340	No
134	Cyanothece_PCC_7822_uid52547	No
135	Cylindrospermum_stagnale_PCC_7417_uid183111	No
136	Desulfitobacterium_hafniense_Y51_uid58605	No
137	Desulfosporosinus_orientis_DSM_765_uid82939	No
138	Desulfotomaculum_gibsoniae_DSM_7213_uid76945	No
139	Desulfurococcus_kamchatkensis_1221n_uid59133	No
140	Dictyoglomus_thermophilum_H_6_12_uid59439	No
141	Eggerthella_lenta_DSM_2243_uid59079	No
142	Eubacterium_limosum_KIST612_uid59777	No
143	Faecalibacterium_prausnitzii_L2_6_uid197183	No
144	Ferroglobus_placidus_DSM_10642_uid40863	No
145	Ferroplasma_acidarmanus_fer1_uid54095	No
146	Fervidicoccus_fontis_Kam940_uid162201	No
147	Filifactor_alocis_ATCC_35896_uid46625	No
148	Fusobacterium_nucleatum_ATCC_25586_uid57885	No
149	Geitlerinema_PCC_7407_uid183007	No
150	Gloeocapsa_PCC_7428_uid183112	No
151	Halanaerobium_hydrogeniformans_uid60191	No
152	Haloarcula_marismortui_ATCC_43049_uid57719	No
153	Halobacteroides_halobius_DSM_5150_uid184862	No
154	Haloferax_volcanii_DS2_uid46845	No
155	Halomicrobium_mukohataei_DSM_12286_uid59107	No
156	Haloquadratum_walsbyi_C23_uid162019	No
157	Halorhabdus_tiamatea_SARL4B_uid214082	No
158	Halorubrum_lacusprofundi_ATCC_49239_uid58807	No
159	Halothece_PCC_7418_uid183338	No
160	Halothermothrix_orenii_H_168_uid58585	No
161	Herpetosiphon_aurantiacus_DSM_785_uid58599	No

162	Hyperthermus_butylicus_DSM_5456_uid57755	No
163	Ignicoccus_hospitalis_KIN4_I_uid58365	No
164	Ignisphaera_aggregans_DSM_17230_uid51875	No
165	Kyrpidia_tusciae_DSM_2912_uid48361	No
166	Leptotrichia_buccalis_C_1013_b_uid59211	No
167	Lysinibacillus_sphaericus_C3_41_uid58945	No
168	Mahella_australiensis_50_1_BON_uid66917	No
169	Megamonas_hypermegale_uid197163	No
170	Melissococcus_plutonius_ATCC_35311_uid66803	No
171	Metallosphaera_sedula_DSM_5348_uid58717	No
172	Methanobrevibacter_ruminantium_M1_uid45857	No
173	Methanocaldococcus_jannaschii_DSM_2661_uid57713	No
174	Methanocorpusculum_labreanum_Z_uid58785	No
175	Methanoculleus_bourgensis_MS2_uid171377	No
176	Methanomassiliicoccus_Mx1_Issoire_uid207287	No
177	Methanoregula_formicicum_SMSP_uid184406	No
178	Methanosaeta_concilii_GP6_uid66207	No
179	Methanosarcina_acetivorans_C2A_uid57879	No
180	Methanosphaera_stadtmanae_DSM_3091_uid58407	No
181	Methanothermobacter_thermautotrophicus_Delta_H_uid57877	No
182	Methanothermococcus_okinawensis_IH1_uid51535	No
183	Methanothermus_fervidus_DSM_2088_uid60167	No
184	Methanotorris_igneus_Kol_5_uid67321	No
185	Microcoleus_PCC_7113_uid183114	No
186	Microcystis_aeruginosa_NIES_843_uid59101	No
187	Moorella_thermoacetica_ATCC_39073_uid58051	No
188	Nanoarchaeum_equitans_Kin4_M_uid58009	No
189	Natrinema_J7_uid171337	No
190	Natronobacterium_gregoryi_SP2_uid74439	No
191	Natronomonas_pharaonis_DSM_2160_uid58435	No
192	Nitratifractor_salsuginis_DSM_16511_uid62183	No
193	Nostoc_PCC_7120_uid57803	No
194	Nostoc_punctiforme_PCC_73102_uid57767	No
195	Oscillatoria_PCC_7112_uid183110	No
196	Pelotomaculum_thermopropionicum_SI_uid58877	No
197	Phycisphaera_mikurensis_NBRC_102666_uid157331	No
198	Picrophilus_torridus_DSM_9790_uid58041	No
199	Pleurocapsa_PCC_7327_uid183006	No
200	Porphyromonas_gingivalis_TDC60_uid67407	No
201	Pseudanabaena_PCC_7367_uid183004	No
202	Psychrobacter_G_uid210641	No
203	Pyrobaculum_aerophilum_IM2_uid57727	No
204	Pyrococcus_abyssi_GE5_uid62903	No
205	Pyrococcus_furiosus_DSM_3638_uid57873	No
206	Pyrolobus_fumarii_1A_uid73415	No
207	Rivularia_PCC_7116_uid182929	No
208	Rubrobacter_xylanophilus_DSM_9941_uid58057	No
209	Selenomonas_ruminantium_lactilytica_TAM6421_uid157247	No
210	Spirosoma_linguale_DSM_74_uid43413	No
211	Stackebrandtia_nassauensis_DSM_44728_uid46663	No
212	Stanieria_cyanosphaera_PCC_7437_uid183115	No
213	Streptococcus_pyogenes_M1_GAS_uid57845	No
214	Sulfobacillus_acidophilus_TPY_uid68841	No
215	Sulfolobus_acidocaldarius_DSM_639_uid58379	No
216	Sulfolobus_solfataricus_P2_uid57721	No

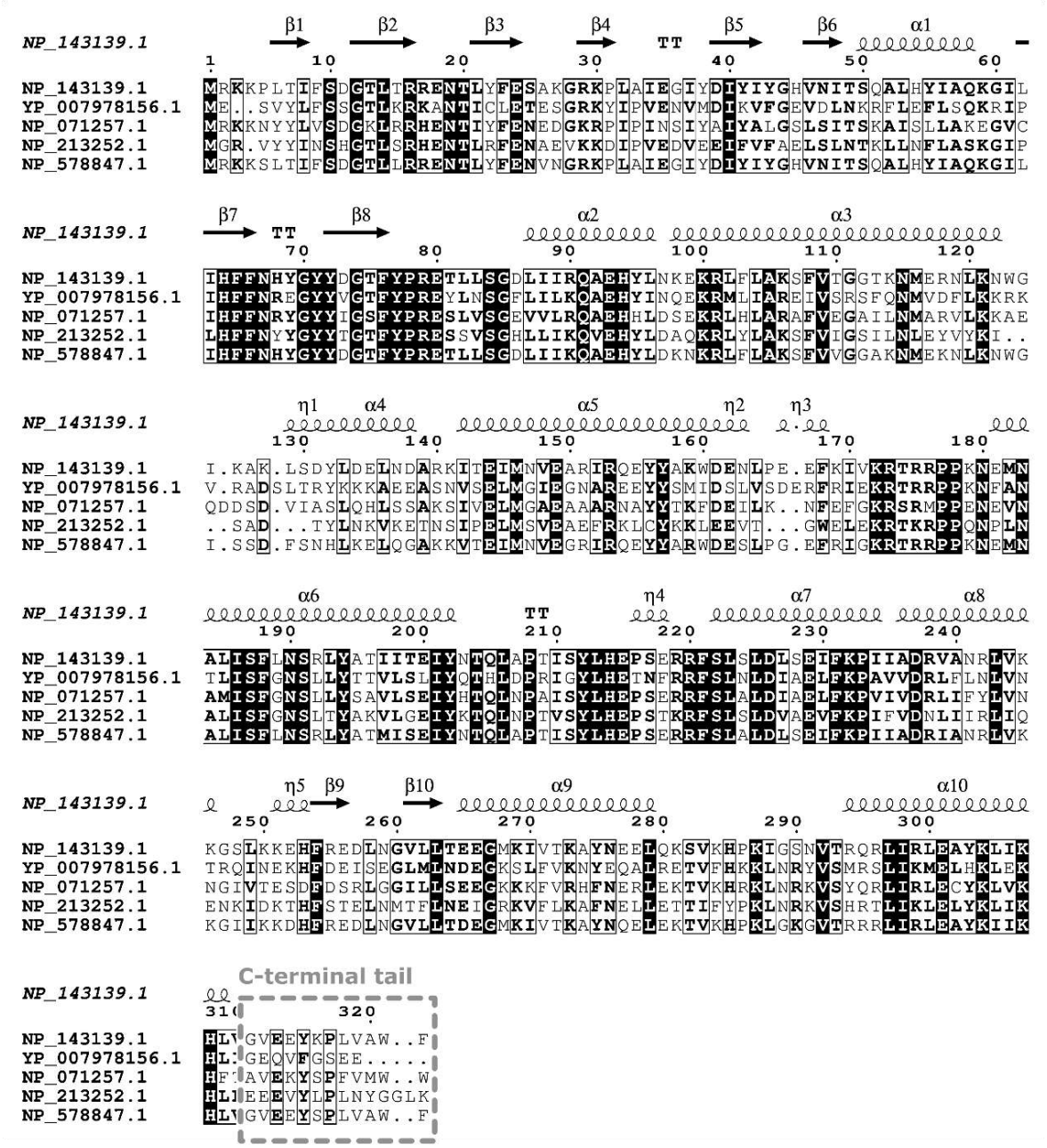
217	Sulfurospirillum_barnesii_SES_3_uid168117	No
218	Synechococcus_PCC_6312_uid182934	No
219	Synechocystis_PCC_6803_uid57659	No
220	Synergistetes_bacterium_SGP1_uid197182	No
221	Syntrophobotulus_glycolicus_DSM_8271_uid63343	No
222	Syntrophomonas_wolfei_Goettingen_uid58179	No
223	Syntrophothermus_lipocalidus_DSM_12680_uid49527	No
224	Tepidanaerobacter_acetatoxydans_Re1_uid184827	No
225	Thermacetogenium_phaeum_DSM_12270_uid177811	No
226	Thermaerobacter_marianensis_DSM_12885_uid61727	No
227	Thermincola_potens_JR_uid48823	No
228	Thermoanaerobacter_tengcongensis_MB4_uid57813	No
229	Thermoanaerobacterium_thermosaccharolyticum_M0795_uid184821	No
230	Thermobacillus_composti_KWC4_uid74021	No
231	Thermococcus_kodakarensis_KOD1_uid58225	No
232	Thermocrinis_albus_DSM_14484_uid46231	No
233	Thermodesulfobacterium_OPB45_uid68283	No
234	Thermodesulfobium_narugense_DSM_14796_uid66601	No
235	Thermofilum_1910b_uid215374	No
236	Thermogladius_1633_uid167488	No
237	Thermomicrobium_roseum_DSM_5159_uid59341	No
238	Thermoproteus_tenax_Kra_1_uid74443	No
239	Thermosediminibacter_oceani_DSM_16646_uid51421	No
240	Thermosphaera_aggregans_DSM_11486_uid48993	No
241	Thermovirga_lienii_DSM_17291_uid77129	No
242	Vulcanisaeta_distributa_DSM_14429_uid52827	No



Appendix Figure 1: Multiple sequence alignment of Cas1/I-A

A representative sequence alignment of Cas1 derived from *Archaeoglobus fulgidus* DSM 4304 (NP_070703.1), *Sulfolobus solfataricus* P2 (NP_342850.1), *Thermococcus kodakarensis* KOD1 (YP_182868.1), *Aeropyrum pernix* K1 (NP_147815.2) and *Methanocaldococcus jannaschii* DSM 2661 (NP_247352.1) is displayed. Using Cas1 structure (PDB ID: 4N06) from *A. fulgidus* DSM

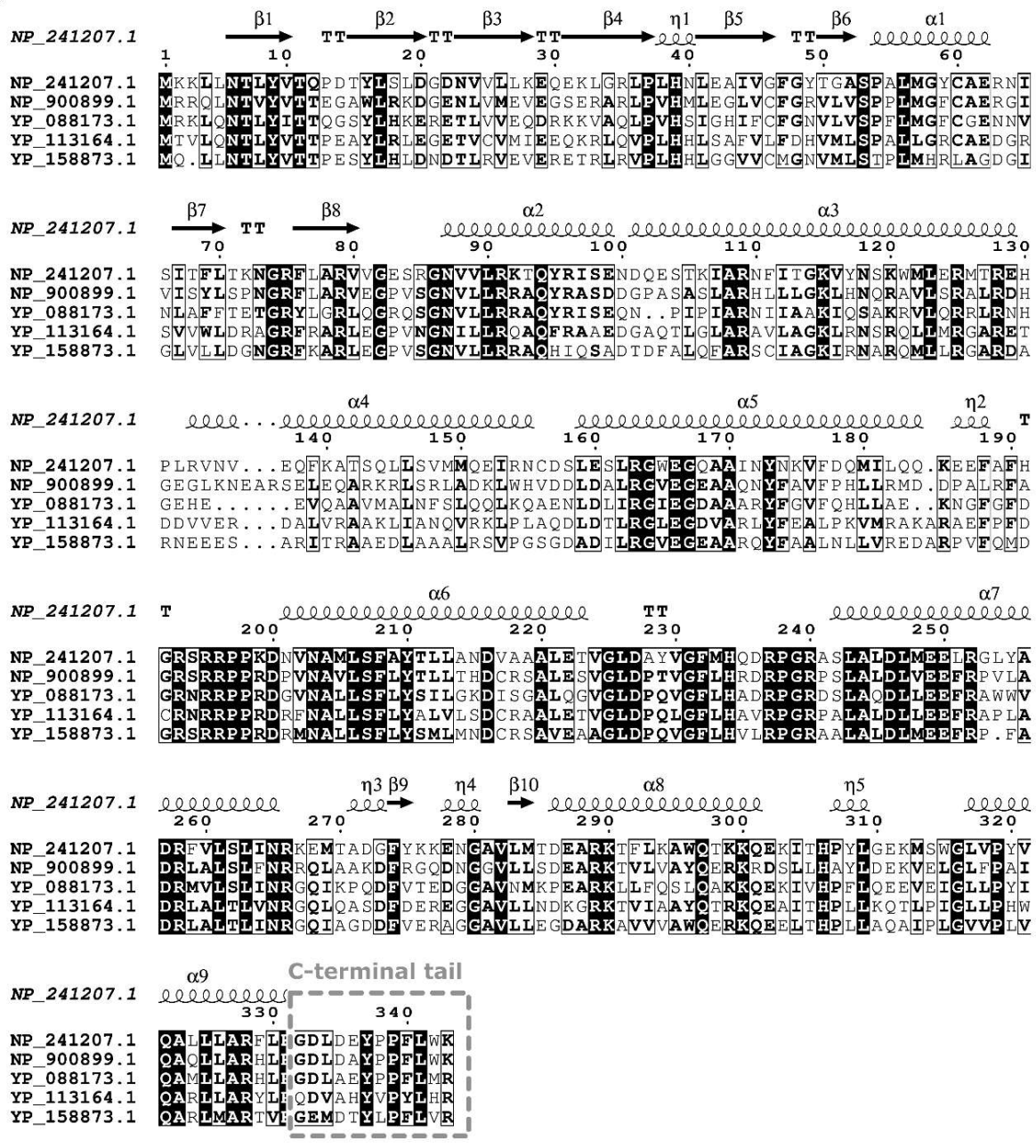
4304 as a reference, positions of various secondary structural elements were mapped onto the sequence alignment by ESpript (Robert and Gouet, 2014). Region corresponding to the C-terminal tail is displayed in Grey box (dotted border), whereas, secondary structural features such as alpha helix (α), beta strands (β), 3_{10} helices (η) and beta turns (TT) are depicted at the predicted positions. Amino acid residues that are completely conserved are highlighted in Black shade with White font, whereas, partially conserved residues are boxed and depicted in Black bolded font.



Appendix Figure 2: Multiple sequence alignment of Cas1/I-B

A representative sequence alignment of Cas1 derived from *Pyrococcus horikoshii* OT3 (NP_143139.1), *Thermotoga maritima* MSB8 (YP_007978156.1), *Archaeoglobus fulgidus* DSM 4304 (NP_071257.1), *Aquifex aeolicus* VF5 (NP_213252.1) and *Pyrococcus furiosus* DSM 3638 (NP_578847.1) is displayed. Using Cas1 structure (PDB ID: 4WJ0) from *P. horikoshii* OT3 as a reference, positions of various secondary structural elements were mapped onto the sequence alignment by ESpript (Robert and Gouet, 2014). Region corresponding to the C-terminal tail is displayed in Grey box (dotted border), whereas, secondary structural features such as alpha helix (α), beta strands (β), 3_{10} helices (η) and beta turns (TT) are depicted at the predicted positions. Amino

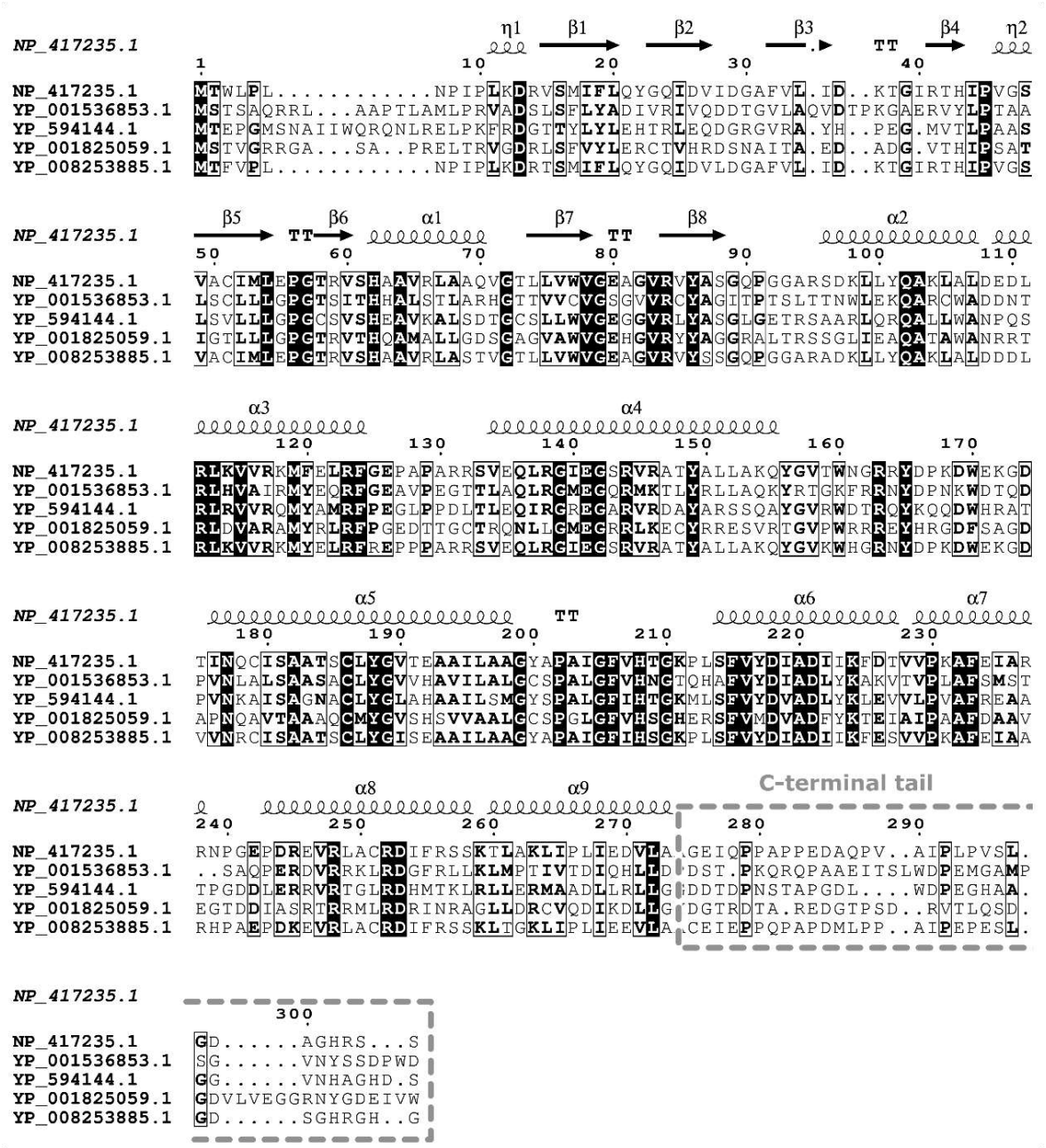
acid residues that are completely conserved are highlighted in Black shade with White font, whereas, partially conserved residues are boxed and depicted in Black bolded font.



Appendix Figure 3: Multiple sequence alignment of Cas1/I-C

A representative sequence alignment of Cas1 derived from *Bacillus halodurans* C-125 (NP_241207.1), *Chromobacterium violaceum* ATCC 12472 (NP_900899.1), *Mannheimia succiniciproducens* MBEL55E (YP_088173.1), *Methylococcus capsulatus* str. Bath (YP_113164.1) and *Aromatoleum aromaticum* EbN1 (YP_158873.1) is displayed. Using Cas1 structure (predicted using I-TASSER) from *B. halodurans* C-125 as a reference, positions of various secondary structural elements were mapped onto the sequence alignment by ESpript (Robert and Gouet, 2014). Region corresponding to the C-terminal tail is displayed in Grey box (dotted border), whereas, secondary structural features such as alpha helix (α), beta strands (β), 3_{10} helices (η) and beta turns (TT) are depicted at the predicted positions. Amino acid residues that are completely

conserved are highlighted in Black shade with White font, whereas, partially conserved residues are boxed and depicted in Black bolded font.



Appendix Figure 4: Multiple sequence alignment of Cas1/I-E

A representative sequence alignment of Cas1 derived from *E. coli* str. K-12 substr. MG1655 (NP_417235.1), *Salinispora arenicola* CNS-205 (YP_001536853.1), *Deinococcus geothermalis* DSM 11300 (YP_594144.1), *Streptomyces griseus* subsp. griseus NBRC 13350 (YP_001825059.1) and *Salmonella enterica* subsp. enterica serovar Typhimurium var. 5- str. CFSAN001921 (YP_008253885.1) is displayed. Using Cas1 structure (PDB ID: 5DQZ) from *E. coli* str. K-12 substr. MG1655 as a reference, positions of various secondary structural elements were mapped onto the sequence alignment by ESpript (Robert and Gouet, 2014). Region corresponding to the C-terminal tail is displayed in Grey box (dotted border), whereas, secondary structural features such as alpha helix (α), beta strands (β), 3_{10} helices (η) and beta turns (TT) are depicted at the predicted

positions. Amino acid residues that are completely conserved are highlighted in Black shade with White font, whereas, partially conserved residues are boxed and depicted in Black bolded font.

List of publications

1. **Yoganand KNR**, Sivathanu R, Nimkar S, Anand B (2017). Asymmetric Positioning of Cas1-2 Complex and Integration Host Factor Induced DNA Bending Guide the Unidirectional Homing of Protospacer in CRISPR-Cas Type I-E system. *Nucleic Acids Res.* 45: 367-381
2. Punetha A, **Yoganand KNR**, Nimkar S, Anand B (2018). Cutting it Right: Plasticity and Strategy of CRISPR RNA Specific Nucleases. *Proceedings of the Indian National Science Academy.* 84: 455-477
3. **Yoganand KNR**, Manasasri M, Nimkar S, Anand B (2019). Fidelity of prespacer capture and processing is governed by the PAM mediated Interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system. *J. Biol. Chem*. 294: 20039-20053.

List of posters presented in conferences

1. **Yoganand KNR**, Anand B (2014). "Molecular characterization of CRISPR adaptation in bacteria". International proteomics conference. Indian Institute of Technology, Mumbai, India, December 7th – 9th, 2014.
2. **Yoganand KNR** and Anand B (2017) "Integration host factor mediated restructuring of CRISPR leader guides the polarized expansion of CRISPR array by deployment of Cas integration complex" 9th RNA Group Meet, Banaras Hindu University, Varanasi, India, October 26th – 28th, 2017.