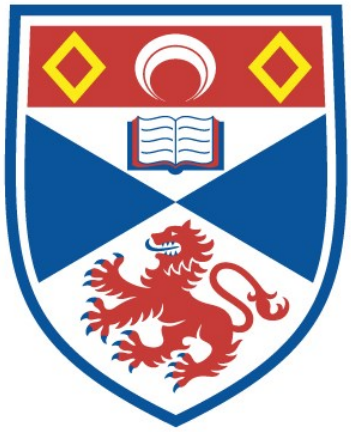
The type I-G CRISPR system: mechanism, structure and application

Qilin Shangguan

A thesis submitted for the degree of PhD
at the
University of St Andrews



2023

Full metadata for this thesis is available in
St Andrews Research Repository
at:
https://research-repository.st-andrews.ac.uk/

Identifier to use to cite or link to this thesis:

DOI: https://doi.org/10.17630/sta/645

This item is protected by original copyright

This item is licensed under a
Creative Commons License

https://creativecommons.org/licenses/by-sa/4.0

Candidate's declaration

I, Qilin Shangguan, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 40,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2019.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 10/07/2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 10/07/2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Qilin Shangguan, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 10/07/2023 Signature of candidate

Date 10/07/2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Qilin Shangguan, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date 10/07/2023 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date 10/07/2023 Signature of candidate

Date 10/07/2023 Signature of supervisor

Table of Contents

Figures and tables

Abbreviations

a-EJ	Alternative end joining
Abi	abortive infection
Acr	anti-CRISPR proteins
ATTR	transthyretin amyloidosis
cA3	cyclic tri-adenylate
cas	CRISPR associated
Cascade	CRISPR-associated complex for antiviral defence
CBASS	cyclic oligonucleotide-based anti-phage signalling system
cGAMP	cyclic GMP–AMP
cOA	cyclic oligoadenylate
CRISPR	clustered regularly interspaced palindromic repeats
crRNA	CRISPR RNA
dCas9	catalytic dead Cas9
ddhCTP	3′-deoxy-3′,4′-didehydro (ddh)-cytidine triphosphate
ddhGTP	3′-deoxy-3′,4′-didehydro (ddh)-guanosine triphosphate
ddhUTP	3′-deoxy-3′,4′-didehydro (ddh)-uridine triphosphate
DSB	dsDNA break
gRNA	guide RNA
HDR	Homology-directed repair
LOAD	late-onset Alzheimer's disease
MGEs	mobile genetic elements
MMEJ	microhomology-mediated end joining
ncRNA	non-coding RNA
NHEJ	nonhomologous end joining
NTP	nucleoside triphosphate
OMVs	outer membrane vesicles
PAC	primed acquisition complex
pAgo	prokaryotic Argonaute
PAM	protospacer-adjacent motif
PAMPs	pathogen-associated molecular patterns
pegRNA	primer editing gRNA
PICIs	Phage-inducible chromosomal islands
PNPase	polynucleotide phosphorylase

pVips	Prokaryotic viperins
Pycsar	pyrimidine cyclase system for antiphage resistance
R-M	restriction-modification
RAMPs	Repair Associated Mysterious Proteins
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNAP	RNA polymerase
RPA	replication protein A
RT	reverse transcriptase
SCD	sickle cell disease
scoutRNA	short-complementarity untranslated RNA
Sie	superinfection exclusion
SpCas9	Cas9 from Streptococcus pyogenes
SRSRs	short regularly spaced repeats
ssODNs	single-strand oligodeoxyribonucleotides
TDT	transfusion-dependent β-thalassemia
TIR	Toll/interleukin-1 receptor
tracrRNA	transactivating crRNA

Acknowledgements

Four years ago, I came to St Andrews 8669.23 km (5386.81 mi) away from my hometown without burying a single thing under a cherry tree. Professor Malcolm White met me. I insisted to call him Professor White back then under my profoundly inscribed view on the righteous manner of calling your supervisor. Time goes, and at the end of my PhD, I get used to calling his name Malcolm, and he becomes not only my supervisor but a friend who supports me in going through life puzzles as well. I respect his enthusiasm that encourages me when I was disappointed by failures, and I am amazed that he can always spot the right direction we need to pursue. My heartfelt thanks go to him for not giving up on me who hesitated to forward at the life cross and for giving me patience and time.

Dr Shirley Graham guided me through the tough beginning in the lab and consistently supports my lab work with her expertise. Dr Sabine Grüschow, her well-rounded protocols make a sophisticated lab process understandable, following her guidance, I overcame so many experimental obstacles. Dr Gaëlle Hogrel initialized the type I-G structure research by contacting Dr Ramasubramanian Sundaramoorthy, who beautifully performed the structure afterwards. My gratitude goes to them for helping to complete the project. I also want to thank Dr Reyes Sanles-Falagan, Dr Rémi Fritzen, Dr Januka Athukoralage and Dr Wenlong Zhu, the previous members of our lab, who gave me adequate instructions in the lab. Ms Haotian Chi, Ms Laura Gaskell-mew, Dr Larissa Krüger, Mr Cooper Brown, Dr Ville Hoikkala and Mr Stuart McQuarrie, the current members of our lab, it is my pleasure, and I am grateful to study and work with them. Thanks to BSRC staff, they built a comfortable platform for us to do research.

My family and friends are always here to support me. My gratitude to them can never be too much. Thanks to my roommate Mr Xupeng Yu, who can stand up with a night owl. Ms Anna Zolotariof, Mr Fillmon Kubrom, Ms Janie Olver and Ms Berta Fatas, it is my luck to meet them.

Studying and living in Scotland, even it is only a short period in my life, changes the way I think, and the life I want to live. The motivation comes from the people I meet and maybe the place I live, where a cherry tree blossoms under the blue sky.

They told me to quote a poem at the end of acknowledgements. Then I did it.

O were my love yon Lilac fair

Wi' purple blossoms to the Spring

And I, a bird to shelter there

When wearied on my little wing

How I wad mourn when it was torn

By Autumn wild, and Winter rude

But I wad sing on wanton wing

When youthfu' May its bloom renew'd

***-Robert Burns**, O were my love yon Lilac fair*

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (REF: BB/S000313/1 to MFW),

This work was supported by the Medical Research Council (REF: MR/S021647/1 to RS)

This work was supported by the China Scholarship Council (REF: 202008060345 to QS.

Research Data/Digital Outputs access statement

Research data underpinning this thesis are available at:

https://doi.org/10.17630/d54ca921-ef39-44bc-8522-4696f2234947

Abstract

CRISPR is originally discovered as an adaptive immune system in prokaryotes. It has been widely repurposed for application in different microbiological fields attributed to its ability to target DNA or RNA in a sequence-specific manner. But there is always an uncharted area of CRISPR systems in nature, awaiting exploration.

The type I-G CRISPR system is one of the subtypes of type I CRISPR systems with a multi-subunit effector complex compared to type II CRISPR-Cas9, a single subunit effector. Characterised by the enigmatic *cas* proteins Csb2 and Cas8g, type I-G system is the least understood type I system and possesses a unique mechanism in CRISPR recognition and interference.

In this thesis, we expressed and reconstructed a type I-G system from *Thioalkalivibrio sulfidiphilus*. We present key insights into the biochemistry and mechanism of the system, and a first view of the structure of the effector complex of type I-G is provided. Heterologous expression of type I-G in *Escherichia coli* provides immunity against mobile genetic elements. Repurposing type I-G for genome editing in *E. coli* with atypical Cas3 generates desirable editing. These observations provide an overview of the type I-G system, potentiating fundamental studies and further applications.

Chapter 1: Introduction

1.1 Bacterial defence systems

Biological defence systems establish a balance between the integration and exclusion of foreign elements. Many species have elaborate defence mechanisms to protect them from invasive mobile genetic elements (MGEs). But MGEs find ways to circumvent these, ingeniously avoiding host immune clearance. Viruses need host cells to live. Therefore, genetic elements carried by viruses consistently invade host cells. In some cases, these genetic elements become part of the host genome, even improving host survival and propagation. Genetic elements of lysogenic phages, for instance, integrated into the host bacterial genome, produce immunity to lytic forms of the phages[1].

Other MGEs, however, take materials and energy from host cells for their own benefits. With the burden imposed by viruses, host cells may undergo dysfunction or even cell death. To strike the balance between the intake and exclusion of MGEs, different layers of immunity systems are deployed to protect host cells. In eukaryotes, innate immunity and adaptive immunity are two primary immune strategies. Invasive nucleic elements are recognized as pathogen-associated molecular patterns (PAMPs) by pattern recognition receptors, triggering immune responses[2]. Innate dendritic cells and natural killer cells clear the infectious cells. Moreover, antibodies produced by adaptive immune system specifically eliminate pathogens and form an immune memory[3-5]. Prokaryotic immunity is different from eukaryotic systems and is specially introduced below.

1.1.1 Overview of prokaryotic defence system

Bacteriophage interacting with bacterial defence systems is a representative model to learn prokaryotic immunity. Phages, as for all other viruses, are dependent on host cells to replicate. Phage infection begins with phage attachment to the bacterial surface, injecting phage genetic elements, DNA, into bacteria cells. Following phage DNA replication and protein expression, assembled phage progeny continue to infect other bacterial cells[6]. Bacterial defences target all phases of phage life cycle from preventing adsorption, blocking injection to inhibiting phage replication. Also, clearance of phage DNA is carried out by defence systems, including restriction-modification (R-M) and CRISPR-Cas systems. Furthermore, drastic immune defences lead to abortive infection (Figure 1.1)[7, 8].

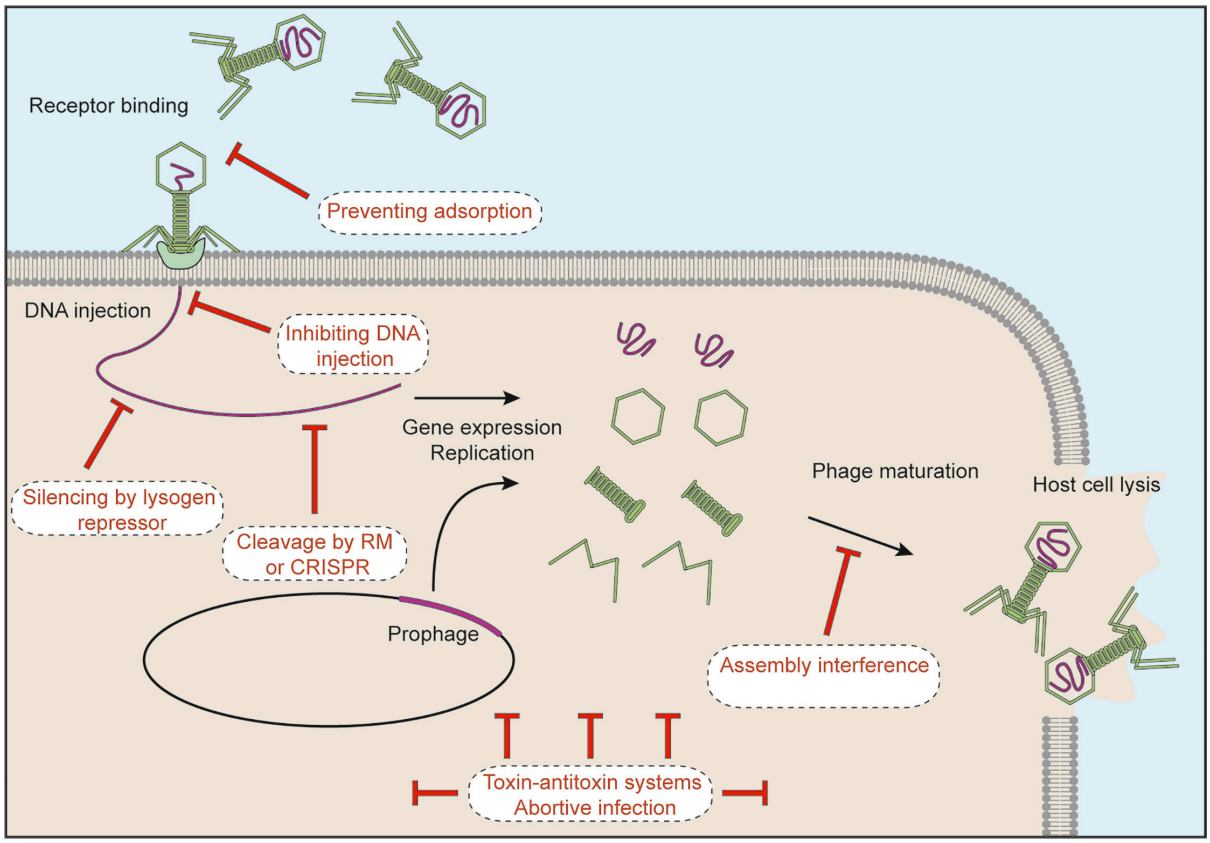


Figure 1.1 Phage life cycle and bacteria defence

Bacterial defence strategies target different phases of phage life cycle. Phage life cycle starts from either direct DNA injection upon attaching the host cell surface or prophage (phage genome that is integrated into host chromosome) induction. Following phage DNA replication and expression, phage particles assembled and released from host cells. Anti-viral defences target each stage of phage's life cycle. Adapted from Rostøl et al., 2019[7].

1.1.2 Blocking adsorption and injection

The start of phage infection is adsorption to the bacterial cell surface. In fact, this surface is a frontline in bacterial immunity. Bacterial cells are protected by biofilms, a physical structure that blocks phage entry[9]. Gram-negative bacteria secrete outer

membrane vesicles (OMVs), acting as decoys to bind extracellular phages. It was reported that *Vibrio cholerae* OMVs inhibit phage infection by binding phages[10].

Receptors on the surface of bacterial cells are the door to the cell interior for phages. Bacteria can hide their surface proteins to block the phage attachment. For example, *Pseudomonas aeruginosa* protects itself from phages through modification on surface receptors. Phages, using type IV pilus protein as a receptor, can be blocked due to glycosylation on the pilus[11]. More interestingly, *E. coli* lytic phage T5, which enters cells through the FhuA protein on the cell surface, encodes a lipoprotein that blocks other phages using FhuA protein from binding to the cell surface. This phenomenon of T5 phage protects bacteria from subsequent infection by other phages, an example of "superinfection exclusion" (Sie)[12]. Receptors can also be modified by a gene mutation. A typical example is that phage lambda adsorption is influenced by lamB (receptors of phage lambda) mutations in *E. coli*[13].

Phage DNA injection can also be targeted. This defence mechanism is always found in superinfection exclusion. Mycobacteriophage Fruitloop expresses a protein gp52, which inactivates Wag31. Wag31, a *Mycobacterium smegmatis* protein controlling cell surface biosynthesis during cellular pole growth, is crucial for Subcluster B2 phages' DNA injection. The inactivation of Wag31 from Fruitloop protects cells from heterotypic phage infection[14].

Once phage escapes the surface defence, innate immune systems targeting phage DNA or interfering phage replication can be activated to protect bacterial cells.

1.1.3 Bacterial immunity

Once phage has successfully attached to cell surface and injected genome into host bacterial cells, bacteria responded by a diversity of bacterial immunity. Immunity of bacteria generally utilizes three different ways to restrict phage propagation: I. Degradation of phage nucleic acids; II. Abortive infection; III. Inhibition of DNA and RNA synthesis (Figure 1.2)[15]. Various defence machinery is produced to counter phage infection with different strategies.

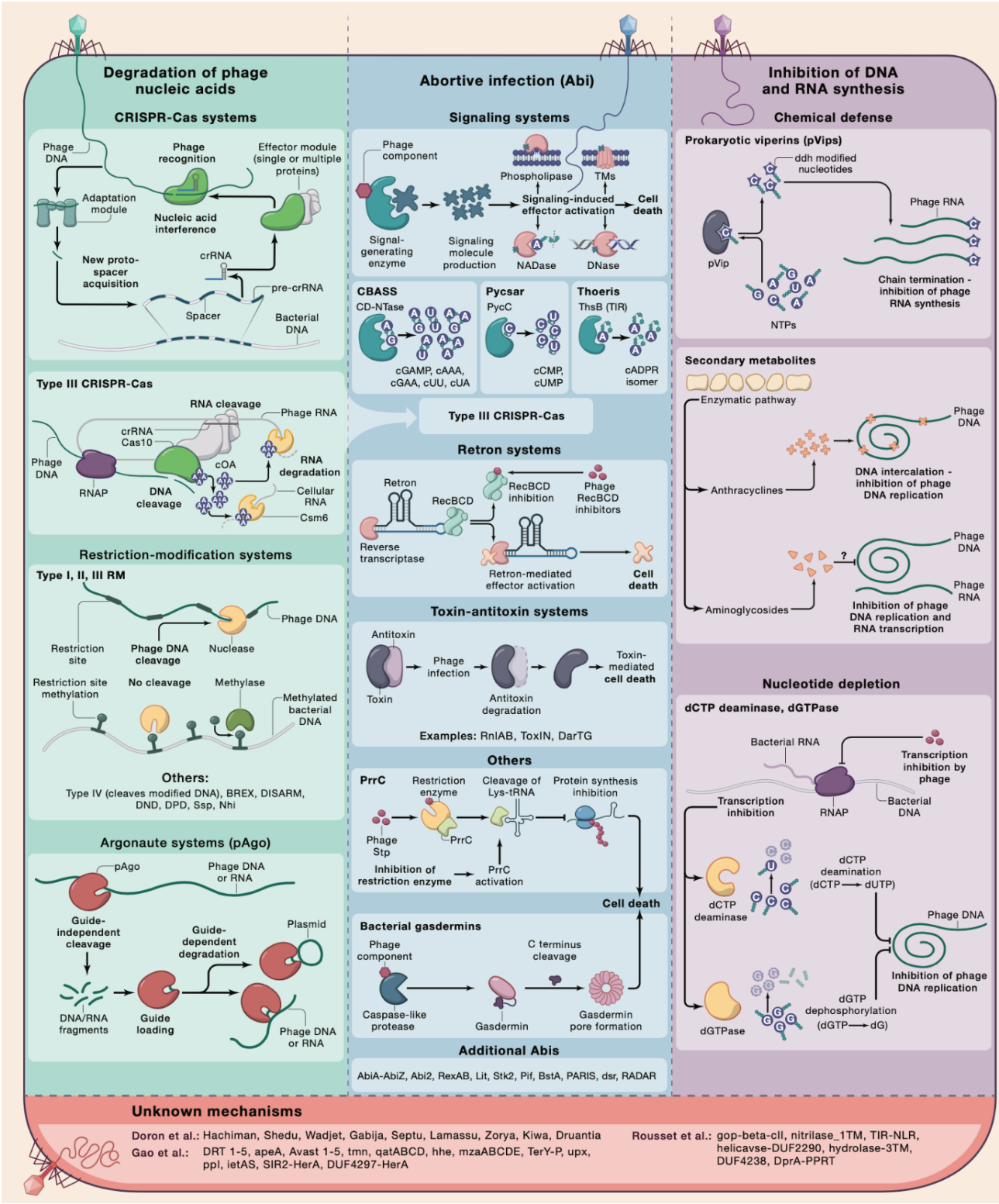


Figure legends next page.

Figure 1.2 The diversity of bacterial immunity.

Green panel (Degradation of phage nucleic acids):

CRISPR systems capture phage DNA by adaptation model and incorporate it into bacterial genome. This incorporated DNA fragment, Spacer, is transcribed with repeat sequence on genome into pre-CRISPR RNA (pre-crRNA), following maturation, pre-crRNA is cleaved into crRNA. crRNA together with CRISPR associated protein (*cas* protein) formed the effector module. The effector specifically targeted and degraded phage DNA by RNA guide. Type III CRISPR effectors recognise phage RNA, besides target nucleic acid cleavage, and generate cyclic oligoadenylate (cOA) as a second signal messenger to activate downstream effector, Csm6 for example. Both phage RNA and cellular RNA are degraded by the effector, resulting in cell growth arrest or cell death. The CRISPR system is deeply elucidated in **section 1.2**.

Restriction-modification systems, phage DNA with restriction sites is cleaved by nuclease, however, host bacterial DNA is modified by methylase, protected from nuclease cleavage. Details in **section 1.1.3.1**.

Argonaute systems, at first, prokaryotic Argonaute (pAgo) non-specifically cleaves phage DNA or RNA, generating DNA/RNA fragments that are being used for guide-dependent degradation. Invasive plasmid, phage DNA or RNA is subsequently degraded by Argonaute system. Details in **section 1.1.3.2**.

Blue panel (Abortive infection, details in **section 1.1.3.3**):

Signalling systems, phage components are sensed by signal-generating enzymes, producing signalling molecules. Different effectors, including Phospholipase, transmembrane proteins, NADase and DNase, are activated by the signalling molecule, causing cell death. Signal-generating enzymes and signalling molecules differ in cyclic oligonucleotide-based anti-phage signalling system (CBASS), Pycsar and Thoeris.

Retron systems, Retron, a DNA-RNA hybrid with reverse transcriptase, “guards” RecBCD. When RecBCD is inhibited by phage protein, Retron and effector are activated, resulting in cell death.

Toxin-antitoxin system, toxin component is repressed by antitoxin, phage infection degrades antitoxin, releasing toxins, leading to toxin-mediated cell death.

PrrC system, phage encodes a short peptide, called Stp, to inhibit restriction enzyme in R-M system, which activates PrrC that cleaves Lys-tRNA. Protein synthesis is aborted, causing cell death.

Bacterial gasdermins, phage component activates Caspase-like protease that removes a C-terminal peptide of gasdermin, leading to a gasdemin pore formation that disrupts the integrity of cell membrane, subsequently causing cell death.

Purple panel (Inhibition of DNA and RNA synthesis, details in **section 1.1.3.4**):

Prokaryotic viperins (pVips) produce 3′-deoxy-3′,4′-didehydro (ddh)-cytidine triphosphate (ddhCTP), acting as a chain terminator of Phage RNA synthesis.

Secondary metabolites, anthracyclines inhibit phage DNA replication by DNA intercalation. Aminoglycosides inhibit phage DNA replication and RNA transcription with an unknown mechanism.

Nucleotide depletion, transcription is inhibited by phage, resulting in dCTP deaminase or dGTPase production, eliminating nucleotide for DNA replication.

Red panel, more defence systems with unknown mechanism is described in **section 1.1.4**.

Adapted from Nitzan Tal and Rotem Sorek[15].

1.1.3.1 Restriction-modification systems

Restriction-modification (R-M) systems were found as a bacterial immune strategy against invasive nucleic acid in 1950s[16]. R-M systems are found in 74% of prokaryotic genomes[17]. In general, the R-M systems include two major components. R stands for restriction endonuclease. M is short for modification, but in most cases, it stands for methyltransferase. Restriction endonucleases can recognize DNA motifs and degrade the DNA, but both host and phage DNA possess these motifs. The modification component marks the DNA through methylation, separating self-DNA from phage DNA. R-M systems modify and degrade target DNA by different mechanisms, classified into four types. In type I, type II and type III R-M systems, host self-DNA is methylated, and non-self DNA without methylation is cleaved. In type IV, non-self DNA is modified while the host DNA remains unaltered, and modified phage DNA is targeted by restriction component[18]. For example, type I R-M enzymes translocate along bacterial DNA. Bacterial DNA is recognized by the methylation state. Enzymes recognize fully or hemi-methylated DNA as self-DNA and modify the hemi-methylated DNA to fully methylated DNA. In contrast, invasive phage DNA without single methylation is cleaved by restriction enzymes distant from the recognition sites[19]. It was also shown that modifications other than methylation can be used to distinguish self and non-self DNA. The *dnd* system, for instance, adds a sulphur group to the host DNA[20].

1.1.3.2 Prokaryotic Argonautes

Argonaute proteins were originally discovered in eukaryotes, where they play a crucial role in RNA interference (RNAi)[21]. Argonaute, as the main functional component,

together with subunits of RNA-induced silencing complex (RISC) and single-strand guide RNA specifically targets RNA sequences, silencing the target RNA expression[22].

Homologous prokaryotic Argonautes were then identified. Considering the lack of RNAi pathways in prokaryotes[23], the functional role of prokaryotic Argonautes was unclear at the beginning. It was hypothesized that Argonautes in prokaryotes provided defence against MGEs, as Argonautes are frequently encoded in defence islands (regions in genome enriched defence genes). Pervasive horizontal gene transfer of Argonautes also suggests its defensive role[24]. Recently, DNA-guided DNA interference by a prokaryotic Argonaute has been reported[25]. One form of prokaryotic Argonautes can non-specifically degrade invasive DNA at first, and the degradation product, as guide DNA, leads to a specific DNA interference to the same invasive DNA[26]. RNA-guided DNA interference has also been demonstrated in a different prokaryotic Argonaute system[27]. More recently, a short prokaryotic Argonaute system was shown to trigger cell death by $NAD(P)^+$ depletion upon high-copy invading DNA detection[28]. It is an abortive infection as discussed below.

1.1.3.3 Abortive infection

There is a drastic defence strategy against phage infection in bacteria. The bacterial cell can activate “programmed cell death” after phage infection instead of trying to survive through defence systems. This is called abortive infection (Abi), which cuts off materials and energy supply before a phage life cycle is completed and protects surrounding cells from phage infection[29]. Abortive infection is often triggered by a specific element such as a phage protein, phage DNA or a cellular phage infection

state[30]. Different types of abortive infection have been shown in **Figure 1.2** blue panel. Below is a more detailed explanation.

Escherichia coli has been widely used in phage infection studies as a model organism[31]. As so, abortive infection was first explored in *E. coli*. Two abortive infection systems (Lit and PrrC system) have been well studied in *E. coli*. The Lit system is activated by phage T4. Lit is a protease of *E. coli* activated by the Gol peptide of the T4 capsid protein, cleaving the ribosomal elongation factor EF-Tu, resulting in translation shutdown, which leads to cell death[32]. PrrC of *E. coli* cleaves the bacterial $tRNA^{Lys}$ when the R-M systems have been suppressed by phage[33]. In addition, more than 20 Abi systems have been described in *Lactococcus lactis*. AbiZ, for example, accelerates the lysis, causing premature lysis of phage-infected cells in *L. lactis*[34].

Toxin–antitoxin systems are defined by a pair of toxin and antitoxin genes, which can be used to execute abortive infection[35]. The ToxN/ToxI TA pair in *Pectobacterium atrosepticum* is a typical TA system. ToxN possesses RNase activity but is sequestered by antitoxin ToxI in the normal cellular state. However, its RNase activity is released after phage infection. Both host and phage transcripts can be degraded by its ToxN[36].

Retrons are genetic elements that were originally discovered in *Myxococcus xanthus*[37]. The general composition of Retron is a non-coding RNA (ncRNA) and a reverse transcriptase (RT), typically generating a chimera RNA-DNA molecule via its RT activity. Retrons have been found in different organisms, but the biological function of Retrons remains unclear until its anti-phage defence role was explored[38, 39]. A retron Ec48 activates abortive infection upon phage infection by guarding RecBCD. Ec48

senses RecBCD inhibition that is caused by phage protein, and its associated effector is activated, leading to abortive infection[39].

A recent study reveals that gasdermin homologs in bacteria provide phage defence and execute cell death, yet another abortive infection with a novel mechanism[40]. In mammals, gasdermin forms cell pores after proper cleavage by caspase protease, which is activated by pathogen infection, causing pyroptotic cell death[41-43]. This strategy probably originated from prokaryotes. Bacterial gasdermin has been identified and presented a similar mechanism of cell death to the mammalian gasdermin. Gasdermin in bacteria is cleaved by caspase-like protease when phage is infected, forming cell pores, and interrupting cell membrane integrity, leading to cell death[40].

As the gasdermin defensive system demonstrates, an immune mechanism in eukaryotes could date back to ancestral prokaryotes. Cell signalling defensive systems has been found in bacteria, showing homology to eukaryotic defences. These systems produce cyclic oligonucleotides as signal molecules upon phage infection, activating the effector to achieve anti-phage defence[44-47]. The same strategy has been adopted by well-known innate immunity cGAS-STING in animals to activate immune response[48]. CBASS (cyclic oligonucleotide-based anti-phage signalling system) is the first elucidated signalling defence system in bacteria. Phage infection induces the production of cyclic oligonucleotides by CBASS cyclase (CD-NTase, cGAS/DncV-like nucleotidyltransferase). The cyclic products vary in different CBASS systems, including cyclic GMP–AMP (cGAMP), cyclic tri-adenylate (cA_3) and others[44, 49]. Those signalling molecules activate CBASS effectors that kill cells in diverse mechanisms. One of them causes cell death by triggering phospholipases, which break cell membrane integrity[45]. NADase has been shown as the effector to execute NAD^+ depletion, leading to cell death[49]. In addition, endonucleases can serve as the effector,

indiscriminately cleaving DNA[50]. Other effectors contain transmembrane domain and membrane ion channel[45]. The death of infected cells inflicted by CBASS system retains the spread of phage, hence providing defence immunity to bacteria. Pycsar (pyrimidine cyclase system for antiphage resistance) is another signalling defence system, characterized by the pyrimidine cyclases that specifically synthesize cCMP and cUMP upon phage infection. There are two effectors in Pyscar signalling, a transmembrane effector and a NADase effector. They induce cell membrane impairment and NAD^+ depletion respectively, resulting in abortive infection[46]. Thoeris system produces an isomer of cyclic ADP-ribose as signalling molecules by ThsB when phage infects. The Toll/interleukin-1 receptor (TIR) domain of ThsB triggers signals production. Following an effector (ThsA, a NADase) activation, infected cells deplete NAD^+, leading to cell death[47]. TIR-domain is identified originally in eukaryotes and has been deeply investigated in eukaryotic immune response[51]. Thoeris provides evidence that prokaryotes and eukaryotes share similarities in signalling defence.

1.1.3.4 Inhibition of DNA and RNA synthesis

Recent studies have shown that bacteria produce small molecules to interrupt phage DNA and RNA synthesis directly. Prokaryotic viperins (pVips) process nucleoside triphosphate (NTP) into 3'-deoxy-3',4'-didehydro (ddh)-cytidine triphosphate (ddhCTP), ddh-guanosine triphosphate (ddhGTP) and ddh-uridine triphosphate (ddhUTP). Those ddh modified nucleotides act as an RNA synthesis terminator, inhibiting phage RNA transcription[52]. Viperin was first discovered in animals as an interferon-induced protein, it produces ddh modified nucleotides to stop viral RNA

transcription[53]. The homology in prokaryotes reveals the origins of the eukaryotic viperins.

Small molecular compounds of bacterial secondary metabolites have been reported to inhibit phage DNA replication and RNA transcription. Anthracyclines, a class of molecules from *Streptomyces,* can be inserted into phage DNA to block phage DNA replication[54]. Aminoglycoside antibiotics produced by *Streptomyces* also interrupt phage life cycle. Phage DNA replication and RNA transcription are inhibited when aminoglycoside antibiotics is present, but the underlying mechanism has not been characterised yet[55].

Another strategy to disrupt phage DNA replication is to prevent phage from acquiring the substrate of DNA synthesis. Two types of deoxynucleotides depletion enzymes induce dNTP shortage to shut down phage DNA replication. This immune response starts when bacterial transcription is suppressed by phage infection. One of the enzymes, dCTP deaminase converts dCTP to dUTP, and another enzyme, dGTPase dephosphorylates dGTP to dG. These two products are no longer used for DNA replication, hence, blocking phage propagation[56].

1.1.4 Diversity of prokaryotic defence systems

Besides the defence systems mentioned above, there are still a variety of defence strategies in prokaryotes. Lysogenic phage inhibits lytic phage by expressing a repressor to maintain lysogeny, which is considered as a form of superinfection exclusion[57]. *Streptomyces* spp. produce small molecules (doxorubicin and daunorubicin) to disturb phage DNA replication without interfering with host DNA replication[54].

A diverse world of prokaryotic defences has been described. The anti-MGEs mechanisms in prokaryotes have been studied for decades. However, the diversity of the field was widely expanded recently. Back to the initial stage of defence system discovery, the two widespread systems, R-M and Abi, were observed co-located frequently on the genome, suggesting genes of defences are clustered in genomic islands[58, 59]. With this assumption, researchers explored bacterial and archaeal genome data bioinformatically and found numerous novel prokaryotic defences in the "defence island" where defence genes enriched[60]. This approach revealed the signalling Abi system, for instance, CBASS that has been experimentally studied (see description above). Further novel defences found in the study have been elucidated, and there are more remaining undissected. But this study is not the end of expanding the prokaryotic defence arsenal. A functional selection study has revealed defences that reside out of the defence island. Intriguingly, those systems are primarily encoded in prophage and MGEs[61]. This raised a question. The so-called "defence system" defends whom from what? We thought the defence system was the immunity of bacteria and archaea against MGEs, the defence line between the host and the invader. But at present, the boundary between these two rivals is blurred. There are MGEs that help the host to counter other MGEs. Three parties in this playground for prokaryotic cells and MGEs: the host, the invasive MGEs and the defensive MGEs The interaction between those three is extensive and can be either antagonism or symbiosis[62]. A representative example is the Phage-inducible chromosomal islands (PICIs). PICIs are mobile genetic elements in bacteria[63]. The expression of PICIs is repressed until bacteria are infected by the helper phage. PICIs then generate proteins and the PICI genome to alter helper phage capsid size and seize the phage capsid,

interfering the phage assembly[64]. In this case, the MGEs PICIs play a defensive role against other phages by seizing phages' assembly components. Moreover, PICIs have been shown to carry defence systems that protect the host from phage competitors without going through PICI expression cycles (no capsid seizing)[65].

1.1.5 Anti-defence systems

The prokaryotic arsenal provides weapons to defence the host from MGEs, but the MGEs can find a way to counterattack. Anti-defence systems are the key to tackling the various defences.

Potentially, where there is a defence system, there is a counterpart anti-defence system. A well-studied example is the anti-CRISPR[66]. The showcase of this anti-defence strategy is that phage encodes anti-CRISPR proteins (Acr), targeting CRISPR associated proteins to repress CRISPR interference. AcrF1, AcrF2 and AcrF3 are three anti type I-F CRISPR proteins. AcrF1 and AcrF2 bind to CRISPR Cascade (CRISPR-associated complex for antiviral defence), and AcrF3 inhibits Cas3[67]. The successful inhibition helps phage infections. A recent study has discovered that in the leading region of plasmids (one of the MGEs), various anti-defence systems were encoded, including anti-CRISPR, anti-SOS, anti-restriction and unknown anti-defences[68]. The discovery of novel anti-defence systems will also lead to the discovery of new defence systems. Anti-defence system plays an important role in shaping the interaction between the host and MGEs[62]. The two sides of this race will keep evolving, and more exciting mechanisms await to be discovered.

1.1.6 Defence and anti-defence

The diversity of defence and anti-defence exists in nature. We may ask why there are myriad defence systems. The simple answer is that there is a variety of MGEs. No bacteria are immune to all MGEs infections, no phages are omnipotent to invade all bacteria. Namely, extensive interaction between the host and MGEs results in diversity. They apply a strategy that makes the community fit in that niche. The situation alters over time, so the defences and anti-defences change. CRISPR-Cas system, for example, adapts spacers to eradicate phage infection, but the spacers can be very dynamic in time despite overall infected microbial being relatively stable[69]. The bacteria phage that evades the host toxin-antitoxin defence loses the ability to invade the other host since they deleted the anti-defence gene that counters the other defence system to maintain the size of the genome for packing[70]. An interesting cooperation between two defence systems was reported recently, CRISPR-Cas13 non-specifically degraded phage and host transcripts upon target recognition, resulting in cell dormancy. RM system cleaves phage genome out and recovers cell from dormancy[71]. We consider the defence and anti-defence as an arms race between bacteria and phages, but if we observe it from an ecological scale, this arms race is not everlasting since it has cost and has to be compromised for each side at a point to get adapted to the current situation[72].

Studying the treasure vault of prokaryotic defences and anti-defence will help us understand the evolution of bacterial interactions with mobile genetic elements and the meaning of the arms race on the ecology level and discover new tools to reshape the development of biology, the medical, and the industrial.

One well-established exploitation of prokaryotic defence systems is repurposing those systems into gene-editing tools. Prokaryotic Argonautes with DNA or RNA guide DNA interference ability are considered as potential genome-editing tools[60]. More strikingly, CRISPR-Cas systems have been successfully applied in gene editing. The CRISPR-Cas systems have become a widely applied gene-editing tool from the initial prokaryotic adaptive immune system over a decade. The following part will introduce CRISPR-Cas systems in detail.

1.2 CRISPR-Cas systems

1.2.1 Discovery of CRISPR

CRISPR has become one of the most exciting advances in the molecular biology field over the last decade. The discovery of CRISPR, however, was initially in the year of 1993. It is in the Mediterranean port of Santa Pola, a beautiful coast in Spain, that Francisco Mojica was working on his doctoral studies, in which he investigated the correlation between salt concentration and genome digestion by restriction enzymes in *Haloferax mediterranei*, an archaeon from Santa Pola's marshes with extreme salt tolerance. During his studies of the genome, he found a palindromic repeated sequence that was separated by spacers on a DNA fragment [73]. A Japanese group reported a similar structure in the late 1980s[74], but no further investigation was carried out on it. Other than that, there were no similar structures in known microbes at that time. Mojica, however, decided to unlock the mystery of the structure. He kept on with the investigation and reported this new class of repeats[75], trying to use bioinformatics for further investigation. After nearly ten years of devotion to this study, he had identified different loci with this structure in 20 microbes[76]. Researchers gradually took notice of this prevalent structure. More features of the repeat locus, such as the repeats were often flanked by associated genes, were characterized. It came to the time to give this mysterious structure a name. Mojica originally called it short regularly spaced repeats (SRSRs). The name was then changed to clustered regularly interspaced palindromic repeats (CRISPR)[77, 78], which has become one of the most known systems in modern biology research.

The functional role of CRISPR was unknown. Different hypotheses were proposed, including DNA repair, gene regulation and other functions[78]. But all of them lacked

experimental evidence. Mojica was still trying to find hints from bioinformatics. As microbial sequence databases expanded, he strikingly found that a spacer of CRISPR locus in an *E. coli* strain had the same sequence of a P1 phage, which infects many other strains of *E. coli* but not the one with the same sequence spacer. Mojica suggested the immune function of CRISPR[79]. A group in France also discovered that CRISPR spacers in *Yersinia pestis* derived from bacteriophage DNA, implying the CRISPR defence function[80]. Another French group reported that CRISPR had spacers of extrachromosomal origin, and the authors speculated that the transcripts from the CRISPR could inhibit phage gene expression in an anti-RNA manner[81]. The crucial experimental evidence appeared in 2007. Philippe Horvath's group proved the immunity feature of CRISPR systems by showing different resistance to phages when integrating new spacers into bacteria or removing particular spacers[82].

The function of CRISPR systems was finally characterized through efforts over two decades. It provides immunity and defence against foreign genetic elements. But how does CRISPR provide immunity? What is the machinery of CRISPR systems? These important questions have occupied scientists for the past 15 years.

1.2.2 Overview of CRISPR

Before we discuss the mechanism of CRISPR systems, a basic understanding of CRISPR structure is necessary. The CRISPR array on prokaryotic genome possesses several unique features. Partially palindromic repeats are interspaced by similarly sized different sequences (spacers). The CRISPR loci are flanked by a leader sequence on one side. Besides, CRISPR-associated (*cas*) genes are invariably adjacent to a CRISPR locus[77] (Figure 1.3).

Back to the discovery story, it had been widely accepted that CRISPR systems had a defence role in prokaryotes. Researchers were seeking more details about this curious system. In 2008, Stan Brouns and his colleagues showed that the immunity in *E. coli* required a 61-nucleotide-long CRISPR RNA (crRNA). The maturation of crRNA relied on the cleavage of a long precursor RNA, the transcript of the CRISPR locus. They also found that Cas proteins formed an effector complex (Cascade). The cascade complex together with mature crRNA is responsible for CRISPR defence. Furthermore, they designed and introduced the first artificial CRISPR arrays, targeting lambda (λ) phage genes, to an *E. coli* strain without lambda phage resistance. The strain gained resistance to lambda phage after introducing sense or anti-sense CRISPR arrays, and the efficiency was higher when sense arrays were introduced. With this observation, they proposed a hypothesis that CRISPR may target DNA instead of mRNA to confer immunity[83].

The indisputable proof came out in the same year. Luciano Marraffini and Erik Sontheimer noticed that a spacer of CRISPR in *Staphylococcus epidermidis* matched a *nickase* (*nes*) gene on plasmids from *Staphylococcus aureus*. The plasmids cannot be transferred to *S. epidermidis.* They tried to reconstruct the CRISPR system of *S. epidermidis in vitro* to elucidate the target of CRISPR effector complex. Unfortunately, this system was too complicated to rebuild at the time. But they figured out another intelligent molecular biology method to prove DNA targeting. The plasmids were introduced to a self-splicing intron. If CRISPR targeted mRNA, the interference would still work because the intron would be spliced. But if the interference was lost, it would prove that the CRISPR effector targeted DNA since the DNA sequence cannot match the spacer sequence. They could not see interference after intron insertion, which means CRISPR targets DNA[84].

We now understand that normally CRISPR immunity could be considered as a three-stage process: adaptation, expression, and interference (Figure 1.3). There are a variety of CRISPR systems in nature. CRISPR systems are widely spread in Bacteria and Archaea. We now have a general impression of CRISPR, but the details of CRISPR systems are quite different. Next, the systemic introduction of CRISPR-Cas will be presented.

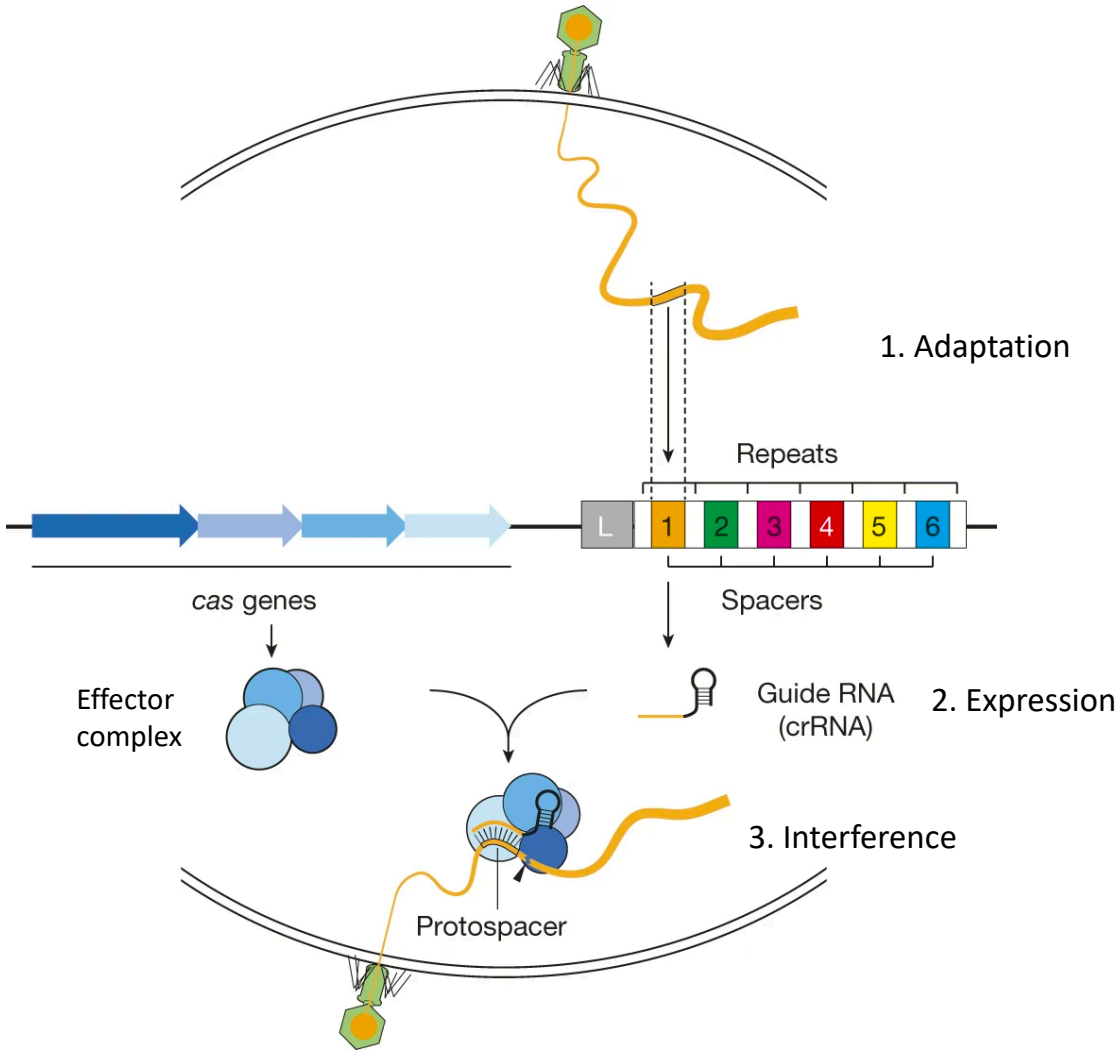


Figure 1.3 Stages of CRISPR-Cas systems

Repeats (white), Spacers (coloured) and leader sequence (L) form the CRISPR locus, which is flanked by CRISPR-associated genes (*cas* genes). Adaptation stage: spacers were acquired from invasive genetic elements, phage DNA for example. Expression stage: Transcripts from CRISPR locus were cleaved to crRNA. Effector complex consisted of *cas* proteins. Interference stage: Effector complex together with crRNA targeted phage DNA protospacer (spacer captured by CRISPR in phage genome). Adapted from Luciano Marraffini[85].

1.2.3 Classification of CRISPR-Cas systems

Diverse CRISPR-Cas systems require a valid classification for systematic research. Classification provides hints for experimental study, and experimental results evolve classification. The *cas* proteins act a key role in CRISPR-Cas classification[86]. They are close to CRISPR arrays on the genome. When the CRISPR function was still a mystery in 2002, they were proposed to be co-functional with CRISPR repeats, and *cas* protein family may be homologous to DNA-helicases and exonucleases[87]. Meanwhile, protein families, named Repair Associated Mysterious Proteins (RAMPs), were identified[88]. The RAMPs were later renamed as Repeat Associated Mysterious Proteins by Haft *et al.* in 2005 without changing the acronym[89] , because they were associated with CRISPR systems. In this report by Haft *et al.*, more *cas* protein families were defined and the original classification of CRISPR-Cas systems was made.

Although the original classification proposed by Haft possessed simplicity, it cannot thoroughly illustrate the relationship between c*as* proteins. Makarova *et al.* proposed a new classification in 2011[90], in which the CRISPR systems were mainly divided into three types: type I CRISPR-Cas systems contain *cas3* gene and components of Cascade-like genes; type II CRISPR-Cas systems contain a single large protein--Cas9; type III CRISPR-Cas systems contain polymerase and RAMP modules. Most CRISPR systems can be classified into three types, according to the signature gene, and even subtypes. All the CRISPR systems including Cas1 and Cas2 shares similarity during adaptation stage. But the expression and interference stages are quite different between various types of CRISPR systems. The classification is widely accepted and used nowadays while it keeps updating.

In 2015, Makarova *et al.* developed the classification of CRISPR systems. The CRISPR-Cas systems were defined into two classes, five types and 16 subtypes. Class 1 includes type I, type III and type IV CRISPR-Cas systems, and they both possess a multi-subunit crRNA-effector complex. Class 2 includes type II and type V CRISPR-Cas systems, and a single subunit crRNA-effector is their feature[86]. Recently, an expansion of the classification systems was released. Class 2 now contains 3 types: type II, type V and type VI. Variants of CRISPR-Cas, lacking the nucleases for interference, were also identified. In addition, numerous ancillary CRISPR-linked genes were discovered (Figure 1.4)[91].

Classification helps us understand the differences between CRISPR-Cas systems. Next, CRISPR machinery will be introduced by comparing different types of CRISPR-Cas systems.

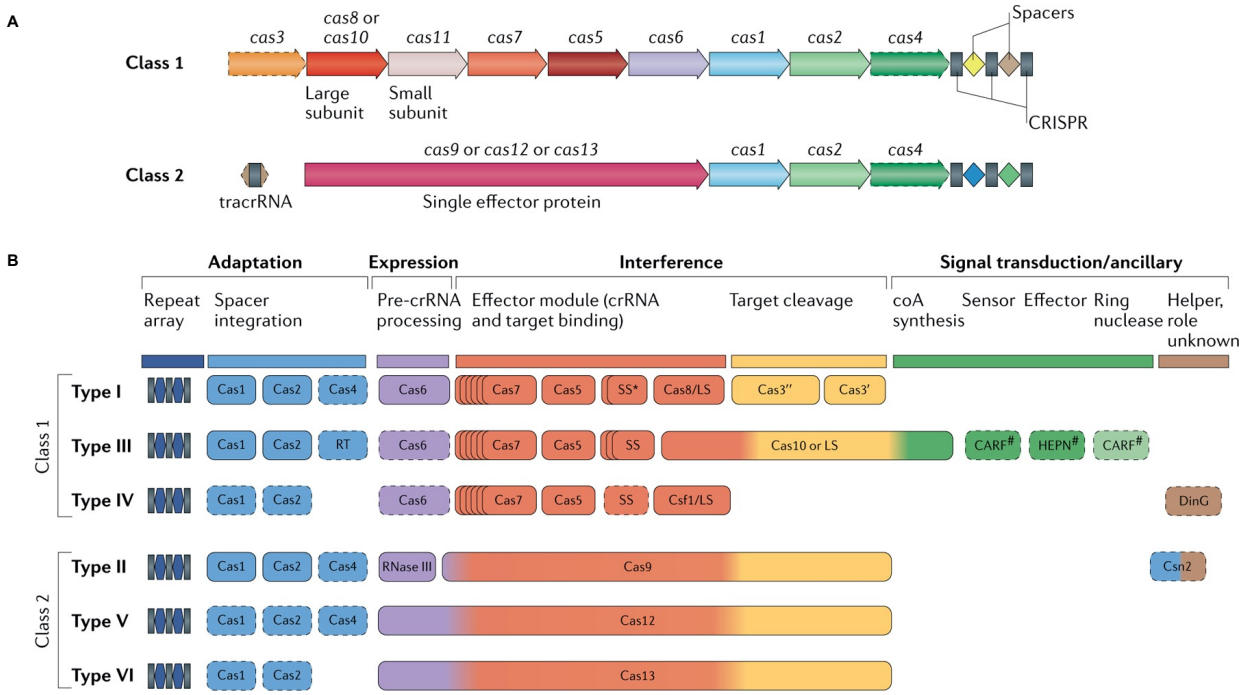


Figure 1.4 Classification of CRISPR systems

(A) A schematic organisation of class 1 and class 2 CRISPR-Cas loci; Multiple cas proteins comprised the effector complex in class 1 CRISPR, the dashed outlined cas3 and cas4 are replaceable or missing in some subtypes. Class 2 CRISPR contains a single effector protein. **(B)** Functional modules of classified CRISPR systems; Dispensable genes are indicated with dashed outlines. * Small subunit might fuse to large subunit; CARF, CRISPR-associated rossmann fold and HEPN, higher eukaryotes and prokaryotes nucleotide-binding, domain proteins are common sensors in type III CRISPR; # Unknow sensor or ring nuclease may involve sensing. Adapted from Kira S. Makarova[92].

1.2.4 Stage one: adaptation

Despite CRISPR being divergent in the interference effector, the adaptation complex is relatively conserved across CRISPR systems, featured by the widely conserved protein, Cas1 and Cas2 (Figure 1.4). During the adaptation stage, mobile genetic elements (MGEs) are captured and integrated into the host genome as spacers. The capture process first involves a DNA repair machinery of the host—RecBCD in Gram-negative organisms (type I-E CRISPR in E. coli for example)[93] and its homologue, AddAB, in Gram-positive organisms (type II CRISPR for instance)[94]. They produce substrates for spacer acquisition (Figure 1.5A). A complex formed by Cas1 and Cas2 facilitates spacer selection. In type I and type II CRISPR systems, a protospacer-adjacent motif (PAM) is crucial for selection and preventing self-targeting. Both type I and type II CRISPR rely on proper PAM as a marker to target and interfere[95]. Therefore, in type I-E system, the Cas1-Cas2 complex recognizes PAM directly and selects the protospacer specifically[96]. Although type II system requires PAM, its Cas1-Cas2 complex lacks the ability to direct recognition of PAM. The PAM-interacting domain of Cas9 interacts with the Cas1-Cas2 complex to accomplish bias acquisition[97] (Figure 1.5B). More proteins, such as Cas4 in type I and accessory proteins, Csn2, in type II, are reported to be involved in the bias spacer selection[97-99]. Although the mechanism of spacer acquisition in type III is still unclear, a study has shown that a reverse transcriptase (RT)-Cas1 fusion protein can acquire spacers from RNA transcripts in the type III CRISPR system[100].

Primed spacer acquisition was observed in type I CRISPR adaptation[101, 102]. This kind of acquisition is compared with “naive acquisition” where initially a spacer from a new genome was incorporated into CRISPR array, but when a spacer from a specific

genome pre-existed in the CRISPR array, the rate of acquisition of additional spacers from this specific genome was increased. This is called primed spacer acquisition or priming[101]. The primed spacer acquisition can be either complete (Figure 1.5C) or partial (Figure 1.5D) matching, and the partial matching enables CRISPR system to adapt spacers to overcome the mutation escapes of MGEs[103]. Primed spacer acquisition is associated with the interference of CRISPR, in type I CRISPR-Cas system, the interference effector complex (Cascade) recognises specific DNA and recruits Cas3, a helicase and nuclease, to destroy target DNA. The product form degradation can be used for spacer acquisition by Cas1-Cas2 complex (Figure 1.5C). In contrast, when PAM in target DNA was mutated and target DNA escaped from surveillance, Cas3 was no longer recruited to Cascade directly. It requires the presence of Cas1-Cas2 to enhance Cas3 recruitment to the mutant site. Cas3 nuclease activity was attenuated by Cas1-Cas2, it translocated in either direction from Cascade and drove spacer acquisition with Cas1-Cas2 integrase (Figure 1.5D)[103]. Type I-F CRISPR in *Thermobifida fusca* shares similarity in spacer primed spacer acquisition. The process requires Cascade, Cas3 and Cas1-Cas2 (primed acquisition complex, PAC), PAC can travel long on the target genome, suggesting long-range spacer acquisition[104].

Recently, type II primed spacer acquisition has been reported[105]. The DNA product from Cas9 degradation provides substrates for the adaptation complex to acquire spacers, enhancing the rate of spacer acquisition[105].

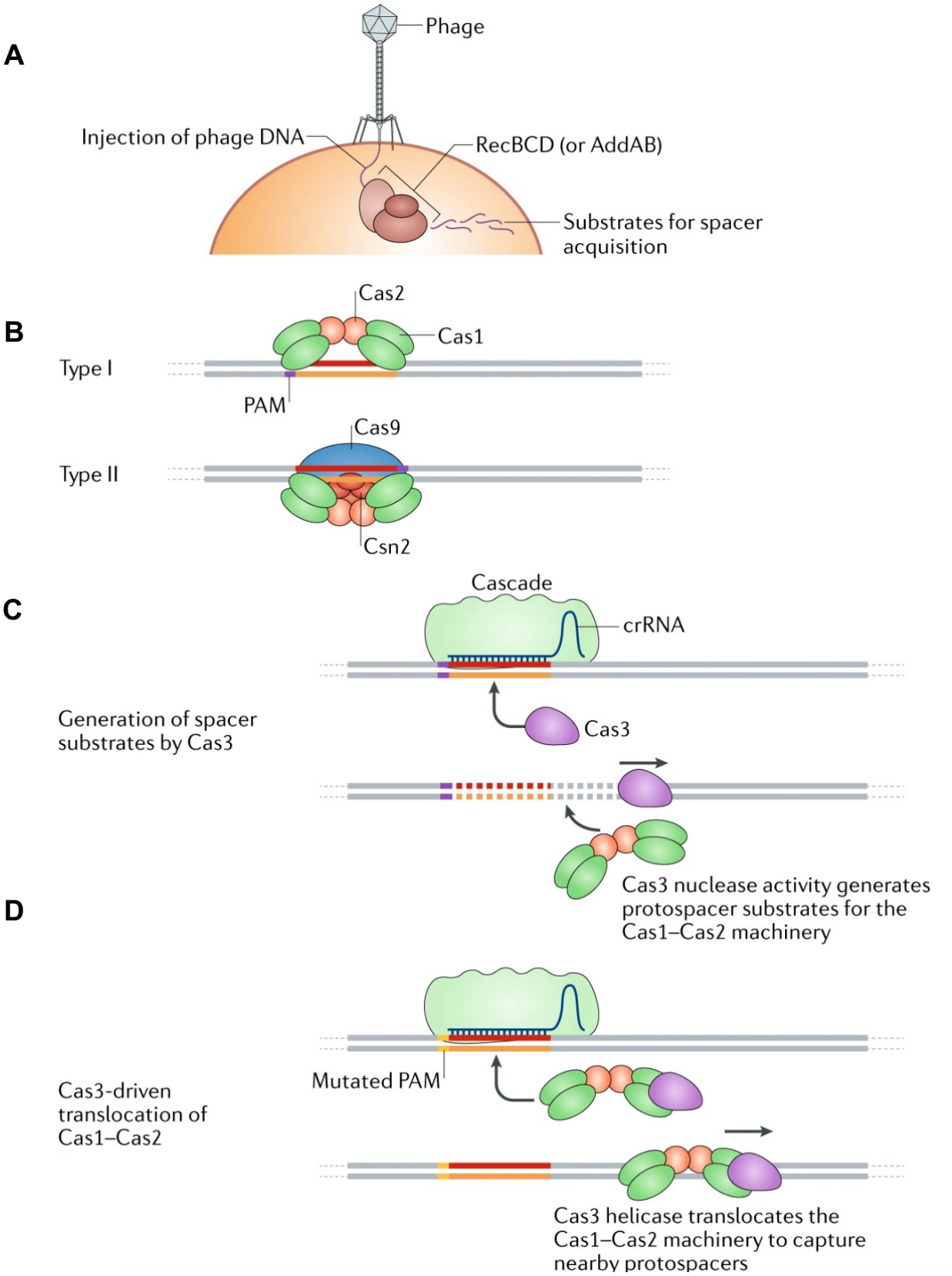


Figure legends next page.

Figure 1.5 Spacer selection and capture.

(A) After viral DNA injection, RecBCD in Gram-negative organisms (or AddAB in Gram-positive organisms) generates substrates for spacer acquisition. **(B)** Protospacer selection, In type I-E CRISPR system, the Cas1-Cas2 complex prefers protospacers with PAM directly. In type II, Cas9 and Csn2 are required for protospacer selection. **(C)** Primed spacer generation with complete matching, Cascade effector binds to target DNA and recruits Cas3 to degrade DNA, generating protospacer substrates for the Cas1-Cas2 adaptation. **(D)** Primed spacer generation with PAM mutation, instead of recruiting Cas3 immediately upon binding, Cas1-Cas2 was present to enhance Cas3 recruitment, and the complex was translocated from the target site to capture nearby protospacers. Adapted from Jon McGinn[106].

Once the heterohexameric complex [$(Cas1_2–Cas2)_2$] is loaded with protospacer, it then acts as the spacer integrase. The integration of new spacers into CRISPR array is polarized. New spacers were incorporated between the leader sequence (an AT-rich sequence preceding CRISPR array) and the first repeat, hence the new spacer become the new first spacer[107, 108]. The protospacer with Cas1-Cas2 complex firstly attacks the site and ligate itself between the leader and the repeat. Subsequently, the second attack and ligation take place at the site between the old spacer and the repeat, generating ssDNA for the repeat sequence which is repaired and ligated eventually (Figure 1.6A)[107]. The polarity of spacer integration is conducted by the Cas1-Cas2 complex that contacts with the leader sequence. However, this is not sufficient for type I CRISPR to achieve the polarized preference. For type I CRISPR, host factors, such as IHF or H-NS, mediates bending of the leader sequence and interacts with Cas1. In contrast to type I CRISPR, the Cas1-Cas2 integrase of type II CRISPR is sufficient to achieve without additional host factors. It interacts with the minor groove of the leader

DNA, named the leader anchoring sequence (LAS) for type II CRISPR, to carry out spacer integration (Figure 1.6B)[108].

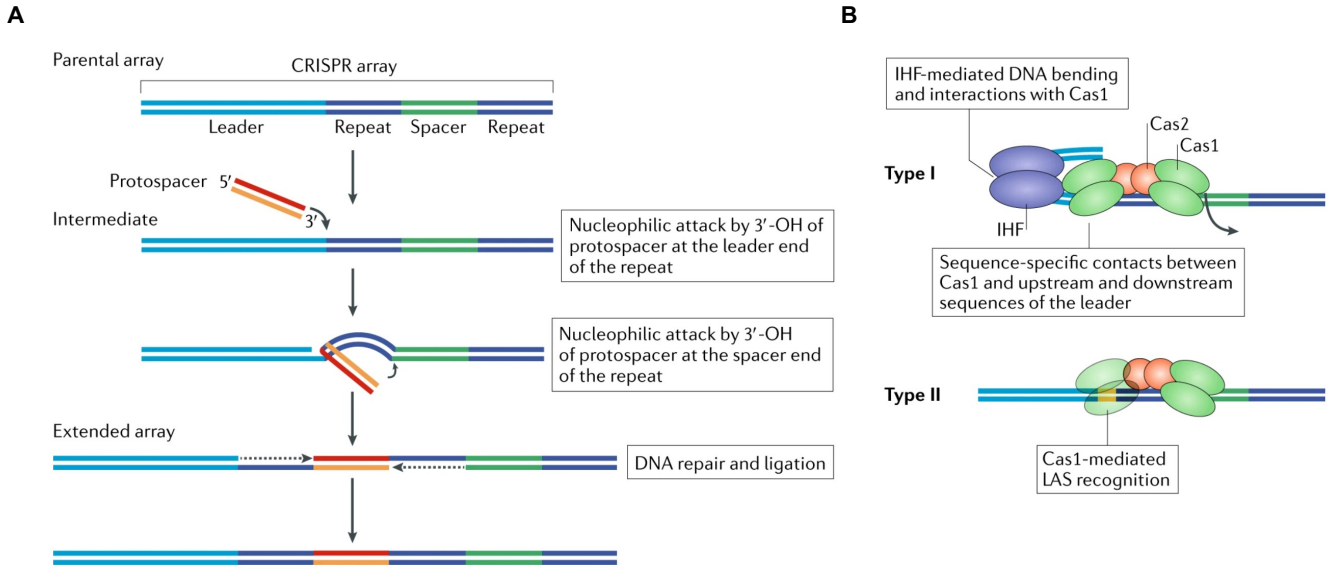


Figure 1.6 Spacer integration

(A) New spacers integrated into CRISPR array; Two cleavage-ligation take place, the first one is at the site between the leader and the repeat, the second one is at the site between the old spacer and the repeat; ssDNA was produced and repaired to achieve spacer integration. **(B)** Mechanisms of the polarized spacer integration; In type I CRISPR, an additional host factor, IHF for instance, is required to interact with the leader and Cas1-Cas2 complex to achieve spacer integration; In type II CRISPR, the Cas1-Cas2 complex is sufficient to integrate new spacer by contacting the LAS (leader anchoring sequence). Adapted from Jon McGinn[106].

We focused on the adaptation of type I and type II systems here. Since Cas1-Cas2 is widely conserved across CRISPR systems, adaptation in other CRISPR systems resembles those two aforementioned systems. But it is noted that different systems inherit specific features under the outline of adaptation.

1.2.5 Stage two: Expression

The expression stage of CRISPR includes expression of pre-crRNA and cas proteins, crRNA maturation and effector ribonucleoprotein (RNP) complex formation. As indicated in the classification of CRISPR, class 1 CRISPR-Cas systems (type I, type III and type IV) possess a multi-subunits effector complex, while class II CRISPR-Cas systems (type II, type V and type VI) utilise a single-unit effector (Figure 1). We will discuss the expression and interference stage by illustrating examples from each type of CRISPR system.

Type I-E CRISPR from *E. coli* is the first biochemically elucidated CRISPR-Cas system[83]. The CRISPR gene cluster of type I-E contains cas gene series, *cas3-cas8-cas11-cas7-cas5-cas6-cas1-cas2*, and the repeat-spacer array (Figure 1.7A). 29 nt repeat forms hairpin and 32 nt spacer sits in between two repeat hairpins post pre-crRNA transcription. Subsequently, the pre-crRNA is processed into mature crRNA by Cas6, where a 5'-8nt handle and a 3'-hairpin are generated (Figure 1.7B). Cas6 remains attached with mature crRNA, and Cas7 forms the backbone along crRNA, together with Cas5, Cas8 and Cas11, a 405 kDa type I-E Cascade (1 Cas5, 1 Cas6, 6 Cas7, 1 Cas8 and 2 Cas11) comprised[83]. Cascade structure elucidated by cryo-EM shows a "seahorse-like" shape RNP complex (Figure 1.7C)[109].

Type I CRISPR has exhibited an elaborate complex, but the complexity of type III CRISPR systems is going even further. Two subtypes (type III-A and type III-B) of type III CRISPR systems were discovered using two sets of *cas* proteins (Csm2-Csm5 for type III-A, Cmr3-Cmr6 for type III-B) respectively for the RNP complex composition (Figure 1.8A)[110-112]. The pre-crRNA of type III systems is cleaved by Cas6, generating an 8nt 5'-tag and a 3'-hairpin[113], which resembles type I CRISPR, but this is not the

end of crRNA maturation in type III CRISPR. The cleaved crRNA is further loaded into Cas10-Cmr/Csm complex[114]. Cas6 is displaced by the complex probably due to physical contact[111]. The Cas10 complex removes the stem-loop (3'-haipin) of crRNA by the interaction between crRNA and Csm3/Cmr4[115]. This interaction defines which part of crRNA is exposed to PNPase (polynucleotide phosphorylase, a host nuclease recruited by Csm5) for crRNA maturation (Figure 1.8B)[116, 117].

Type IV CRISPR expression has similar characteristics to type I and type III CRISPR, *cas* proteins bind to crRNA, forming a crRNP complex. However, the composition of type IV complex is much reductive (Cas5-Cas7-Csf1), where Csf1 is a replacement of the larger subunit in type I (Cas8) or type III (Cas10)[118, 119].

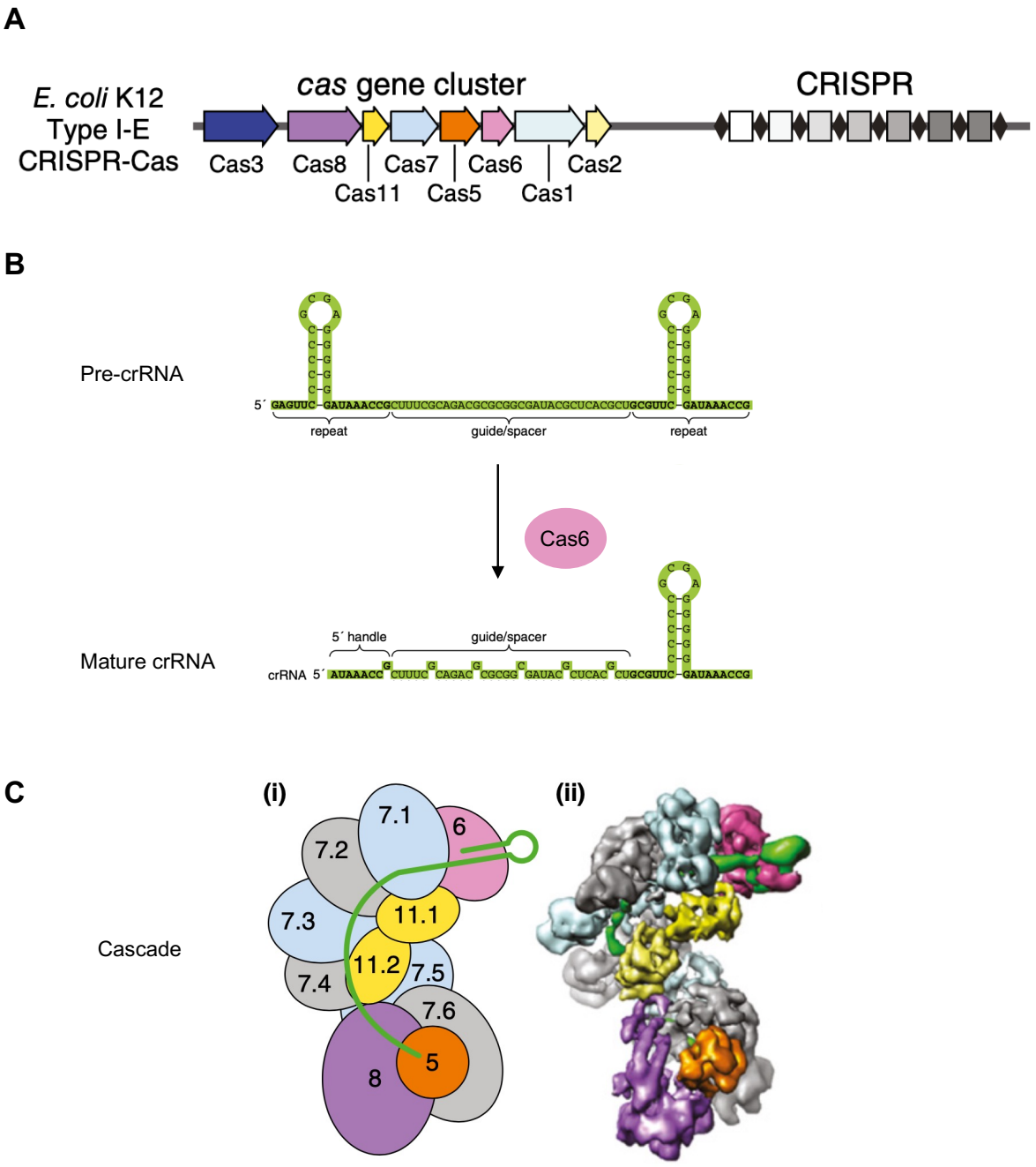


Figure 1.7 type I CRISPR expression

(A) type I-E CRISPR array from *E. coli* K12 strain genome. **(B)** crRNA maturation; pre-crRNA was processed to mature crRNA by Cas6; mature crRNA contains an 8 nt 5'-handle and a 3'-hairpin. **(C)** Structure of type I-E Cascade; (i) A model representation; (ii) A cryo-EM structure. Adapted from John van der Oost[120].

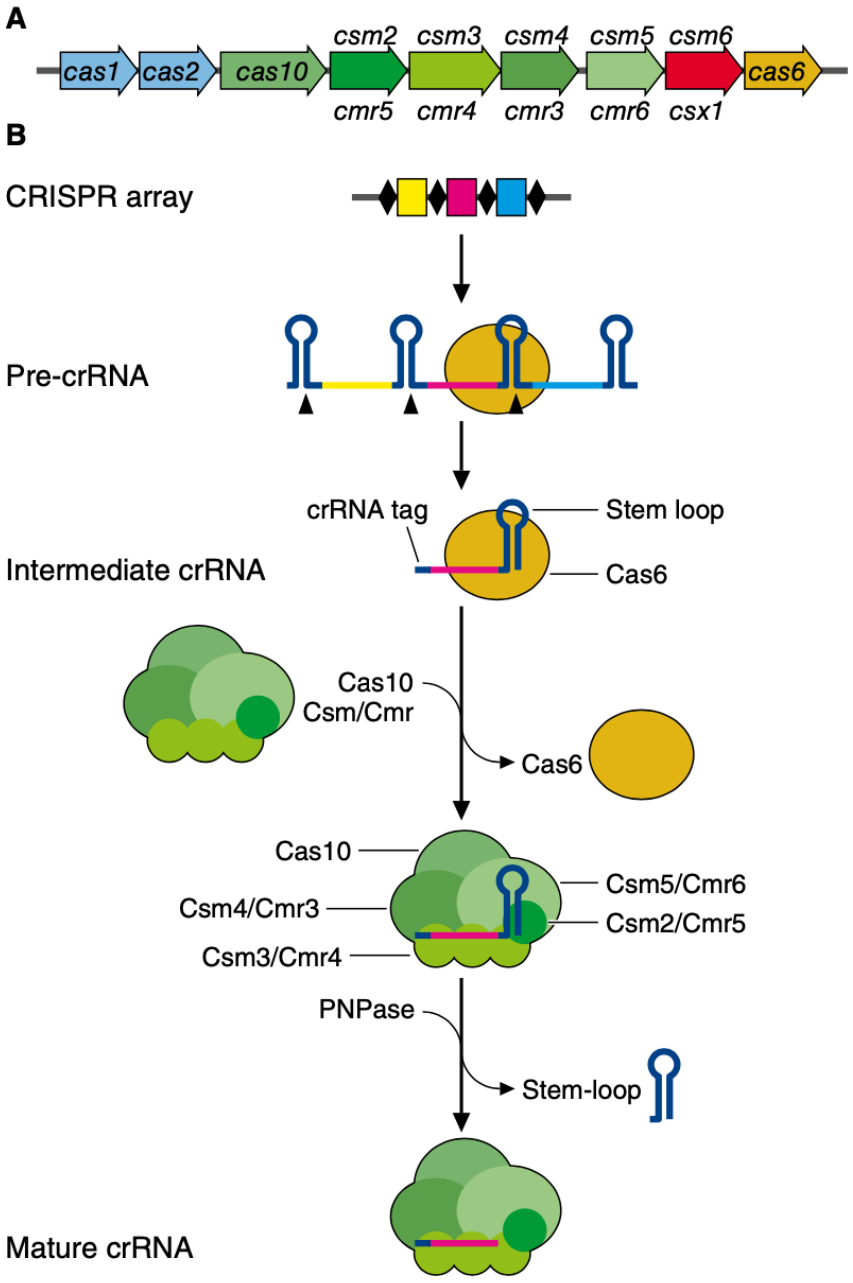


Figure 1.8 type III CRISPR expression

(A) type III CRISPR cluster, *csm2-csm5* for type III-A CRISPR, *cmr3-cmr6* for type III-B CRISPR. **(B)** crRNA maturation; pre-crRNA was first processed to mature crRNA by Cas6, generating an 8 nt 5'-tag and a 3'-stem loop; crRNA was further transferred to Cas10-Csm/Cmr complex; The complex removed the stem-loop of crRNA by recruiting a host nuclease polynucleotide phosphorylase (PNPase). Adapted from Luciano A. Marraffini[121].

Multi-subunit crRNP complexes of Class 1 systems show the complexity of CRISPR. On the other hand, Class 2 CRISPR has revealed the simplicity of CRISPR. In class 2 systems, the multi-subunit complex is substituted by a single protein. A paradigm is Cas9 protein from type II CRISPR. Type II CRISPR cluster consists of the signature gene *cas9* and a *tract* gene across all subtypes of type II systems (Figure 1.9A)[91]. The crRNA maturation in type II applies a unique mechanism, where pre-crRNA is not cleaved by *cas* protein, but an RNA, named transactivating crRNA (tracrRNA), plays an indispensable role in the process[122-124]. Pre-crRNA, however, is cleaved by an endogenous RNase III, and the cleavage requires a formation of RNA duplex including the pre-crRNA and the tracrRNA, which contains a sequence (anti-repeat) that is complementary to the repeat sequence in the pre-crRNA. The cleavage by RNase III produces short crRNA with a 5' overhang that is subsequently processed by an unidentified nuclease to finish crRNA maturation (Figure 1.9B)[122].

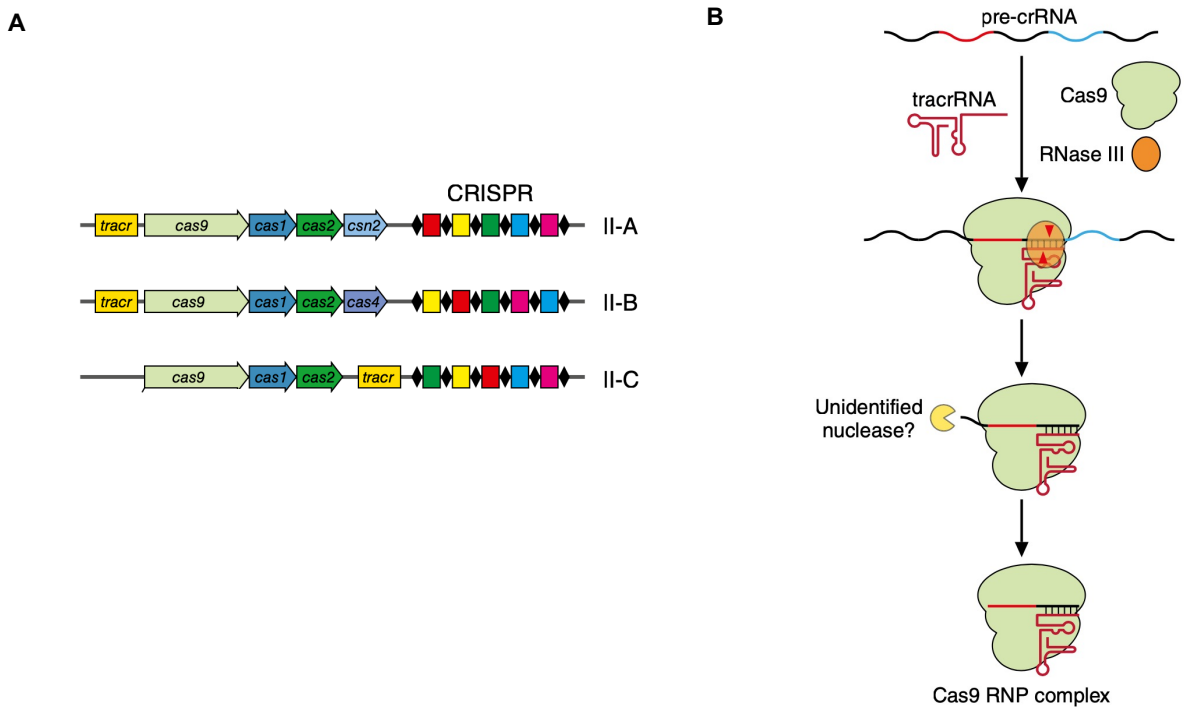


Figure 1.9 type II crRNA maturation

(A) type II CRISPR cluster, *cas9* and *tracr* gene exist in all subtypes of type II CRISPR. **(B)** crRNA maturation; pre-crRNA formed a duplex with tracrRNA, RNase III cleaved the duplex into a short piece; An unidentified nuclease was recruited to further process 5' end of crRNA. Adapted from Tautvydas Karvelis and Virginijus Siksnys[125].

Type V CRISPR features the signature protein Cas12. Cas12 protein possesses a striking diversity that includes 12 subtypes currently (type V-A to type V-K and type V-U, U stands for uncharacterised)[91]. Type V systems utilise three different approaches to achieve crRNA maturation (Figure 1.10). Pre-crRNA can be cleaved directly by Cas12 to generate mature crRNA (Figure 1.10A)[126-128] or perform the type II-like cleavage that requires tracrRNA and additional host factor (Figure 1.10B)[127].

Furthermore, the scoutRNA (short-complementarity untranslated RNA, a unique RNA found in type V system) is required to self-cleave pre-crRNA without additional host factors in some subtypes (Figure 1.10C)[129].

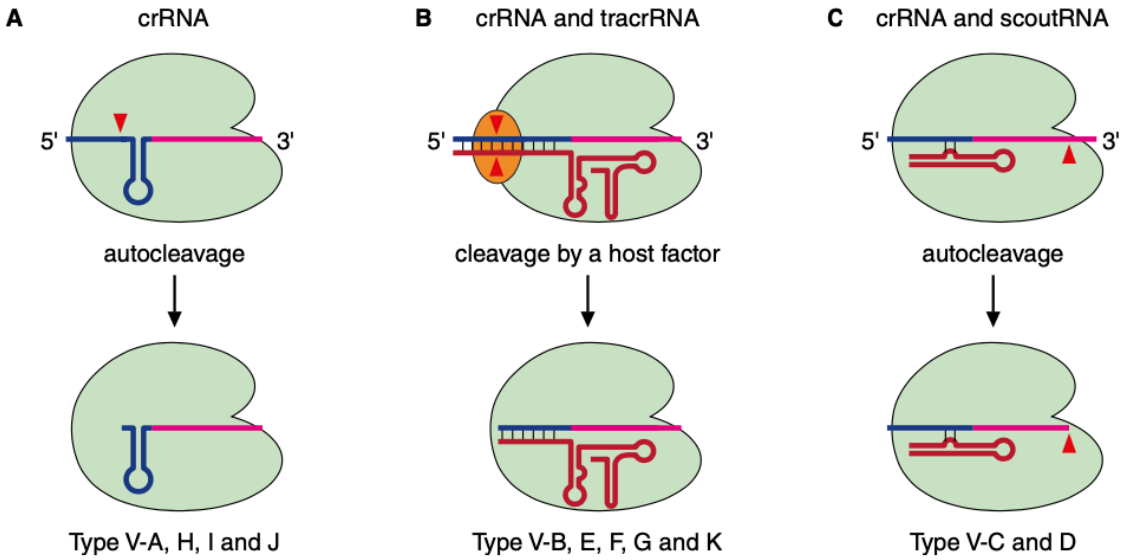


Figure 1.10 Three different ways of type V crRNA maturation

(A) Cas12 (green) directly cleaves pre-crRNA into mature crRNA. **(B)** tracrRNA and a host factor enable crRNA maturation. **(C)** scoutRNA without additional host factor activates autocleavage. Adapted from Morgan Quinn Beckett et al[130].

The crRNA maturation of type VI CRISPR is similar to the first approach of type V systems, pre-crRNA is cleaved by the signature cas protein Cas13 that contains an RNase domain (Figure 1.11)[131].

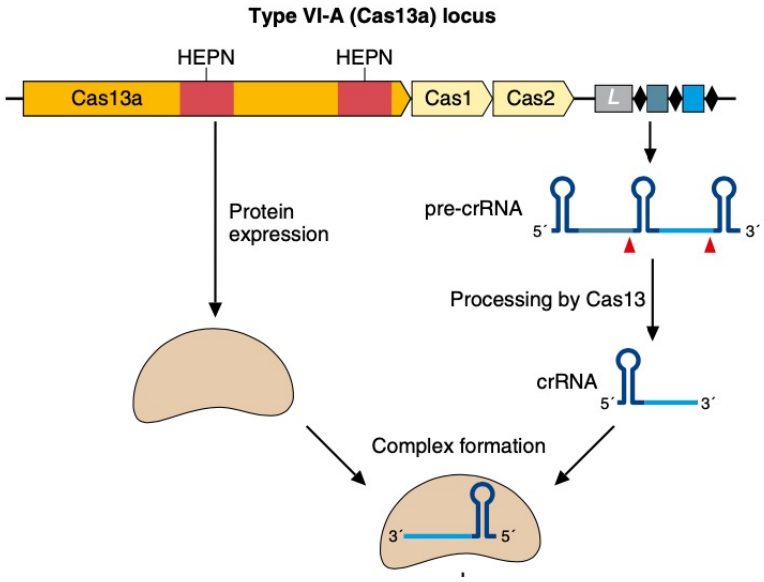


Figure 1.11 type VI expression

pre-crRNA is processed by Cas13 to mature crRNA and the complex formed. Adapted from Omar O. Abudayyeh and Jonathan S. Gootenberg[132].

1.2.6 Stage three: Interference

After the expression of the *cas* protein and crRNA complex (RNP complex), the CRISPR machinery is prepared to perform interference by targeting DNA or RNA derived from MGEs.

Cascade, RNP complex of type I systems, specifically target dsDNA (Figure 1.12). The interference starts from locating Cascade on the target dsDNA under the guide of crRNA. PAM on dsDNA is crucial for targeting non-self-DNA, instead of self-DNA, as mentioned in the adaptation stage[95]. Scanned PAM changes the state of Cascade, enabling target dsDNA unwinding[133]. A seed sequence in crRNA is subsequently base paired with the target DNA. This seed sequence is part of the crRNA, 6 nt to 8 nt at the 5' end of the spacer sequence, and of importance to initialise R-loop formation (a

state that non-target strand of target dsDNA is displaced, spacer in crRNA base paired with the target strand)[133-135]. Single mismatch in seeding abrogates further interference, but sporadic mismatch at downstream of seed sequence is tolerated and R-loop can be formed if not too many mismatches occur[101, 136]. Cas3, a key protein in target dsDNA degradation, is recruited to the Cascade when R-loop is present. Cas3 is comprised of an ATP-dependent helix domain and an HD-nuclease domain that unwinds target dsDNA, reeling and looping the target strand, and cleaving the non-target strand of the target dsDNA (Figure 1.12). The target strand not cleaved by Cas3 may be further degraded by host nuclease[103, 137, 138]. Hence, the target dsDNA is specifically degraded by the Cascade-Cas3.

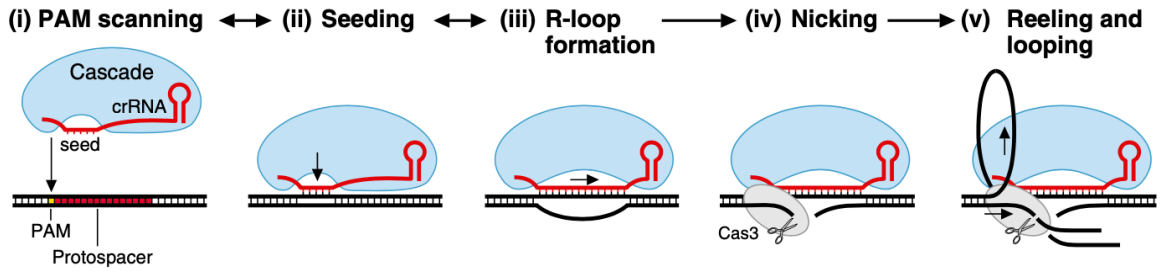


Figure 1.12 type I interference

(i) Cascade scanned PAM on the target dsDNA. (ii) Seed sequence base paired with target dsDNA upon PAM recognition. (iii) crRNA fully paired with target strand to form R-loop. (iv) Cas3 was recruited to Cascade to cleave non-target strand. (v) Cas3 unwound target DNA, reeling and looping target strand with the degradation of non-target strand. Adapted from John van der Oost[120].

Different from type I systems, the target of type III effector complex is RNA instead of DNA, namely, the transcript[110]. Upon recognition of the specific transcript that is

sequence-complementary to the guide RNA, Cas10-Csm/Cmr complex triggers a series of non-specific degradation to achieve type III interference. Firstly, the HD domain of Cas10 is activated to non-specifically cleave ssDNA[139-141]. In addition, Cas10 possesses a polymerase/cyclase palm domain that is activated to catalyse ATP into cyclic oligoadenylates (cOA). cOA acts as signalling molecules that convert inactive type III accessory protein (Csm6/Csx1) to active state. Activated Csm6/Csx1 degrades transcripts of host and invader indiscriminately with its RNase activity[142, 143]. Eventually, the transcript that binds to the Cas10-Csm/Cmr complex is cleaved by Csm3/Cmr4, which ends the Cas10 complex activation (Figure 1.13A)[110, 144, 145].

The degradation of target transcript marks the end of Cas10 activation, non-specific ssDNA cleavage and production of cOA stops to avoid further harm to the host, but the produced cOA keeps activating Csm6/Csx1, which generates continuously non-specific RNA degradation that is not beneficial for the host. To protect host from overreaction, a ring-nuclease is expressed to eradicate cOA[146]. Thus, the non-specific degradation is controlled.

Since type III complex targets transcripts, self-targeting is not likely to take place unless bidirectional transcription generates a complementary transcript. But even in the presence of an antisense transcript, the Cas10 complex can prevent itself from activation. This unique recognition requires a 3' anti-tag on target transcript. The anti-tag is not base-paired with the crRNA 5' tag (Figure 1.13B). Mismatches between the tag and anti-tag lead to Cas10 activation. While the transcript is completely complementary to crRNA 5' tag (Figure 1.13C), the Cas10 is locked into a static state that is incapable of activation[147].

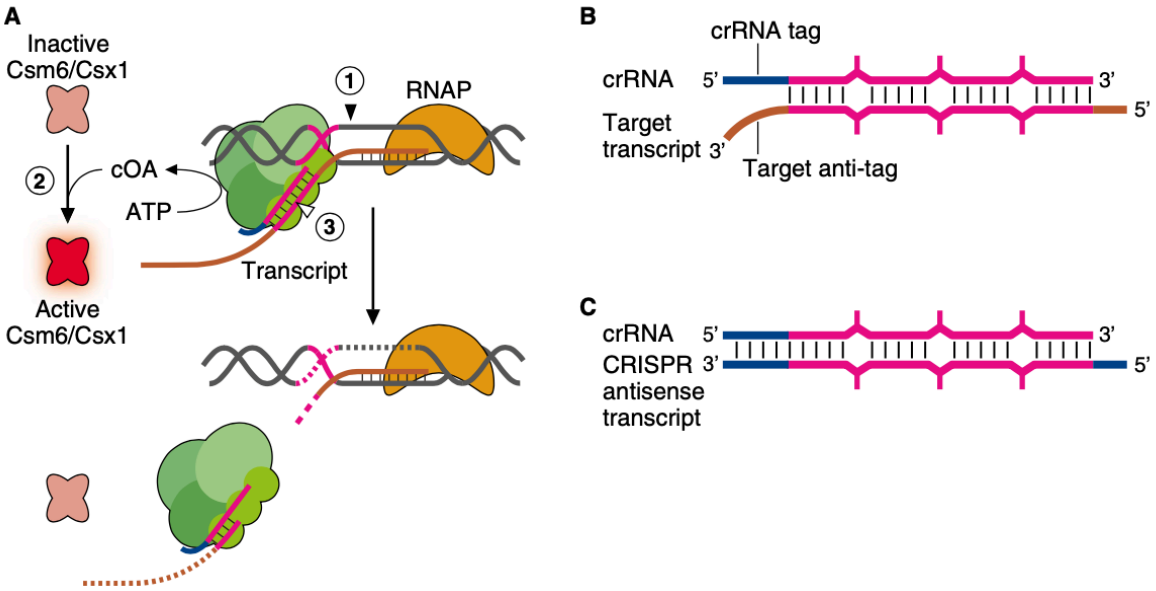


Figure 1.13 type III interference

(A) Cas10-Csm/Cmr complex (green) recognised target transcript, activating Cas10 to non-specifically cleave ssDNA (1); In addition, Cas10 generated cyclic oligoadenylates (cOA) to active accessory protein Csm6/Csx1 that non-specifically degraded RNA (2). Target transcript was cleaved by the complex to deactivate Cas10 (3). **(B)** Spacer of crRNA in the complex base-paired with target transcript with a flipped 6th base; Target transcript possessed a 3' anti-tag that is not complementary to the crRNA 5' tag. **(C)** CRISPR antisense transcript that harbours a complementary 3' sequence was incapable to activate Cas10. Adapted from Luciano A. Marraffini[121].

Since type IV systems lack a nuclease domain in CRISPR cluster, compared with type I (Cas3) and type III (Cas10), the interference of type IV was speculated to inhibit MGEs replication, plasmid mostly, by blocking the movement of replication fork or transcription without cleavage[148].

Multiple subunits of the effector work together in a coordinated manner to achieve the clearance of target DNA or RNA in class 1 systems. Class 2 systems, on the other hand, use a well-rounded single-unit effector to perform interference.

Cas9 of type II systems is comprised of nuclease (NUC) and recognition (REC) lobes. The formation of Cas9 RNP complex results in a conformation change of Cas9 and the complex is activated to scan PAM sequences in DNA[149]. The recognition of PAM via the PAM interaction (PI) domain bends dsDNA and triggers R-loop formation. Like the R-loop initiation in type I systems, the R-loop formation starts at the seed sequence proximal to PAM and extends to the end distal to PAM[150]. The formation of R-loop leads to a conformation change of NUC, in which the HNH domain is placed for complementary strand cleavage while the non-target strand is cleaved by RuvC domain[150, 151]. Hence, dsDNA break is introduced to the target by Cas9 RNP complex (Figure 1.14).

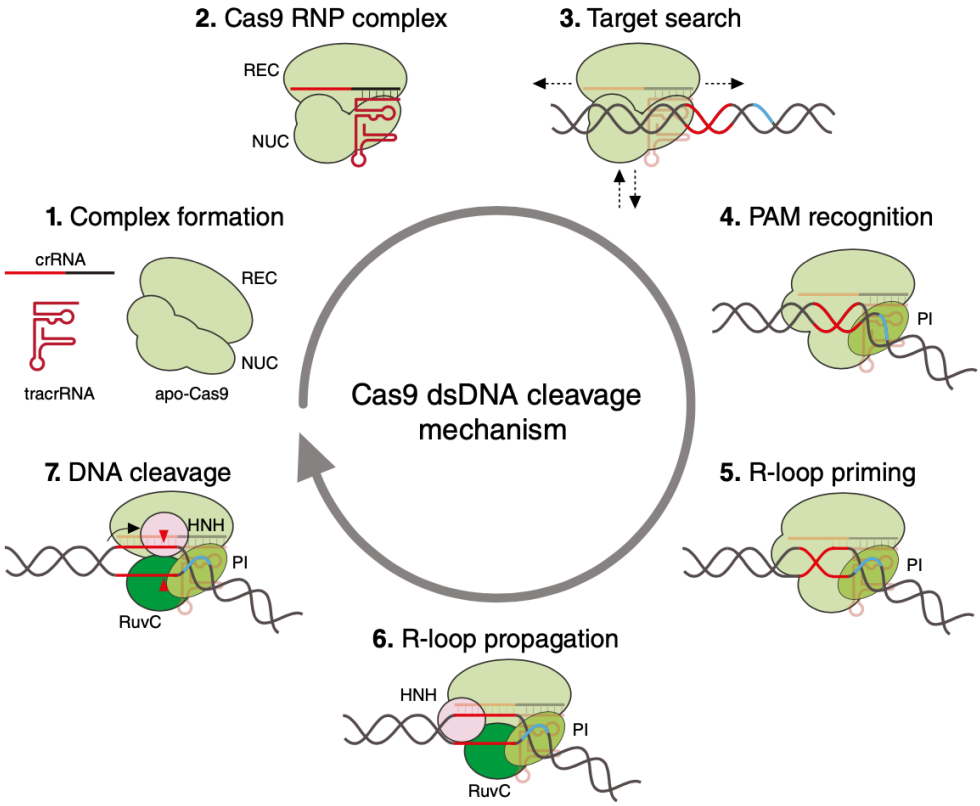


Figure 1.14 type II interference

1. Two lobes, REC and NUC, comprised Cas9, and stay relaxed without crRNA-tracrRNA duplex binding (apo-Cas9); 2. Cas9 RNP complex has a conformation change and is activated; 3. The RNP complex searching target by 3D collision and 2D diffusion; 4. PAM is recognised by PAM interaction (PI) domain; 5. R-loop formation is initialized by seed sequence priming; 6,7 R-loop formation drives a conformation change of NUC, where HNH domain cleaves target strand and RuvC domain cleaves non-target strand. Adapted from Tautvydas Karvelis and Virginijus Siksnys[125].

Type V interference shows the diversity of type V CRISPR. The target range of type V systems is broad, including dsDNA, ssDNA and ssRNA. Cas12 RNP complex can target dsDNA with PAM sequence, in this case, no tracrRNA is present, generating a staggered dsDNA break with a 5' overhang on the target strand (Figure 1.15A)[126, 152-

[154]. Some subtypes of type V CRISPR target ssDNA or ssRNA with tracrRNA-crRNA-Cas12 complex. PAM recognition is not necessary for single strand targeting (Figure 1.15B)[127, 155]. In addition, target-activated Cas12 shows non-specific single-strand nucleic acid cleavage. The collateral cleavage is carried out by Cas12 in both dsDNA targeting and single-strand nucleic acid targeting[127, 156].

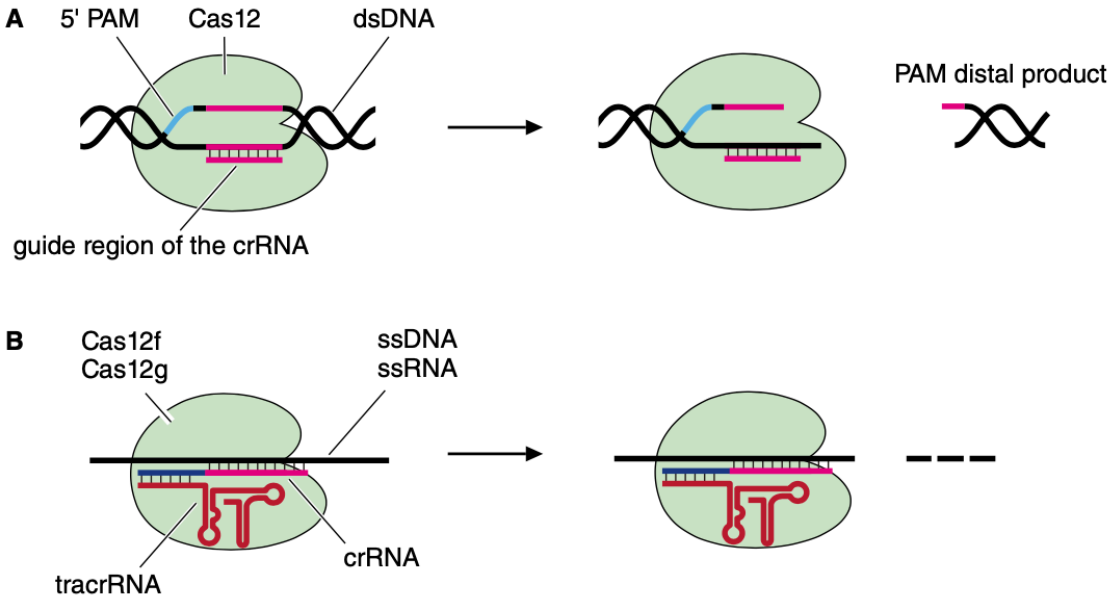


Figure 1.15 type V interference

(A) Cas12 RNP complex recognises 5' PAM in dsDNA and produces staggered dsDNA break. **(B)** Subtypes of type V systems target ssDNA or ssRNA with an RNP complex that contains a tracrRNA-crRNA duplex. Adapted from Morgan Quinn Beckett et al[130].

The target of type VI systems is RNA. When phage infects host cells, phage RNA or transcripts from phage DNA is captured by the Cas13 RNP complex and cleavage by Cas13 RNase domain[157]. In addition to the target cleavage, collateral cleavage of RNA is triggered upon target ssRNA binding. Non-specifical degradation of host mRNA

leads to abortive infection which arrests cell growth and restricts phage infection (Figure 1.16)[131].

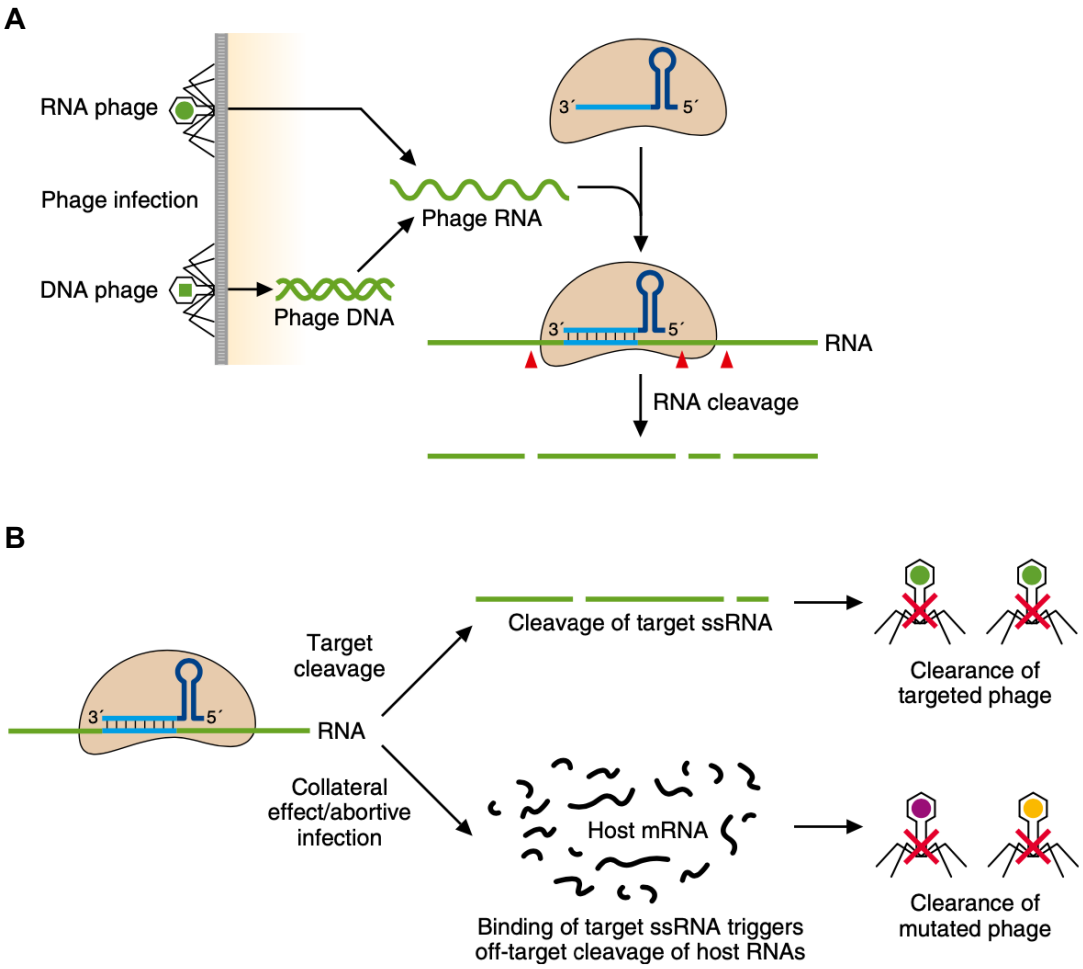


Figure 1.16 type VI interference on phage infection

(A) Cas13 RNP complex recognises phage RNA and cleaves target RNA. **(B)** Cas13 target cleavage clears phage ssRNA; In addition, collateral cleavage is activated to degrade host RNA to retain the phage infection; Mutated phage is eradicated via the collateral effect. Adapted from Omar O. Abudayyeh and Jonathan S. Gootenberg[132].

1.3 Applications of CRISPR

From what we have discussed above, CRISPR systems can be viewed as adaptive immune systems that target nucleic acid in a sequence-specific manner via an RNA guide. Two keywords from CRISPR are 'sequence-specific' and 'nucleic acid targeting'. It opens the opportunity to repurpose this prokaryotic immune system for gene engineering. The last decade has seen the development of CRISPR-Cas systems in gene editing, and myriad achievements have been made. In this section, we will introduce the principle of utilising CRISPR-Cas systems in gene engineering and demonstrate successful cases of CRISRP-Cas application.

1.3.1 The principle of CRISPR-Cas application

CRISPR interference results in DNA or RNA cleavage, which eradicates invasive nucleic acid and builds immunity in the prokaryotes. The fundamental role of CRISPR-Cas can be repurposed to target genome, leading to gene depletion. Taking a step further, when targeting host genome, and generating dsDNA break (DSB), host DNA repair systems take actions to repair DNA damage, and the outcome of repair is controllable via manipulating the repair process. This established the base for utilising CRISPR-Cas systems in desired gene editing.

Cas9 is the most broadly applied CRISPR-Cas system owing to its simplicity. We will take Cas9 as a paradigm to explain how CRISPR-Cas facilitate gene engineering. CRISPR-Cas9 is introduced to eukaryotes and exhibits target editing via a pre-designed guide RNA(gRNA)[158, 159]. Cas9 produces DSB upon binding to target dsDNA, nonhomologous end joining (NHEJ) is the major pathway to repair the DNA damage

created by Cas9 targeting in eukaryotes[160]. NHEJ requires KU70-KU80 heterodimers that bind to DNA ends, preventing DNA resection. DNA ligase IV is then recruited to ligate the damage ends with 0 to 4 nt microhomology at the damage site. But if the damaged DNA has already been processed before KU protection and ligation is not available, nucleases can be recruited to process DNA, making it available for ligation. In this process, insertions or deletions (indels), are incorporated into the damaged DNA. If the repair is faithful, Cas9 keeps cleaving target dsDNA until unfaithful repair takes place (Figure 1.17A)[161, 162]. In the absence of KU, the alternative end joining (a-EJ), also termed microhomology-mediated end joining (MMEJ), carries out DNA repair by end resection and ligation. The resection is driven by CtIP-MRN, which generates 5 to 25 bp of microhomology used for following ligation by DNA ligase IIIα (Figure 1.17A)[163, 164]. The “error prone” property of NHEJ and a-EJ enables Cas9 editing to introduce mutation or deletion at target site. However, the outcome of editing is not well controllable. Homology-directed repair was exploited to perform a more precise editing. Compared with NHEJ, KU-mediated end protection is not present in HDR, instead, CtIP-MRN resects DNA ends from DSB, replication protein A (RPA) binding to DNA strand to prevent fusing of overhang. RPA is then replaced by RAD51, enabling repair with a donor dsDNA template (Figure 1.17B)[165, 166]. The donor templates provide engineers with a controllable platform where desired editing can be inscribed. Ideally, any desired outcomes can be gained by HDR, but HDR suffers a low efficiency in practical applications since the damage of DNA is primarily repaired via NHEJ rather than HDR *in vivo*[167]. One direction of Cas9 editing development is enhancing HDR frequency or inhibiting NHEJ.

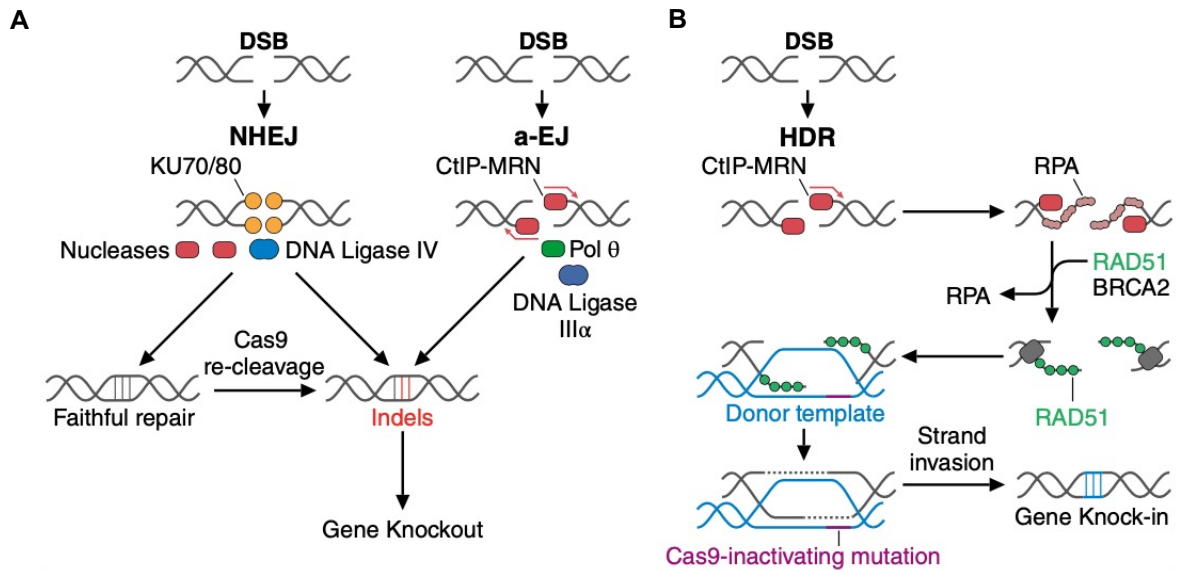


Figure 1.17 Genome editing strategies by exploiting endogenous DNA repair pathways.

(A) dsDNA break (DSB) created by Cas9 targeting activates endogenous DNA repair; Nonhomologous end joining (NHEJ) is a repair pathway where KU70/80 protect DNA ends from resection, and DNA ligase IV ligate the end of DNA to achieve repair; DNA ends can be processed before KU protection, which leaves an end that is not available for ligation, nucleases process it to fit DNA ligation, in this process, insertions or deletions can be introduced to achieve gene knockout. Alternative end joining (a-EJ) is another pathway that repairs DNA damage with Pol θ and DNA ligase IIIα, but the DNA end is not protected by KU, instead, resected by CtIP-MRN. **(B)** Homology-directed repair (HDR) pathway also utilises resected DNA ends, but the overhangs are protected by replication protein A (RPA) from binding. RPA is subsequently replaced by RAD51 under the meditation of BRCA2. RAD51 enables damaged DNA to be repaired with a donor template that can be provided artificially. Cas9-inactivating mutation can be introduced into the donor to cease editing after a successful gene knock-in. Adapted from Peter Lotfy and Patrick D. Hsu[168].

Through the exploitation of endogenous repair pathways, Cas9 editing is capable of modifying the genome, but there are three concerns in practical application: off-target effects, potential toxicity of DSB and various outcomes of editing. To overcome those detriments and improve Cas9 editing efficiency, approaches that generate gene editing without DSBs have been explored. The first step is to abolish Cas9 endonuclease activity but keep its ability of recognition. Mutations in Cas9 nuclease active sites convert Cas9 into either a nickase (Cas9n) or catalytically dead Cas9 (dCas9), and these two mutants are capable of targeting dsDNA without generating DSBs[158, 169]. The Cas9 variants were further fused with ssDNA deaminases that convert DNA bases (C to T or A to G), enabling mutation in target sequence, which is termed base editing[170]. Base editing changes DNA bases without introducing DSBs, but the conversion is limited to specific bases (C to T or A to G). A different method, primer editing was developed to expand the range of editing outcomes. In primer editing, Cas9 nickase variant is fused to a reverse transcriptase, and a primer editing gRNA (pegRNA) that carries a desired editing sequence in addition to basic gRNA allows the nicked strand to be repaired by reverse transcription based on the provided sequence[171]. The base editing and primer editing without DSBs may get around the toxicity of gene editing and have potential to be widely applied.

1.3.2 Utilising CRISPR-Cas systems

CRISPR-Cas application has touched every corner of modern biological study. Fundamental research has been shaped owing to the development of CRISPR-Cas. One core question in genetic studies is the association between genetic variants and phenotypes. Gaining the variants is essential for studying the functional association.

CRISPR-Cas9 has been used to generate variants that can be used for phenotyping and functional study (Figure 1.18A). One example is the genetic study of late-onset Alzheimer's disease (LOAD). The *APOE* gene is polymorphic and related to differential risk for LOAD. Researchers applied Cas9 to target *APOE* gene with a donor single-strand oligodeoxyribonucleotides (ssODNs), which generates variants that can be used for functional comparison with wild-type. The variant shows many features associated with neurodegeneration, revealing the *APOE* functional role in LOAD[172, 173].

CRISPR-Cas9 has also been used in transgenic animal model generation. Cas9 RNP is electroporated or microinjected into animal zygotes with or without donor DNA. The effector RNP cleaves target locus before cell division, which reduced the chance of generating mosaic editing outcomes, improving the efficiency compared with canonical workflow of transgenic animal model production (Figure 1.18B) [174-176].

Another expertise of CRISPR-Cas gene engineering is multiplexed genome editing which requires targeting multiple genes simultaneously. CRISPR-Cas processes pre-crRNA into mature crRNA where multiple spacers with different targets can be generated, enabling multi-targeting. Cas12a is one of the most streamlined Cas proteins in multiplexed genome targeting since Cas12 generates mature crRNA without host RNase or a tracrRNA compared with Cas9 (Figure 1.18C) [126, 128].

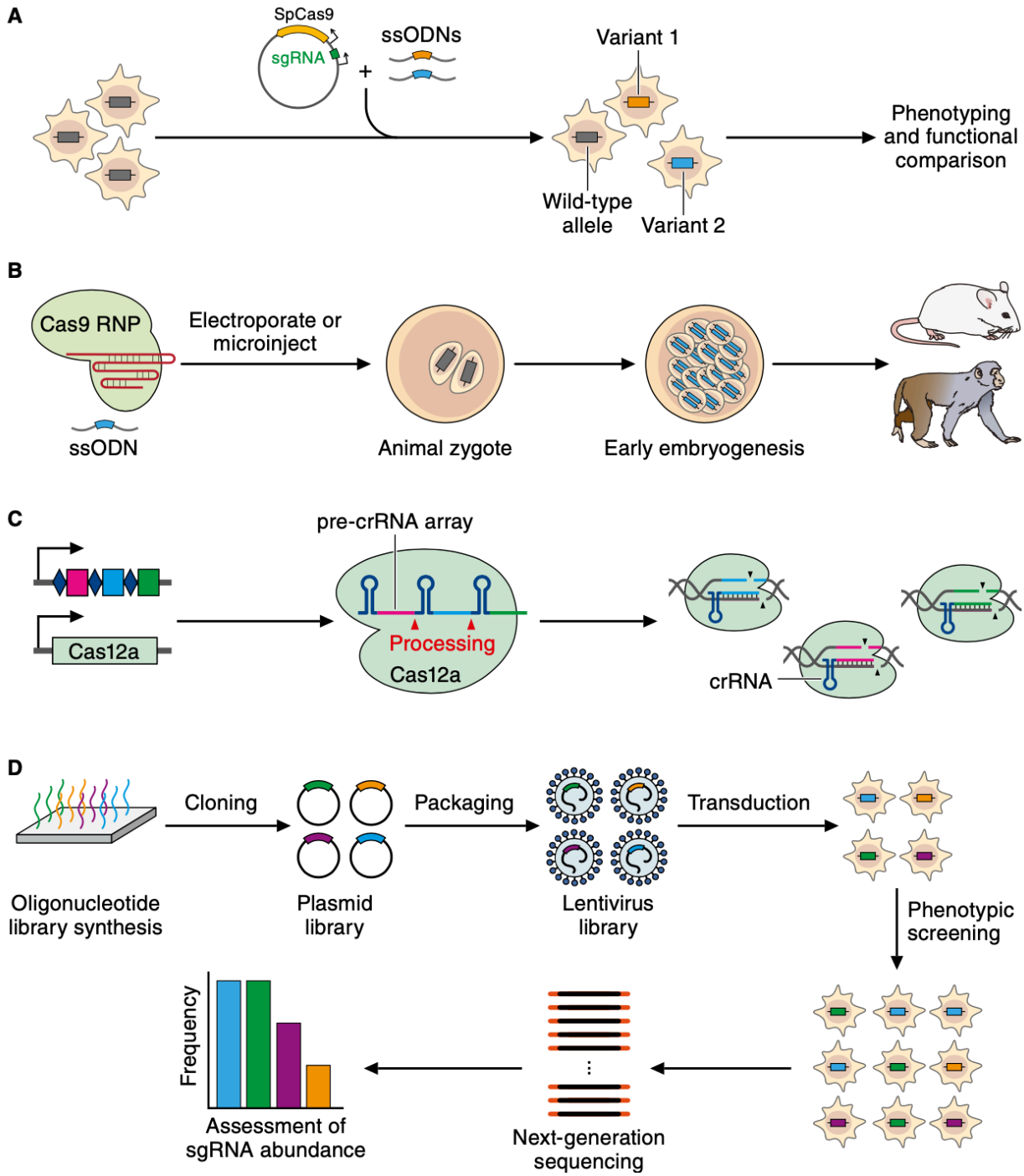


Figure legends next page.

Figure 1.18 CRISPR-Cas application in fundamental research

(A) CRISPR-Cas9 is applied in genetic studies; SpCas9 (Cas9 from *Streptococcus pyogenes*) is transfected into cells with donor (ssODNs) to target specific genes and generate variants that can be used for phenotyping and functional comparison. **(B)** Cas9 application in Transgenic animal models; Cas9 RNP with or without donor is electroporated or microinjected into animal zygotes to edit target gene before cell division. **(C)** Cas12a is used for multiplexed gene editing; Cas12a processes pre-crRNA into gRNA, which produces different spacers that drive the multiplexed gene editing simultaneously. **(D)** Genome-scale screens by CRISPR-Cas; Oligonucleotides encoding sgRNA (single guide RNA) are synthesized for broad genome targeting; They are cloned into a plasmid with Cas gene and further packed into lentivirus that is transduced into cells; Phenotypic screening with selective pressure generates cells for next-generation sequencing; Assessment of sequencing data reveals the differences of sgRNA abundance between the treated population and initial population, pointing to the essential genes in the specific process. Adapted from Peter Lotfy and Patrick D. Hsu[168].

Genome-scale screening has been reshaped by the introduction of CRISPR-Cas in recent years. One well-presented example is gene screening in cancer cells. The screens start with an oligonucleotide library where thousands of oligonucleotides that encode sgRNA, targeting various genes, are synthesized. The oligonucleotides are further cloned into vectors that express Cas proteins, and the vectors are packed into lentivirus which acts as a cargo and is introduced into cells by transduction. Applying selection pressure on cells and performing phenotypic screening allow subsequent next-generation sequencing to gain data for assessment of sgRNA abundance, either depletion or enrichment compared with the initial population, thereby revealing essential genes for different process[177].

In addition to the genome-editing, CRISPR-Cas systems have also been repurposed for programmable gene regulation. The dCas9 (catalytic dead Cas9 variant) possesses targeting ability while losing the endonuclease activity. When relocated dCas9 RNP to the upstream area of a gene of interest, it inhibits target gene transcription by sterically blocking the RNA polymerase (RNAP) (Figure 1.19A)[178, 179]. By fusing an effector (either activator or repressor) to dCas9, extensive gene repression or activation can be achieved (Figure 1.19B)[180].

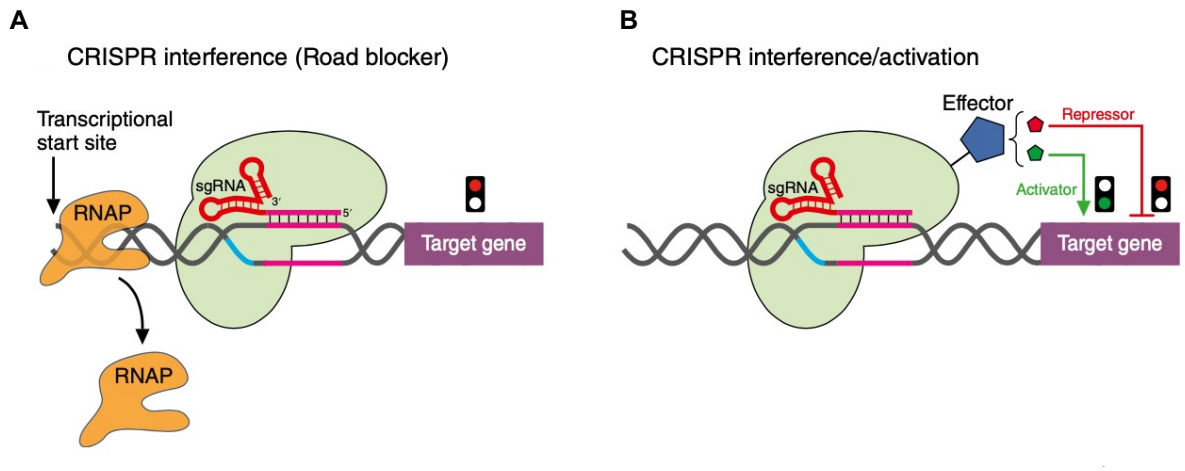


Figure 1.19 Programmable gene regulation by CRISPR-Cas systems

(A) dCas9 RNP binding to the upstream of target gene blocks transcription by displacing RNAP. **(B)** dCas9 RNP is fused to an effector that either activates or represses target gene expression. Adapted from Jasprina N. Noordermeer, Crystal Chen, and Lei S. Qi[181].

CRISPR-Cas is a prokaryotic immune system that protects host from MGEs invasion. If the CRISPR is artificially manipulated to target its own chromosome (self-targeting), host cells would be killed[182]. This has led to the utilisation of CRISPR-Cas systems for

antimicrobials development. Both class 1 and class 2 CRISPR are capable of forcing cells into committing “suicide”. Several studies have been proposed to repurpose CRISPR-Cas for antimicrobials[183, 184].

We focused on applying the interference of CRISPR-Cas to various aspects of modern biology research. But the vault of CRISPR is not only interference stage, the adaptation stage of CRISPR also holds its unique property that can be exploited for biological information recoding[185, 186]. CRISPR adaptation apparatus, Cas1-Cas2, capture DNA and write it into CRISPR array. If the Cas1-Cas2 was leveraged for capturing DNA whose expression was regulated based on biological signals, this temporal biological information can then be recorded into genome, following sequencing, the temporal information attributed to signal change is able to be of readout. Under this rationale, a workflow was developed where biological signals induce the expression of intracellular DNA that is subsequently recorded by Cas1-Cas2 into cell genome. This method provides a way to record dynamic information of cell status[185].

It is no more than two decades since the discovery of CRISPR-Cas systems, but CRISPR-Cas has exhibited its omnipotent potential. In 2021, CRISPR-Cas9 mediated genetic therapy has been reported in transthyretin amyloidosis (ATTR)[187], transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)[188]. Although this pioneering work is restricted to a small group of patients or individuals, the results support further research on applying the gene editing therapeutic strategy in genetic disease.

1.4 Type I CRISPR-Cas systems

1.4.1 Subtypes of type I CRISPR-Cas

Type I systems are the most abundant CRISPR-Cas in the microbial world[86]. The signature gene of type I systems is *cas3*, which encodes the effector to degrade target dsDNA. Type I systems also possess a stunning diversity in terms of gene compositions and operonic organizations. There are 7 subtypes of type I systems (type I-A to type I-G) with distinct features (Figure 1.20)[91]. The prototype of type I consists of 8 *cas* proteins (Cas1 to Cas8), with which type I-B fits the best, where Cas1, Cas2 and Cas4 carry out the adaptation[189], while the rest of *cas* proteins form the effector complex for interference[190].

Type I-A is considered as a derivative of type I-B, where Cas3 is split into Cas3' and Cas3'', and the large subunit of the effector complex (Cas8) is split into a small subunit (Cas11) and the remaining large subunit[86]. Experimental studies have revealed the adaptation of type I-A that also involves Cas1, Cas2 and Cas4[191]. Cas3 helicase domain and nuclease domain are split into two distinct units, Cas3' and Cas3'' respectively, and reconstitution of type I-A *in vitro* shows interference activity[192]. The structure of type I-A effector complex has been elucidated, revealing a well-organized complex that contains Cas5, Cas6, Cas7, Cas8, Cas11 and even Cas3[193].

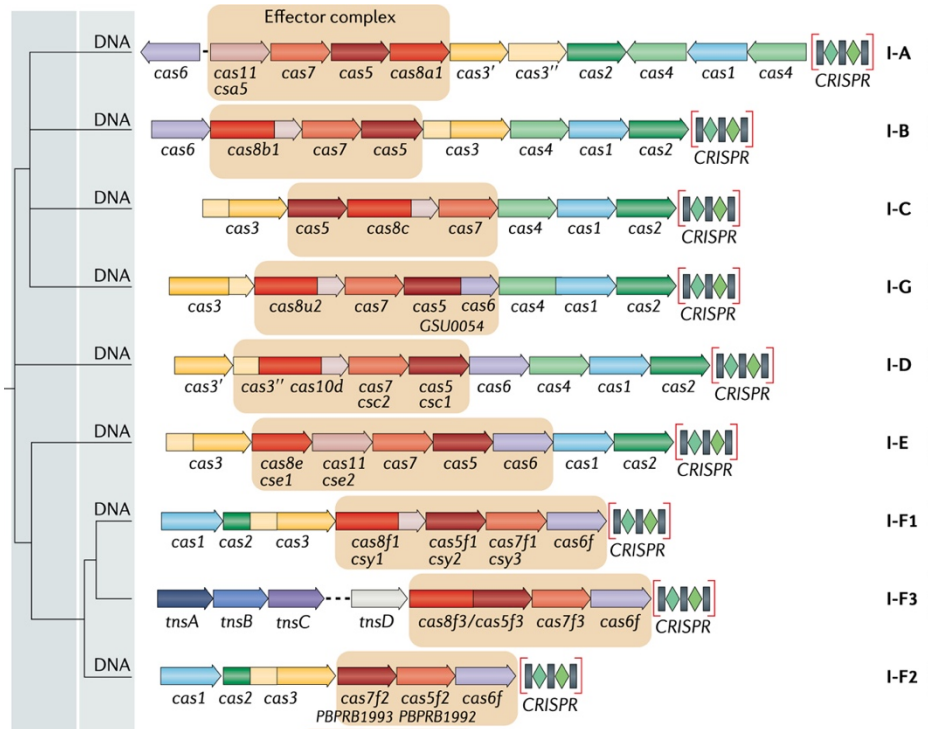


Figure 1.20 Subtypes of type I CRISPR

Representative CRISPR cluster of type I-A to type I-G; *cas3*, Signature gene of type I; *cas1*, *cas2* and *cas4*, CRISPR adaptation genes; *cas5*, *cas6*, *cas7*, *cas8* and *cas11*, crRNA maturation and effector complex formation. I-B, the prototype of type I CRISPR, contains *cas1* to *cas8*; I-A, signature gene: split *cas3*, *cas11* present; I-C, lack of *cas6*; I-D, signature gene: *cas10d*; I-G, signature gene: *csb2* (fusion of cas5 and cas6); I-E, lack of cas4, *cas11* present; I-F signature gene: fusion of *cas2* and *cas3*; Variants of I-F: lacks *cas3* and utilised by transposons or lacks *cas8* and *cas11*. Adapted from Kira S. Makarova[91].

Type I-C lacks *cas6* gene compared with other type I systems. Cas6 is the protein that cleaves pre-crRNA into mature crRNA, however, this role has been substituted by Cas5 in type I-C[194]. The structure of type I-C effector complex shows that a small subunit, translated from a Cas8 inner ORF, is critical to stabilizing R-loop formation[195]. Type I-D is considered to be an evolutionary intermediate between type I and type III

since the nuclease domain of Cas3 is shifted to the larger subunit Cas8, which shares the similarity to type III Cas10, accordingly the large subunit is termed Cas10d[86]. Experimental data shows that type I-D cleaves both dsDNA and ssDNA, which holds both canonical type I dsDNA cleavage and type III ssDNA cleavage feature[196].

Type I-E is the first elucidated CRISPR system[83]. The study on type I-E reveals fundamental biological mechanisms of CRISPR immunity, which has been described in **section 1.3**. Compared with the prototype I-B, I-E cluster contains distinct large subunit Cas8 and small subunit Cas11, similar to type I-A. And no Cas4 is present in the adaptation gene locus[86].The signature gene of type I-F is the fusion of cas2 and cas3. However, there are variants of type I-F that lack *cas3*, losing the ability of interference. They are instead utilised by transposons for RNA-guided DNA integration[197]. Structural studies have been carried out on type I-F[198], type I-F transposon variant[199] and a variant that lacks both Cas8 and Cas11[200]. The structure of the transposon variant shows that Cascade of type I-F binds to transposition protein, which provides the basis for RNA-guided transposition[199]. The variant that lacks Cas8 and Cas11 is structurally demonstrated, showing that Cas5 and Cas7 replace the functional role of Cas8 and Cas11, the Cas5 and Cas7 in this variant type I-F is structurally different from the canonical ones[200].

The last subtype of type I CRISPR is type I-G whose signature gene is *csb2*, a fusion of *cas5* and *cas6*. cas4 and cas1 are also fused into one gene in type I-G. Type I-G is used to be named type I-U, U stands for uncharacterized[86]. It is the least understood subtype among all type I systems. The details of type I-G interference and underlying mechanisms remain unexplored.

1.4.2 Cas3, the signature protein of type I CRISRP

Cas3 is the signature protein of type I CRISPR systems. It is a dual-functional protein with HD-nuclease and helicase activity[201]. The nuclease activity is activated by single-strand DNA (ssDNA), and the double-strand DNA (dsDNA) helicase activity requires the presence of ATP and ssDNA. The Helicase domain of Cas3 is a typical superfamily 2 (SF2) helicase with Walker A and B boxes involved in the binding and hydrolysis of ATP[201-203]. *In vitro* reconstitution of type I-E Cascade, coupling with Cas3 from *Streptococcus thermophilus*, shows that target dsDNA is specifically degraded, and the direction of Cas3 cleavage activity is 3' to 5' of DNA[204, 205]. Cascade-Cas3 initialises the target dsDNA degradation by nicking in the proto-spacer of the non-target strand (NTS), which generates ssDNA as substrates for Cas3 nuclease. The effector complex stays in place while target dsDNA is unwound by Cas3 helicase activity and NTS is reeled to Cas3 nuclease domain, subsequently degraded[104, 203]. The degradation of NTS results in the exposure of target stand (TS) that acts as a platform for loading the same or other Cas3, following further degradation[204]. The Cascade-Cas3 is functionally activated when introduced to a heterologous organism, for instance, *S. thermophilus* type I-E system provides heterologous protection in *E. coli*[206].

1.4.3 Application of type I CRISPR-Cas

CRISPR application is often taken as a synonym for CRISPR-Cas9. However, the type II CRISPR, Cas9 is only the tip of the iceberg in natural CRISPR occurence. In fact, Type I CRISPR is the most prevalent (50%) CRISPR system in bacteria[86]. So, instead

of pursuing the heterologous editing tool, repurposing endogenous type I systems for self-genome editing is performed in practice.

Industrial bacterial strains and medical pathogens have been modified by endogenous type I CRISPR. Examples include type I-F in *Zymomonas mobilis*[207], an ethanol-producing bacterium; type I-B in *Clostridium tyrobutyricum*[208], a butanol-producing strain; type I-C in the pathogenic bacterium, *Pseudomonas aeruginosa*[209]; type I-B in *Clostridium difficile*, a human pathogen[210].

Not only for the endogenous application, but type I CRISPR can be also introduced into heterologous organisms for genome editing. Type I-C and I-F from *Pseudomonas* are packed and transformed into heterologous bacterial hosts to achieve genome engineering[209, 211]. Furthermore, there are successful utilizations of type I in eukaryotic organisms[193, 209, 212-218].

Compared with small DSB generated by Cas9 editing, type I CRISPR destructively cleaves the target DNA, generating long-range deletion on genome targeting (Figure 1.21A)[212, 214]. This deletion outcome can be altered by removing Cas3, in this manner, Cascade targets dsDNA and blocks binding from other proteins, RNAP for instance, leading to repressing the expression of target gene (Figure 1.21B)[219, 220].

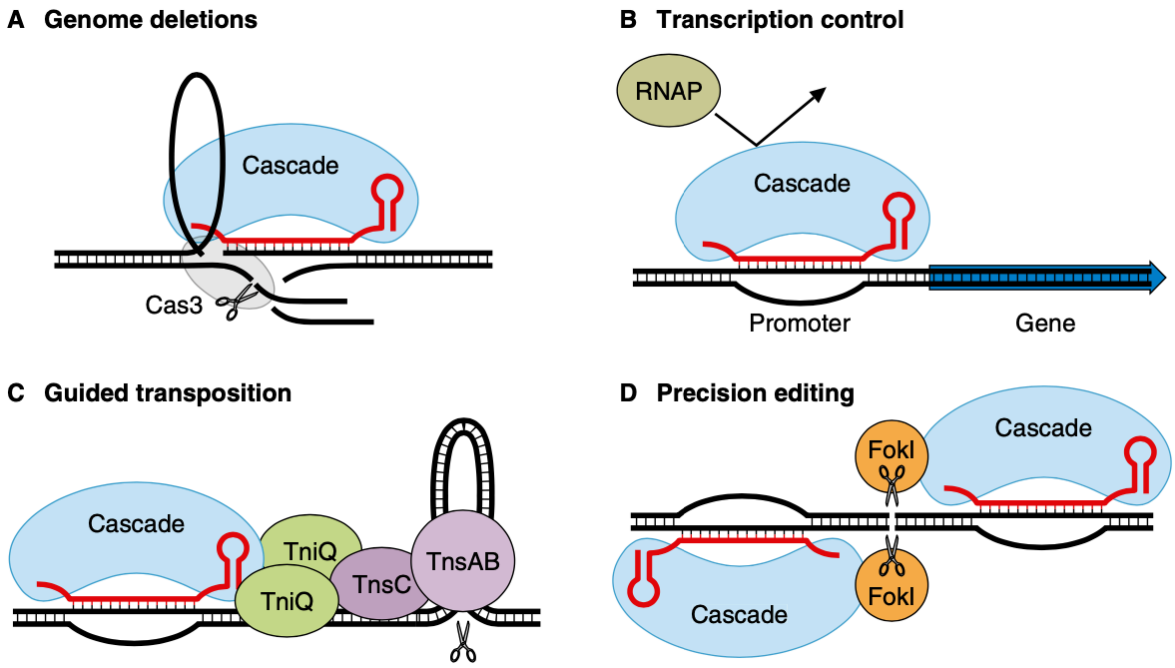


Figure 1.21 Type I CRISPR application

(A) Genome deletion by type I Cascade and Cas3. **(B)** Cascade binds to target promoter to block the binding of RNAP (RNA polymerase). **(C)** Genome editing with type I Cascade and transposons associated proteins. **(D)** Cascade is fused to FokI, a restriction enzyme; the dimeric Cascade/crRNA-complex allows target editing in the centre of the DNA. Adapted from John van der Oost[120].

As mentioned above, there are variants of type I systems that couple with transposons. This combination allows donor dsDNA to be incorporated into target site without DSB, which has potential to generate genome editing under lower toxicity and higher efficiency (Figure 1.21C)[197]. One example of type I precision editing is to fuse a restriction enzyme, FokI, to Cascade. By using a dimeric Cascade/crRNA-complex, the target editing on the centre site flanked by the complex can be obtained (Figure 1.21D)[221].

1.5 Significance and aims of the thesis

The fundamental studies on CRISPR systems reveal an elegant machinery of the prokaryotic adaptive immune system. The discovery of underlying mechanisms drives the repurposing of CRISPR systems in gene engineering. Modern genetic modifying technologies have been shaped since the appearance of CRISPR. It has been extensively applied in every aspect of biological research. However, there are still mysterious areas existing in this exciting field.

Type I-G CRISPR, the protagonist of this thesis, is the least understood type I CRISPR. It possesses a unique gene organisation and a signature gene, but the expression and interference of type I-G CRISPR have not been investigated. With the study on type I-G, we will elucidate its basic biological mechanism and complete the understanding of type I CRISPR systems. It will also provide opportunities to utilize this system in gene engineering.

The aim of the thesis:

1. Elucidating the expression and interference of type I-G CRISPR.
2. Obtaining the structure of type I-G effector complex.
3. Application of type I-G CRISPR in heterologous genome editing.

Chapter 3 focuses on the reconstruction of type I-G CRISPR for mechanism investigation. In Chapter 4, we will reveal the structure of type I-G effector complex. Chapter 5 will open the door to applying type I-G in genome editing.

Chapter 2: Materials and methods

2.1 In vitro and in vivo construction of type I-G system

2.1.1 Cloning

2.1.1.1 Vectors for single Cas protein expression

Synthetic genes encoding Cas proteins (Cas8g, Csb2, Cas7 and Cas3 from *Thioalkalivibrio sulfidiphilus*) were obtained from Integrated DNA Technologies (Coralville, IA, USA). Restriction enzyme sites were added when necessary and codon usage was optimized for *Escherichia coli*. *cas8g*, *cas7* and *cas3* genes were digested with *Nco*I and *BamH*I (Thermo Scientific) and ligated into pEV5HisTEV[222] to produce the vector that allows expression of individual proteins with N-terminal TEV cleavable His_8-tags. *csb2* was cloned into the pET-Duet (Novagen, Merck Millipore) vector via ligation after *Nde*I and *Xho*I (Thermo Scientific) digestion. Site directed mutagenesis of cas genes was carried by standard protocols using Phusion enzyme (Thermo Scientific).

2.1.1.2 Vectors for pre-crRNA generation

A CRISPR array containing six identical spacers targeting the *tetR* gene flanked by seven repeats was cloned into pCDF-Duet (Novagen, Merck Millipore) by ligation after *Nco*I and *Sal*I digestion. The other two CRISPR arrays: *lacZ* target (five repeat, four spacer) and *lpa* target (four repeat, three spacer) were prepared using the same method. Sequence details in Table 1.

2.1.1.3 Vectors for multiple Cas proteins expression

To express the type I-G complex for *in vivo* studies, vector pACE-M1 (MultiColi™, Geneva Biotech, Genève, CH) was assembled by SLIC (sequence and ligation independent cloning). DNA fragments encoding *cas8g*, *csb2* and *cas7* were amplified with PCR prior to SLIC and ligation into pACE, placing these three genes under control of a single T7 promoter. The *cas*3 gene was digested with *Nco*I and *Sal*I and incorporated into vector pRAT under the control of the araBAD promoter. Plasmid *lacZ*-pRAT was described previously[223]. All final constructs were verified by sequencing (GATC Biotech, Eurofins Genomics, DE)

2.1.2 Oligonucleotides

All 6-FAM™-labelled and non-labelled DNA or RNA substrates were purchased from Integrated DNA Technologies (Leuven, BE). Where required, oligonucleotides were 5'-end-labelled with [γ-^{32}P]-ATP (10 mCi ml^{-1}, 3000 Ci mmol^{-1}, Perkin Elmer) with polynucleotide kinase (Thermo Scientific). Duplex DNA was obtained by annealing equimolar amount of complementary ssDNA in 10mM Tris-HCl, 50mM NaCl, pH7.5, 95 °C for 5min, slowly cooling down overnight to room temperature in a heat block. All oligonucleotide sequences can be found in Table 1.

2.1.2.1 Oligonucleotides purification

Purchased oligonucleotides were submitted for gel purification before using. Lyophilised DNA substrates were resuspended in TE-NaCl buffer (10 mM Tris-HCl, 1 mM EDTA, 10 mM NaCl) or double distilled water (ddH_2O), and RNA substrates in

RNase-free H_2O to a concentration of 500 µM and stored at -20℃ until required. 2 µl of the oligonucleotide (500 µM) was diluted with 8 µl RNase-free water and 10 µl denaturing loading buffer (100% formamide, no dye, a dye loading buffer (100% formamide, 0.25% bromophenol blue and 0.25% xylene cyanol) was running alongside the no dye lane as a visual sign) and the mixture was heated at 90℃ for 5 min. The solution was then cooled on ice before being loaded on a pre-run denaturing polyacrylamide-TBE gel (20% polyacrylamide, 7 M urea). Gels were run in 1X Tris-Borate-EDTA running buffer (100 mM Tris (pH 8), 90 mM M boric acid, 1 mM EDTA) at 30W and 45℃ for between 1.5 and 3 hours, depending on oligonucleotide length. Substrates were visualised using UV shadowing (Minerallight USV-54 UV wand) and the substrate band was excised. The gel band was soaked in 400 µl TE-NaCl buffer/ddH_2O/RNase-free H_2O overnight at 4℃. The supernatant was then decanted and filtered before the nucleic acid was extracted by ethanol precipitation as described below.

2.1.2.2 Ethanol precipitation

Ethanol precipitation of DNA/RNA substrates was carried out by adding 2 volumes of cold (4℃) 100% ethanol and 0.1 volume of 3 M (pH 5.2) sodium acetate. The solution was then centrifuged at 13,200 rpm and 4℃ (Eppendorf fixed angle F-45-24-11 rotor) for 30 min, before the supernatant was decanted. 2 volumes of cold 70% ethanol was added to the nucleic acid pellet and the solution was centrifuged for a further 30 min (Eppendorf fixed angle F-45-24-11 Rotor, at 13,200 rpm). The ethanol was carefully decanted and the pellet was air-dried and resuspended in the desired volume of RNase-free water (RNA substrates) or TE- NaCl buffer.

2.1.2.4 *In vitro* transcription and RNA extraction

In vitro transcription was performed with MEGAscript™ T7 Transcription Kit, Invitrogen™. In details, linear DNA templates containing a T7 promoter were amplified from pCDF, pre-crRNA vectors. 0.1 to 0.2 µg template was mixed with 8 µl NTP mix, 2 µl 10X reaction buffer and 2 µl enzyme mix, added water up to 20 µl. The reaction tube was gently mixed and incubated at 37°C for 4 hours. The volume of product was adjusted to 180 µl by adding 160 µl nuclease-free water. 20 µl of 3 M sodium acetate, pH 5.2 (Final concentration 0.3M) or 20 µl of 5 M ammonium acetate was added, mixed thoroughly. 400 µl (2x volume) of phenol:chloroform:isoamyl alcohol pH 8.0 was added and shaken vigorously inside fume hood for 15 seconds. The tube was centrifuged at maximum speed in a microcentrifuge at 4° C, for 2 minutes and the top aqueous phase was transferred to a new tube. 2x volume of chloroform:isoamyl alcohol pH 8.0 was added and shaken vigorously inside fume hood for 15 seconds, following centrifugation at 4° C, for 2 minutes. The top aqueous phase was transferred to a new tube. Glycogen (final amount 20 µg) (optional) and 2x volume of ethanol was added for ethanol precipitation as described above.

2.1.3 Protein expression and purification

2.1.3.1 Expression

pEV5HisTEV vectors harbouring *cas* genes were transformed into *E. coli* C43(DE3) for protein expression. Cells grew in LB culture containing 50 µg ml^{-1} kanamycin overnight. Following a 100-fold dilution, cells were grown at 37 °C, 180 rpm to reach an OD_{600} of 0.6~0.8, induced by 400 µM IPTG, followed by overnight protein

expression at 25 °C. The cells are harvested by centrifugation at 4000rpm, 4°C for 10 min, using a pellet scraper to remove the pellet from the 1 litre pot and weigh. Cell pellets can be used immediately or stored frozen until needed. Pellets resuspend better from frozen.

2.1.3.1 Purification

Day 1:

1. Cell pellet is defrosted and resuspended in 3-5 volumes of lysis buffer, mls per gram of pellet containing Lysis buffer (50mM Tris–HCl pH 7.5, 0.5 M NaCl, 10 mM imidazole, and 10% glycerol) supplemented with EDTA-free protease inhibitor tablets (Merck; 1 tablet per 100 ml buffer) and lysozyme (1 mg/ml).

2. Cells are lysed by sonicating for 6x1 min with 1 min rest intervals on ice at 4°C with the medium probe.

3. Lysed cells are ultracentrifuged at 40000 rpm, 4°C for 30 min to pellet cell debris. Remove tubes quickly after centrifugation to stop ‘jelly’ layer from resuspending and clogging filters.

4. Cleared cell lysate is then filtered using a 0.45 µm syringe filter to remove any precipitated material and loaded onto a 5 ml HisTrap FF crude column (GE Healthcare) equilibrated with Wash buffer (50 mM Tris–HCl pH 7.5, 0.5 M NaCl, 30 mM imidazole, and 10% glycerol). Load the lysate with the pump (prewash with water and wash buffer before equilibration and loading).

5. Unbound protein is washed away with up to 20 column volumes (CV) of Wash buffer prior to elution of the bound his-tagged protein using a linear gradient of Elution buffer (50mM Tris–HCl pH 7.5, 0.5 M NaCl, 0.5 M imidazole, and 10% glycerol). A stepped gradient AKTA program (Histrap first Nickle) was used for first nickle run.

6. After SDS-PAGE showing the trace of different eluted fractions. The fraction containing protein of interest are pooled and concentrated to 5 ml.

7. Remove the his-tag by incubating concentrated protein with TEV protease (1mg per 10mg protein) during dialysis in at least 100x volume of Wash buffer (50 mM Tris–HCl pH 7.5, 0.5 M NaCl, 30 mM imidazole, and 10% glycerol) at RT overnight. Equilibrated gel filtration column (HiLoad 16/60 Superdex pg 200, GE Healthcare) with GF buffer (20 mM Tris-HCl, 250 mM NaCl, pH 7.5) for the next day.

Day 2:

1. The TEV-cleaved protein is recovered using a 5ml HisTrapFF crude column and loaded via the 10 ml loading loop. The cleaved protein does not bind the column and therefore can be eluted using 20 ml Wash buffer (50 mM Tris–HCl pH 7.5, 0.5 M NaCl, 30mM imidazole, and 10% glycerol) into a 50 ml tube, then concentrated to 2ml. The TEV protease contains an uncleavable his-tag and therefore binds the HisTrap column. 100% Elution buffer was used to wash the bound protein off into a separate 50 ml tube to run on a gel to determine the percentage cleavage. (The AKTA program Histrap second Nickle).

2. The protein is further purified by gel filtration chromatography with the equilibrated GF column and GF buffer (20mM Tris–HCl pH 7.5, 0.25 M NaCl). (The AKTA program S200).

3. Run the eluted fractions on SDS-page, collect and concentrate the fractions containing protein of interest. Pure protein was finally concentrated in Amicon Ultra centrifugal filter (Merck-Millipore).

4. Aliquot proteins and freeze with liquid Nitrogen.

A schematic figure is shown in Figure 2.1.

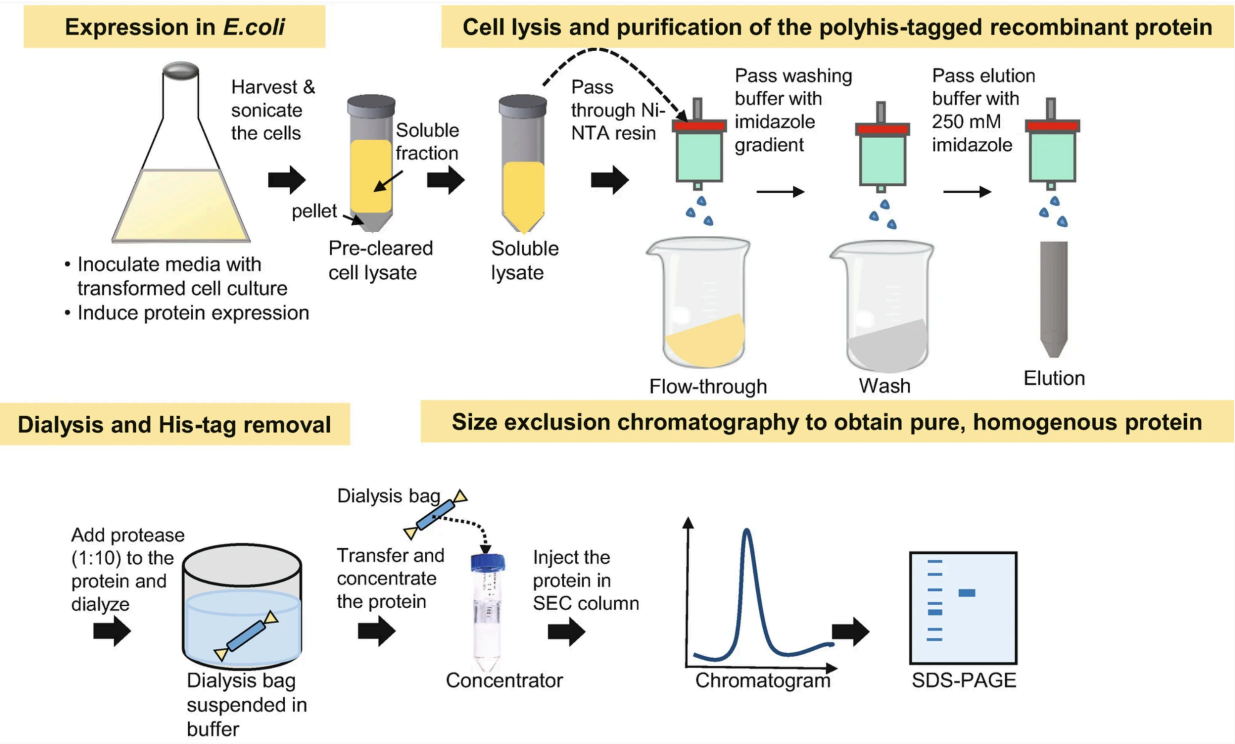


Figure 2.1 Protein expression and purification

A schematic figure protein expression and purification. Cas protein was expressed in *E. coli*. Soluble lysate after sonicating was submitted to Ni-NTA affinity column, polyhis-tagged Cas protein was eluted with a high imidazole concentration buffer. The polyhis tag of eluted Cas protein was then removed by TEV protease while dialysing. Tag removed protein was then concentrated and submitted to size exclusion column (SEC) and the eluted protein was analysed on SDS-PAGE. Adapted from Nitu Singh and Kakoli Bose[224]

2.1.4 CRISPR repeat cleavage assay

36 nt 5'-6-FAM™-labelled CRISPR repeat RNA was incubated with Csb2 or other Cas protein in 20 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, 5 mM EDTA, 0.1 U/µl RNase inhibitor, pH 7.5 for 5 min at 37 °C, at a final concentration of 50 nM RNA and 0.5 µM protein. Reaction was stopped by adding 1µl 0.5M EDTA formamide and heat denaturing at 95 °C for 3 min. Product was loaded to a denaturing gel (20% acrylamide, 7M Urea), running in 1 X TBE buffer, and visualized by scanning (Typhoon FLA 7000, GE Healthcare). RNA ladders were obtained by alkaline hydrolysis of the CRISPR repeat RNA (Thermo Fisher Scientific, RNA Protocols).

2.1.5 pre-crRNA cleavage

643nt pre-crRNA was *in vitro* transcribed with [α-^{32}P]-ATP (10 mCi ml^{-1}, 3000 Ci mmol^{-1}, Perkin Elmer) as described above, followed by phenol-chloroform extraction and ethanol precipitation. Reactions contained 1 µM pre-crRNA in reaction buffer (20 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, 0.1 mg/ml BSA, 5 mM $MgCl_2$, 1 mM EDTA, 0.1 U/µl RNase inhibitor, pH 7.5). Wildtype or variant H503A Csb2 was added to the reaction at a progressively higher concentration followed by incubation at 37 °C for 30 min. Cleavage products were resolved in a denaturing gel as described above.

2.1.6 Csb2 N-terminal domain and C-terminal domain expression

A stop codon was introduced by mutagenesis at R260 of Csb2 to allow the expression of the N-terminal domain on pEV5HisTEV vector. For the C-terminal domain, primers

with NcoI and BamHI were used to amplify C-terminal sequence fragment encoding from R262 to the end of the *csb2* gene, which was cloned into the pEV5HisTEV vector by digestion and ligation. Both N- and C-terminal domains were expressed and purified as described for the full-length proteins above. Primer sequences are shown in Table 1.

2.1.7 Fluorescence anisotropy

The method is adapted from Sarah L. Reid[225]. In detail, 25 nM 5'-6-FAM™-labelled CRISPR hairpin or repeat RNA was suspended in a quartz cuvette with 50 mM NaCl, 1 mM DTT, 5 mM $MgCl_2$, 1 mM EDTA, 0.1 U/µl RNase inhibitor (Thermo Scientific). The initial anisotropy of the RNA was measured in a Varian Cary Eclipse fluorescence spectrophotometer (Agilent Technologies), using the Eclipse ADL application. The measurement was carried at 37 °C, exciting the fluorescein-labelled RNA at 480 nm and monitoring emitted fluorescence at 525 nm. Csb2 was titrated at progressively higher concentrations into the sample. All points of titration were carried out with automatic polarizers. The anisotropy value of each point was plotted against Csb2 concentration, and a curve was fitted to get K_D value using equation below.

$$A=Amin+((D+E+K_D)-((D+E+K_D)^2-(4DE))^{1/2})(A_{max}-A_{min})/(2D)$$

Where A is anisotropy of free RNA, D is total RNA concentration, E is total protein concentration. A_{max} and A_{min} is maximum and minimum anisotropy. The equation assumes a RNA : protein binding stoichiometry of 1:1[225].

2.1.8 Effector complex reconstruction

Effector complexes were assembled by mixing individual pure protein subunits with pre-crRNA, obtained by *in vitro* transcription (MEGAscript™ T7 Transcription Kit, Invitrogen™) in the combinations noted in the results. The Cas protein combinations were incubated with pre-crRNA in 20 mM Tris-HCl, 250 mM NaCl, 1 mM DDT, 1 mM EDTA and 0.5 U/µl RNase inhibitor pH 7.5 for 1 h at 37 °C, filtered by centrifugation at 10000 x g for 10 min, then loaded onto a Superose 6 10/300 increase (GE Healthcare) column for gel filtration in GF buffer buffer (20 mM Tris-HCl, 250 mM NaCl, pH 7.5). Fractions containing the complex were pooled and concentrated using a centrifugal filter (Vivaspin® 500, MW cutoff 30,000 Dalton, Vivaproducts).

2.1.9 Electrophoretic mobility shift assay (EMSA) of short dsDNA

10 nM [γ -^{32}P]-labelled dsDNA was mixed with 0.8 µM effector complex in 10 µl reaction buffer (20 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, 0.1 mg/ml BSA, 5 mM $MgCl_2$, 1 mM EDTA) at 37 °C for 30 min, loaded onto an 8 % acrylamide 1 x TBE gel with Ficoll loading buffer, and electrophoresed at 200 V for 1 h, followed by visualization (Typhoon FLA 7000, GE Healthcare).

2.1.10 Plasmid DNA binding and cleavage assays

In vitro reconstructed effector complex was mixed with target or control plasmid (2 nM), incubated in 10 µl reaction buffer (20 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, 0.1 mg/ml BSA, 5 mM $MgCl_2$, 1 mM EDTA) at 37 °C for 1h. ATP was present at 2 mM where indicated. Reactions were analysed by separation on a 0.8 % Agarose gel, running in

1 X TBE buffer at 10 mA overnight, post-stained with SYBR green for 30 min, and visualized by scanning (Typhoon FLA 7000, GE Healthcare). Open circular plasmid control was obtained by incubation with the *Nt.BspQI* nickase (New England Biolabs), and linear plasmid by cleavage with *Bam*HI

2.1.11 Short dsDNA cleavage assay

0.3μM *in vitro* reconstructed effector complex was mixed with 16.6 nM [γ -^{32}P]-labelled dsDNA and incubated in reaction buffer (20 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, 0.1 mg/ml BSA, 5 mM $MgCl_2$, 1 mM EDTA) at 37 °C for 1h, supplemented with 2 mM ATP as indicated. Reactions were stopped by addition of 0.5 M EDTA, an equal volume of formamide and denaturation at 95 °C for 3 min. Cleavage products were separated on a S2 sequencing gel (20% acrylamide, 7 M Urea), 90W, 70 min and visualized by scanning as above. A Maxam-Gilbert G+A ladder was acquired by incubating 5 ng ^{32}P-labelled oligonucleotide with 1 μg calf thymus DNA and 0.4 % formic acid in 10 μl TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 7.5) at 37 °C for 25 min, followed by addition of 150 μl 1 M piperidine and heating at 95 °C for 30 min. The ladder was ethanol precipitated before resuspension into loading buffer and loading onto the gel.

2.1.12 Plasmid challenge assay

The method was described previously[223]. For the type I-G system, pACE-M1 and pCDF with targeting CRISPR array were co-transformed into *E. coli* C43 (DE3). Transformants were selected by 100 μg ml^{-1} ampicillin and 50 μg ml^{-1} spectinomycin.

Competent cells were prepared by diluting an overnight culture 50-fold into fresh, selective LB medium. The culture was incubated at 37 °C, 220 rpm to reach OD_{600} 0.4 to 0.5. Cells were collected by centrifugation and the pellet resuspended in an equal volume of 60 mM $CaCl_2$, 25 mM MES, pH 5.8, 5 mM $MgCl_2$, 5 mM $MnCl_2$. Following incubation on ice for 1 h, cells were collected and resuspended in 0.1 volumes of the same buffer containing 10 % glycerol. Aliquots were stored at -80 °C. pRAT Plasmid with or without Cas3 was transformed to the competent cells. Transformation mixture with LB medium was incubated with shaking for 2.5 h after heat shock. A total of 3 µl transformation product was applied in a 10-fold serial dilution to LB agar plates supplemented with 100 µg ml^{-1} ampicillin and 50 µg ml^{-1} spectinomycin when selecting for recipients only; transformants were selected on LB agar containing 100 µg ml^{-1} ampicillin, 50 µg ml^{-1} spectinomycin, 25 µg ml^{-1} tetracycline. LB agar plates containing 0.2 % (w/v) D-lactose and 0.2 % (w/v) L-arabinose were used for induction. Plates were incubated at 37 °C for 16–18 h. The experiment was performed with two biological replicates and at least two technical replicates.

2.1.13 Phage propagation

E. coli phage P1 (DSM5757) was obtained from DSMZ and stored at 4°C.

Bottom LB agar with 10 mM $MgCl_2$, 5 mM $CaCl_2$ and 10 µg/ml Ampicillin was set as bottom agar in 9 cm petri dishes. 3 ml melted top agar (LB agar : L-broth = 1 : 1, around 42°C) with 10 mM $MgCl_2$ and 5 mM $CaCl_2$ was inoculated with 1/100th volume of *E. coli* LMG194 culture (mid- to late-log phase) and $1/1000^{th}$ to 1/100th volume of phage P1 stock ($\geqslant 10^{11}$ PFU/ml) and spread onto bottom agar. The plates were incubated upside down at 37 °C for 16 h. 5 ml SM buffer (100mM Sodium chloride,

10mM Magnesium sulphate, 50mM Tris-HCl, pH 7.5) was added to overnight plates and incubated carefully at 16 °C for 15 to 30 min with gentle shaking (180 rpm). SM buffer containing the phage was collected, filter-sterilised, aliquoted and stored at 4 °C.

To determine the phage titre, 30 ml top agar as used above was inoculated with 1/100th volume of a mid- to late-log phase of *E. coli* C43(DE3) culture and spread to three plates (10ml each). 3 µl of a serial 10-fold dilution of the above prepared P1 stock in SM buffer was applied on each plate, incubating upside down at 37 °C for 16 h. Phage titre was calculated by counting the plaques.

2.1.14 Phage immunity assay

The method was described previously[226]. For the type I-G system, pACE-M1, pCDF-Lpa (CRISPR array targeting phage P1 *Lpa*) and pRAT-Cas3 were co-transformed to *E. coli* C43 (DE3). Cells were selected by 100 µg ml^{-1} ampicillin, 50 µg ml^{-1} spectinomycin and 12.5 µg ml^{-1} tetracycline. The cells were grown overnight at 37 °C in LB medium containing 50 µg ml^{-1} ampicillin, 25 µg ml^{-1} spectinomycin and 12.5 µg ml^{-1} tetracycline. The overnight culture was diluted to OD_{600} of 0.1 by LB medium supplemented with the antibiotics, 10 mM $MgSO_4$, 0.2 % (w/v) D-lactose and 0.2 % (w/v) L-arabinose. In uninduced tests, D-lactose and L-arabinose were absent. 160 µl of diluted culture was infected with 40 µl diluted bacteriophage P1 to give MOIs of 1, 0.1 and 0.01 in a 96-well plate. The OD_{600} of the culture in the plate was measured by a FilterMax F5 Multi-Mode Microplate Reader (Molecular Devices) every 20 min over 20 h. The experiment was carried out with two biological replicates and three technical replicates. The OD_{600} was plotted against time using Graphpad Prism 8.

2.2 Structure of type I-G effector complex

2.2.1 Type I-G effector complex preparation for cryo-EM

Type I-G complex containing Csb2, Cas7 and Cas8g was obtained as describe in section 2.1.8. The sample was aliquoted and flash-freezing with liquid nitrogen, storing in -80°C. Sample was sent for cryo-EM preparation in dry ice. The complex sample was diluted to 1 mg/ml for cryo-EM grid preparation.

2.2.2 Assays for Cas8g mutants

Complexes with mutated Cas8g were obtained using the same method in section 2.1.8. EMSA was described in section 2.1.9. Plasmid challenge assay and phage immunity assay were described in section 2.1.11 and 2.1.12 respectively.

2.2.3 Structure prediction

Alphafold strucute predition[227] was performed on the online server of Alphafold2: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb. The protein sequences were entered in “query_sequence” and parameters was set to default. For Csb2 structure comparison, Alphafold generated predicted structure was submitted to DALI server[228] for PDB search. The hits were ranking by Z-score. Structure was analysed in ChimeraX[229].

2.3 Genome editing in prokaryotes by type I-G system

2.3.1 Cloning

2.3.1.1 Genome targeting vectors

For *E. coli* genome editing, the pM2 vector was constructed based on the pACE-M1. The original T7 promoter was replaced by an araBAD promoter using overlap PCR extension with overlap primers. Restriction sites (*NcoI* and *SalI*) were introduced for further construction. The *cas3* gene was digested with *Nco*I and *Sal*I (Thermo Scientific) and ligated into the promoter-swapped vector to generate the pM2 vector. Site directed mutagenesis of the *cas3* gene in pM2 was carried out using standard protocols with Phusion enzyme (Thermo Scientific). Two *Bpi*I restriction sites with type I-G repeat sequence were introduced to pRAT-Duet MCS-I to get the spacer replaceable backbone of the pSPACER plasmid. 5'-phosphorylated oligos of CRISPR spacers were annealed and ligated into the *Bpi*I digested pSPACER backbone to obtain the pSPACER vector with target spacer.

2.3.1.2 HDR vectors

For experiments that required a DNA donor, we constructed the pHR vector by introducing homologous template into the pSPACER vector. Two 615 bp homologous arms (donor) for homologous directed repair (HDR) were PCR amplified from the *E. coli* MG1655 genome. The donor was incorporated into pSPACER MCS-II using restriction sites *Nde*I, *XhoI* and *XhoI AvrII*, followed by ligation. All final constructs were verified by sequencing (GATC Biotech, Eurofins Genomics, DE). Primers and

synthetic genes were obtained from Integrated DNA Technologies (Coralville, IA, USA), sequence can be found in Table 2 and Table 3.

2.3.2 Genome targeting by the type I-G CRISPR system

pM2 was transformed into *E. coli* MG1655. Transformants were selected using 100 µg ml^{-1} ampicillin. Competent cells were prepared by diluting an overnight culture 50-fold into fresh, selective LB medium. The culture was incubated at 37 °C, 220 rpm to reach OD_{600} 0.4 to 0.5. Cells were collected by centrifugation and the pellet resuspended in an equal volume of 100 mM $CaCl_2$, 40 mM $MgSO_4$. Following incubation on ice for 30 min, cells were collected and resuspended in 0.1 volumes of the same buffer containing 10 % glycerol. Aliquots were stored at -80 °C. 60 ng pSPACER or pHR was transformed into 60 µl competent cells. 400 µl LB medium was added after heat shock and cells incubated at 37 °C for 80 min. 100 µl aliquots of cells were applied onto 10 cm petri dishes in a 10-fold serial dilution for colony number counting and the number was corrected for dilution and volume to obtain colony-forming units (cfu) $0.1ml^{-1}$. The LB agar plates contained 100 µg ml^{-1} ampicillin, 12.5 µg ml^{-1} tetracycline, 1 mM IPTG, 0.2 mg/ml X-gal and 0.2 % (w/v) L-arabinose for induced plates.

2.3.3 Tiling PCR

Transformants were submitted for colony PCR with sets of primers (**table 3**). 10 µl MyTaq™ Red Mix (Bioline, Meridian bioscience) was used with 2 µl 20 µM primer mix,

colonies were added into the reaction. PCR products were analysed by separation on a 0.8 % agarose gel, running in 1 X TBE buffer.

2.3.4 Assays for Cas3 mutants

Cas3g mutated complexes were obtained using the same method in section 2.1.8. Plasmid challenge assay and phage immunity assay were described in section 2.1.11 and 2.1.12 respectively.

E. coli strains and plasmid used in this thesis can be found in Appendix Table 4 and Table 5.

Chapter 3: *In vitro* and in vivo construction of type I-G system

This chapter is adapted in part from the published manuscript: Structure and mechanism of the type I-G CRISPR effector[230].

3.1 Introduction

The type I-G gene locus in *Thioalkalivibrio sulfidiphilus* was investigated for elucidating type I-G system (Figure 3.1). This locus contains three canonical sections of CRISPR system, the CRISPR array, the adaptation *cas* genes and the interference *cas* genes. It also includes non-canonical genes (two toxin-antitoxin systems and an xpf-like gene) between *cas4-1* and *csb2*. The toxin-antitoxin system might be co-regulated with type I-G CRISPR as an immune defence. The xpf-like gene is uncharacterised but most likely encodes a DNA endonuclease involved in prespacer processing for adaptation. However, type I-G adaptation has been elucidated in another organism (*Geobacter sulfurreducens*), showing the fusion protein of Cas4 and Cas1 (Cas4-1) is crucial for spacer acquisition and the mutation of cas4 domain significantly decreases the rate of gaining spacer[231].

We have discussed how the Cas1-Cas2 complex acquires a spacer and incorporates it into CRISPR array. Cas4 protein has been found in several CRISPR systems related to the adaptation complex. In type I-A and I-D, Cas4 is not indispensable for spacer acquisition, but it involves functional PAM selection and prespacer processing[99, 232]. Another study in type I-C from *Bacillus halodurans* observes strong interaction between Cas4 and Cas1, showing Cas4 is associated with Cas1-Cas2 complex in EM structure[233]. Cas4 and Cas1 are fused in a single gene *cas*4-1 in type I-G system. The

Cas4 domain is crucial for PAM selection (TTN) and unfused Cas4 leads to a decrease in acquisition frequency[231], suggesting an evolutionary route of CRISPR adaptation.

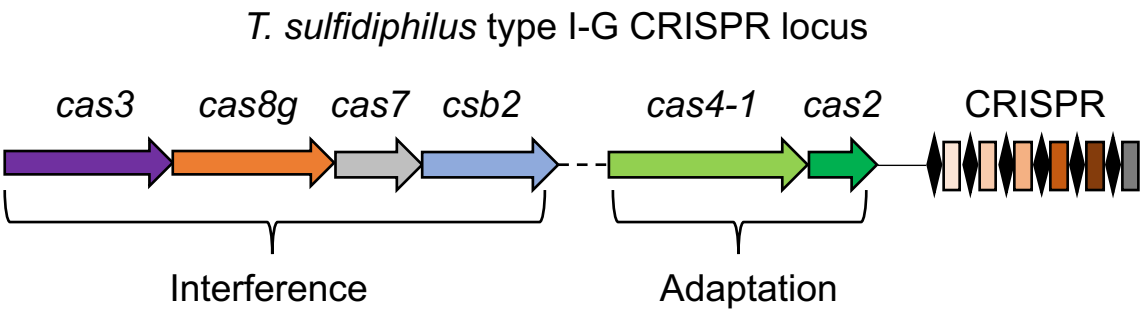


Figure 3.1 Type I-G gene locus of *Thioalkalivibrio sulfidiphilus*

There are three canonical sections: CRISPR array, adaptation, and interference. Between csb2 and cas4-1, there are 5 genes, including two toxin-antitoxin systems and an xpf-like gene.

csb2, the signature gene of type I-G, is in the interference section. It is taken as a fusion gene of cas5 and cas6. We will start from here to reveal this uncharacterized system.

3.2 Result

3.2.1 crRNA maturation in type I-G

3.2.1.1 Expression and purification of *cas* proteins

To study type I-G interference, we first investigated crRNA maturation, the basis of type I-G effector complex formation. To dissect this process, synthetic *cas* genes were first cloned into the pEV5HisTEV vector for protein expression. The vector contains a TEV protease cleavable His-tag, allowing further removal of His-tag after His-pulldown. Individual *cas* gene in the expression vector was then transformed into *E. coli* protein expression strain (C43 or Bl21 star) for protein expression. Through Ni-NTA affinity chromatography, Cas protein was isolated and pooled for TEV protease cleavage which removed the His-tag. The TEV cleaved protein was then submitted for size exclusion chromatography (Details in Figure 2.1). Eluted protein was monitored by SDS-PAGE. All three Cas proteins (Cas8g, Csb2 and Cas7) were successfully expressed and generated a defined peak on size exclusion chromatography. SDS-PAGE gel showed the purified protein from the eluted peak (Figure 3.2).

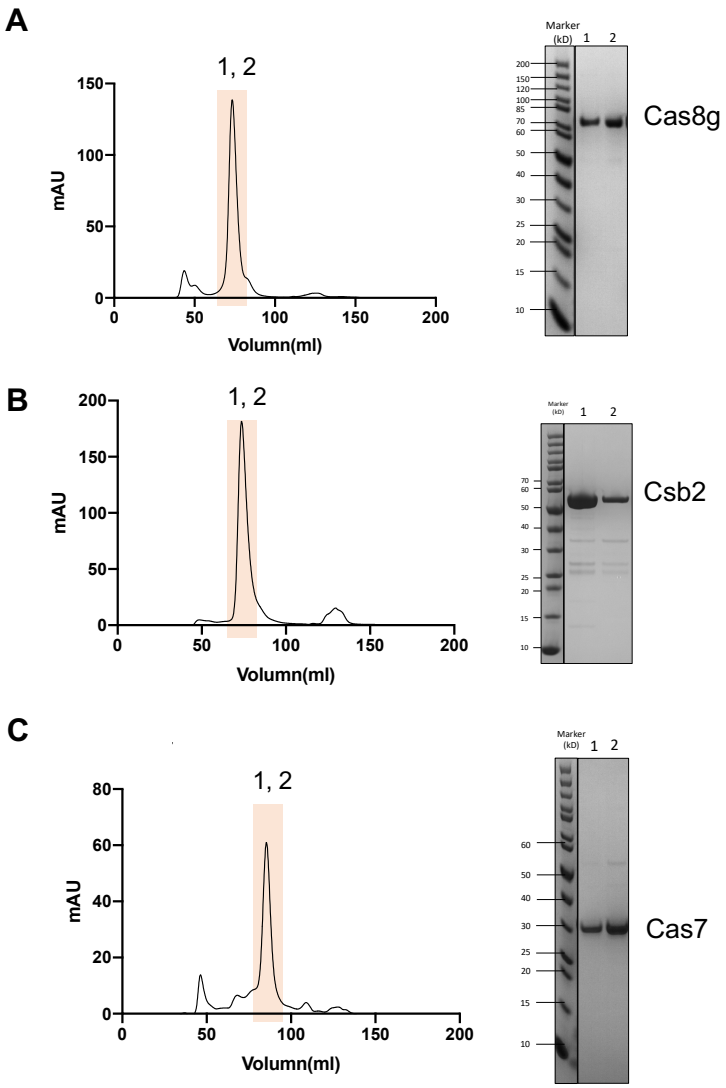


Figure 3.2. *cas* protein purification.

(A) (B) (C) Chromatograms of Cas8g, Csb2 and Cas7 purification by gel filtration chromatography, purification monitored by SDS-PAGE; lane 1-2, corresponding to the peak shown in the chromatogram.

3.2.1.2 Csb2 cleaves pre-crRNA into mature crRNA

The canonical crRNA maturation is carried out by Cas6 in type I systems, except type I-C where no Cas6 is present and Cas5 executes crRNA maturation. Csb2 is annotated as a fusion of Cas5 and Cas6. It is highly likely that Csb2 cleaves pre-

crRNA into mature crRNA in type I-G, however, the large subunit Cas8g is also a mysterious protein that might contribute to this process. Accordingly, we performed crRNA cleavage assay with all the Cas proteins to elucidate which one is the key to the crRNA maturation.

We first incubated Csb2, Cas7 and Cas8g individually with a 36 nt crRNA repeat. The crRNA repeat sequence was originally obtained from NCBI (NC_011901.1, *T. sulfidiphilus* genome). But we noted a directional error of the repeat sequence on the current database, hence, the crRNA repeat sequence was reverted for correction. With the corrected crRNA repeat, we observed a cleavage on crRNA repeat in Csb2 incubation and no cleavage in both Cas7 and Cas8g incubation (Figure 3.3A). The cleavage generated a 28 nt product, enabling us to map the cleave site on the crRNA. As shown in Figure 3.3C, the cleavage point was located 4 nt 3' of the base of the hairpin formed by pre-crRNA, an unusual feature as most Cas6 enzymes cleave at the base of hairpin[234]. But it still produces a canonical 8 nt handle at 5' end. To confirm the Csb2 activity and active site, we detected two conserved histidine (H331 and H503) in Csb2 protein sequence by multiple sequence alignment. Mutated variants of Csb2 were then purified as for wild-type Csb2 and incubated with crRNA repeat. The H331A variant showed no significant difference from wild-type Csb2, while H503A variant completely aborted the activity of crRNA cleavage (Figure 3.3B).

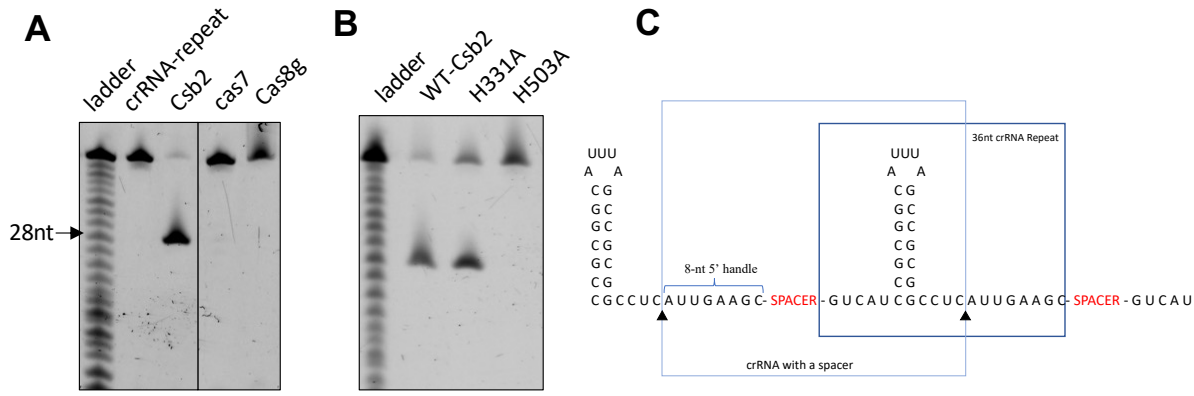


Figure 3.3. Csb2 generates mature crRNA in type I-G systems.

(A) A synthetic RNA species corresponding to the 36 nt CRISPR repeat (50 nM) was incubated with 0.5 μM purified recombinant Csb2, Cas7 and Cas8g at 37 °C for 10 min and analysed by denaturing gel electrophoresis. Only Csb2 generated a cleavage product. The ladder was obtained by alkaline hydrolysis of the crRNA. **(B)** The H503A variant of Csb2 does not cleave the repeat. **(C)** Schematic showing the pre-crRNA with Csb2 cleavage sites indicated.

To investigate pre-crRNA cleavage in more detail, we generated a 643 nt radioactively labelled RNA by *in vitro* transcription. The pre-crRNA comprised of 7 repeats and 6 spacers and is theoretically cleaved by Csb2 into a 72 nt mature crRNA (Figure 3.4A). Wild-type Csb2 and variants were submitted to process the pre-crRNA, a series of products were observed on the denaturing gel with a major product corresponding to the expected size of mature crRNA. A high concentration of Csb2 is required for producing the final product; intermediate products accumulated when Csb2 was at a low concentration. The H503A variant no longer processed the pre-crRNA (Figure 3.4B), which is consistent with the previous crRNA repeat cleavage assay. The product of long pre-crRNA cleavage is less pure than that of the single repeat cleavage

(Figure 3.3), suggesting a more complicated structure of RNA might hinder the crRNA maturation.

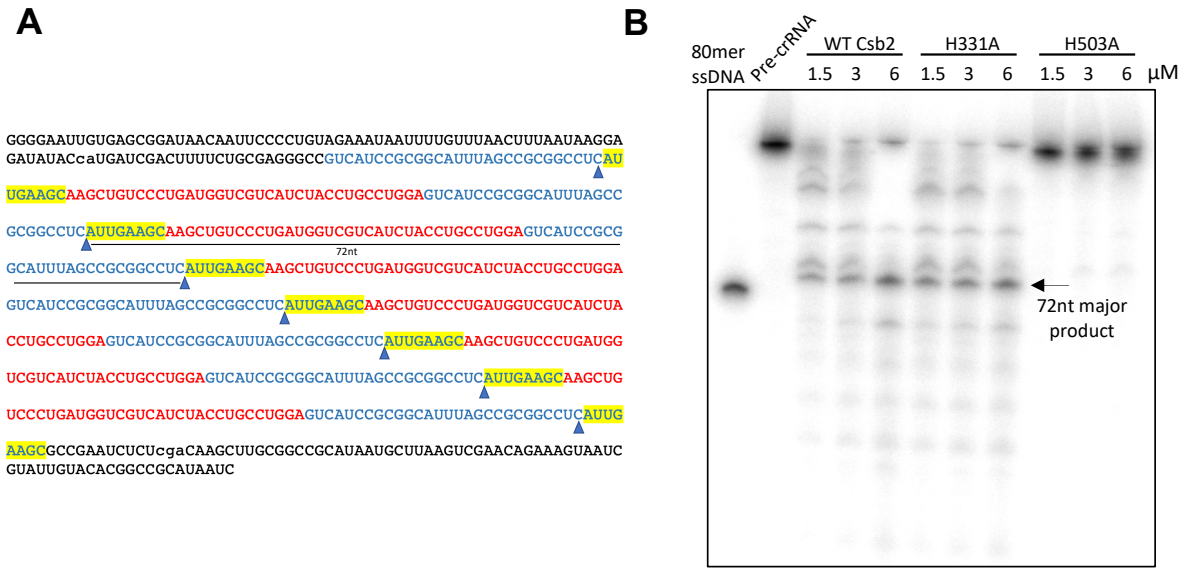


Figure 3.4. pre-crRNA cleavage by Csb2

(A) The sequence of this 643 nt pre-crRNA; Repeat sequence in blue, spacer sequence in red, 8 nt 5' handle highlighted in yellow, arrow indicates the cleavage site. **(B)** 643nt pre-crRNA (1 μM) was *in vitro* transcribed with [α-^{32}P]-ATP and cleaved by WT-Csb2 or variants of Csb2 and products were analysed by denaturing gel electrophoresis.

3.2.2 Csb2, the fusion of Cas5 and Cas6

3.2.2.1 Structural analysis of Csb2

Csb2 cleaves pre-crRNA into mature crRNA. It provides evidence for the assumption that Csb2 is a fusion of Cas5 and Cas6. But how the Csb2 interacts with the crRNA remains unclear. To further investigate it, we used Alphafold to model Csb2[227]. This model shows two distinct domains of Csb2 connected by a linker (Figure 3.5). Although the orientation of the two domains cannot be predicted, the model does

support the fusion theory with histidine 503 of Csb2 positioned in the C-terminal domain. The model was then submitted to DALI[228] for structural homologues search. The first 150 residues of the N-terminal domain match to Cas5 subunit of type I and type III systems from the DALI search, giving a reasonable Z-scores in range 5 to 10. And the C-terminal domain of Csb2 has a strong match (Z-score 12.4) to the Cas6b protein from *Methanococcus maripaludis.*

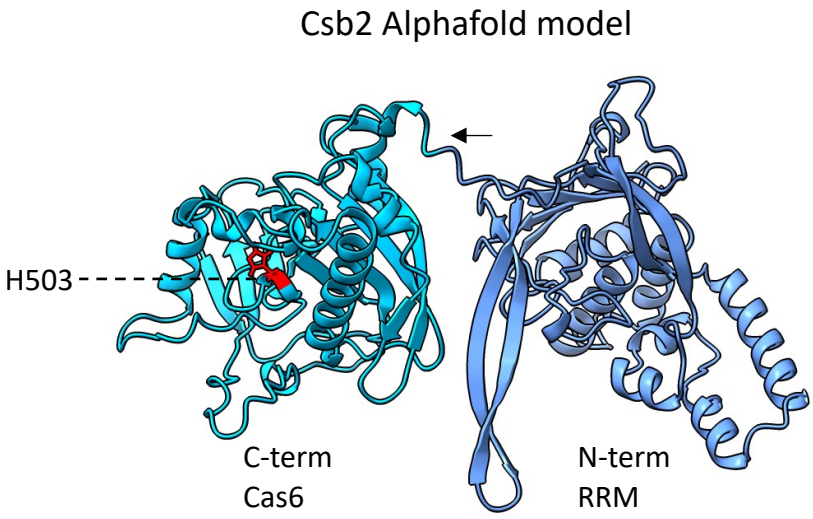


Figure 3.5. Csb2 Alphafold model structure

The sequence of Structural model of the Csb2 protein by Alphafold suggests a two-domain structure with a C-terminal Cas6-like domain, joined by a linker (Indicated by arrow) sequence. The relative orientation of the two domains cannot be predicted. RRM, RNA recognition motif.

By structural comparison between the N-terminal domain of Csb2 and Cas5 from *E. coli*, the N-terminal domain does hold common structural features of Cas5, an RRM (RNA recognition motif) fold and extended beta-hairpin. But at the end of N-terminal domain, the structure diverts with a more extended alpha-helix (Figure 3.6A). For C-terminal domain, the comparison with Cas6b supports the strong match, and the

binding site of crRNA could be estimated by the overlay of the structures, which is consistent with the Csb2 catalytic site histidine 503 (Figure 3.6B).

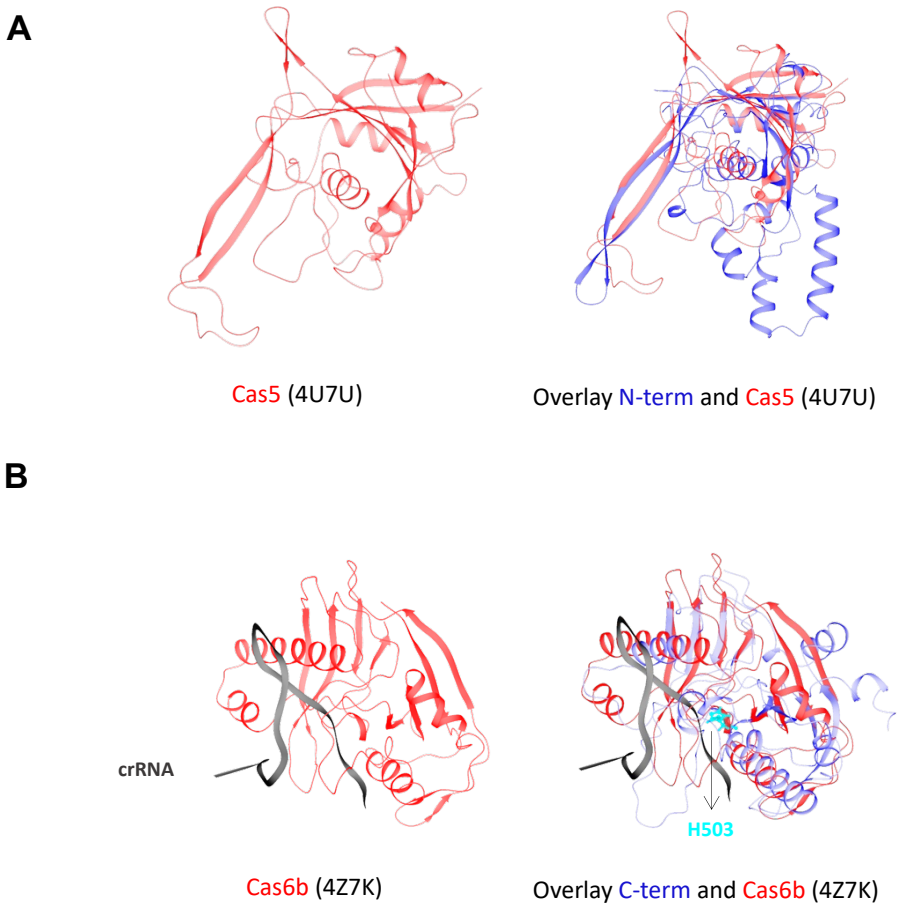


Figure 3.6. Alignment of Csb2 N-term and C-term

(A) Structural comparison of the modelled Csb2 N-terminal domain (blue) with the Cas5 protein from *Escherichia coli* (red), PDB: 4U7U. **(B)** Structural comparison of the modelled Csb2 C-terminal domain (blue) with the Cas6b protein from *Methanococcus maripaludis* (red), PDB: 4Z7K, reveals the likely site of crRNA hairpin binding (black, from the Cas6b structure) adjacent to the position of the H503 (cyan) active site residue.

3.2.2.2 Csb2 binding affinity with crRNA repeat

The structure model and the overlay provide us with a working model where Csb2 is associated with the 3' hairpin of crRNA. It is a canonical model in type I systems that Cas6 cleaves pre-crRNA and remains associated with the 3' hairpin, however, Cas5 is found associated with 5' handle after crRNA maturation in type I-C[195]. This brought up uncertainty since Csb2 is the fusion of Cas5 and Cas6, two distinct domains linked together. To resolve this uncertainty and figure out where Csb2 is associated after cleavage, we processed to investigate the binding affinity of Csb2 for crRNA 3' hairpin or 5' handle with fluorescence anisotropy. Csb2 was titrated at progressively higher concentrations into fluorescein-labelled (FAM) crRNA repeat, and a high binding affinity (K_D=55 nM) was observed (Figure 3.7A). The fluorescein-labelled crRNA 3' hairpin or 5' handle was submitted to the same anisotropy assay. Csb2 showed a higher binding affinity (K_D=43 nM) for crRNA 3' hairpin (Figure 3.7B), while no binding affinity was observed for 5'-8 nt handle (Figure 3.7C). To exclude the possibility that the position of fluorescein-label blocked the binding, the label was moved to 3' end, and no binding was detected (Figure 3.7D). This indicates that Csb2 remains associated with 3' hairpin after cleavage, a canonical Cas6 feature. But as we showed in the Csb2 structure comparison, Csb2 does possess a Cas5-like N-terminal domain. What is the functional role of the Cas5-like domain? Is it indispensable in crRNA maturation? We dissected Csb2 further to resolve these questions.

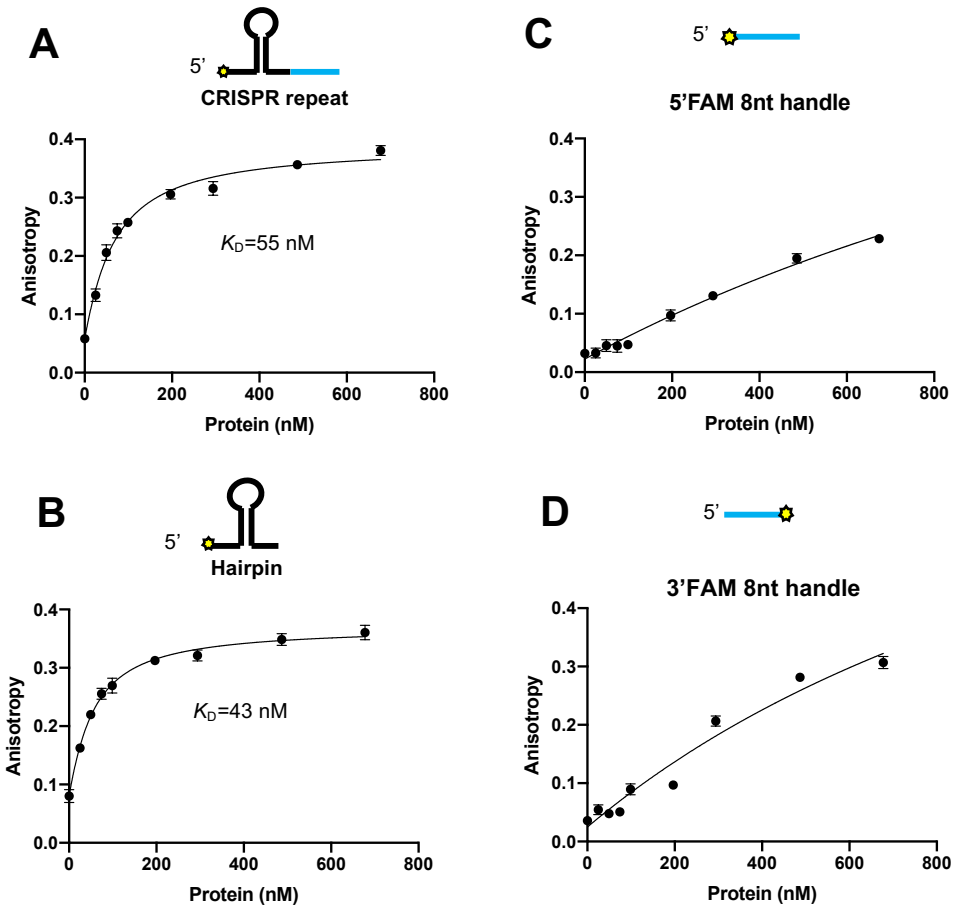


Figure 3.7. Csb2 binding affinity with CRISPR repeat, 3' hairpin or 5'- 8 nt-handle.

(A) (B) Fluorescence anisotropy analysis of Csb2 binding to the CRISPR repeat and hairpin, the dissociation constant (K_D) is 55 nM and 43 nM respectively. **(C) (D)** Fluorescence anisotropy analysis of Csb2 binding to the 8nt handle; The FAM-label is either 5' end or 3' end; The position of the FAM label is indicated with a yellow star. Data points and error bars represent the mean of five technical replicates and standard deviation.

3.2.2.3 The two domains of Csb2

The Csb2 structure model shows two distinct domains, a Cas5-like N-terminal domain and a Cas6-like C-terminal domain, joined by a linker. Based on the model, we

separated the Csb2 domain at R260 in the linker. The N- and C-terminal domain were cloned and purified successfully, suggesting that the stability of folding each domain is independent. The purified N- and C-terminal protein were submitted to crRNA repeats cleavage. The Cas6-like C-terminal domain demonstrated similar cleavage activity to the intact Csb2, while the Cas5-like N-terminal cannot cleave the crRNA repeat at all (Figure 3.8A). The binding affinity assay was carried out with those two domains. For intact crRNA repeat and 3' hairpin, C-terminal domain shows a high binding affinity (K_D=8 nM and K_D=3 nM for crRNA repeat and 3' hairpin respectively), even higher the intact Csb2. However, the N-terminal domain has no binding with crRNA (Figure 3.8B&C). Both N- and C-terminal domain are not associated with 5' -8 nt handle (Figure 3.8D). These results confirm that Csb2 binds with crRNA 3' hairpin after crRNA maturation through its Cas6-like C-terminal domain and suggest that Cas5-like N-terminal domain is not necessary for crRNA maturation. The functional role of N-terminal domain is still unclear, and a reasonable guess is that it might be involved in the effector complex construction, we will investigate it in the reconstruction of type I-G effector complex.

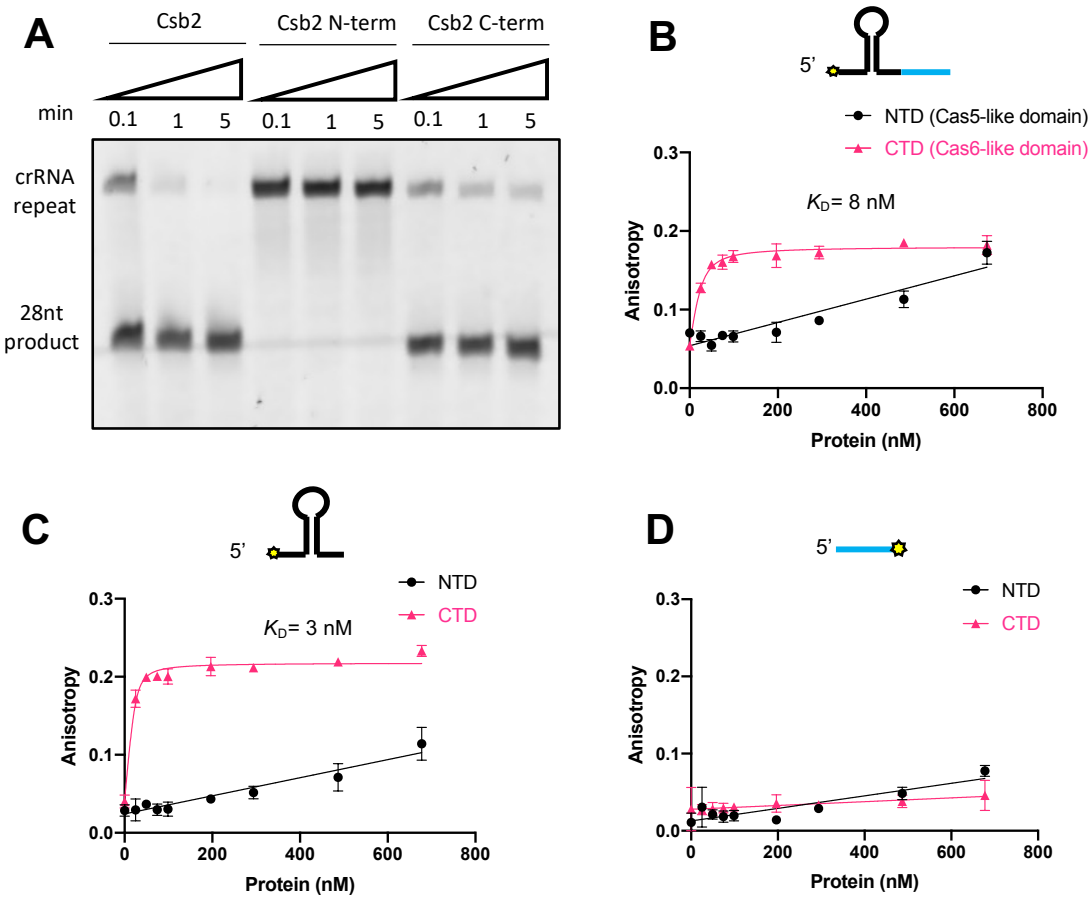


Figure 3.8. Binding affinity of the two domains of Csb2

(A) crRNA repeat cleavage with N- and C- terminal domain of Csb2. **(B) (C)** Fluorescence anisotropy analysis of N- and C- terminal domain of Csb2 binding to the CRISPR repeat and hairpin, the dissociation constant (K_D) of C-terminal domain binding is 8 nM and 3 nM respectively. **(D)** Fluorescence anisotropy analysis of N- and C- terminal domain of Csb2 binding to the 8nt handle. The position of the FAM label is indicated with a yellow star. Data points and error bars represent the mean of five technical replicates and standard deviation.

3.2.3 *In vitro* reconstruction of type I-G effector complex

3.2.3.1 The formation of type I-G effector complex

Type I systems expressed an effector complex to target dsDNA. The complex is first termed as Cascade (CRISPR-associated complex for antiviral defence) when type I-E CRISPR was investigated. Upon targeting dsDNA, Cascade recruits Cas3 to degrade target dsDNA. To investigate the type I-G system, we sought to reconstruct the Cascade of type I-G *in vitro*. Since Cas proteins were expressed and purified separately, we proceeded to incubate Cas proteins with *in vitro* transcribed pre-crRNA to form the effector complex *in vitro*. Csb2 and Cas7 were first incubated with pre-crRNA. Theoretically, Csb2 performs pre-crRNA cleavage, generating mature crRNA, and Cas7 binds to the mature crRNA, forming the backbone of the effector complex. After incubation, the Cas protein and crRNA mix were submitted for size exclusion chromatography. As shown in Figure 3.9A, an early eluted peak was observed compared to the no pre-crRNA control, suggesting a large molecular weight complex eluted. The eluted protein from the early peak was then run on SDS-PAGE to identify the composition. Csb2 and Cas7 were both detected in this early eluted fraction, showing that the Csb2, Cas7 and crRNA form the backbone of the effector complex by *in vitro* incubation. We then introduced Cas8g, the large subunit of the effector complex, into the incubation. Cascade was obtained by the same procedure (Figure 3.9B). We also attempted to introduce Cas3 into the incubation, surprisingly, Cas3 was incorporated into Cascade without target dsDNA present (Figure 3.9C), suggesting Cas3 in type I-G is a stable component of the effector complex instead of being recruited to the Cascade upon target binding.

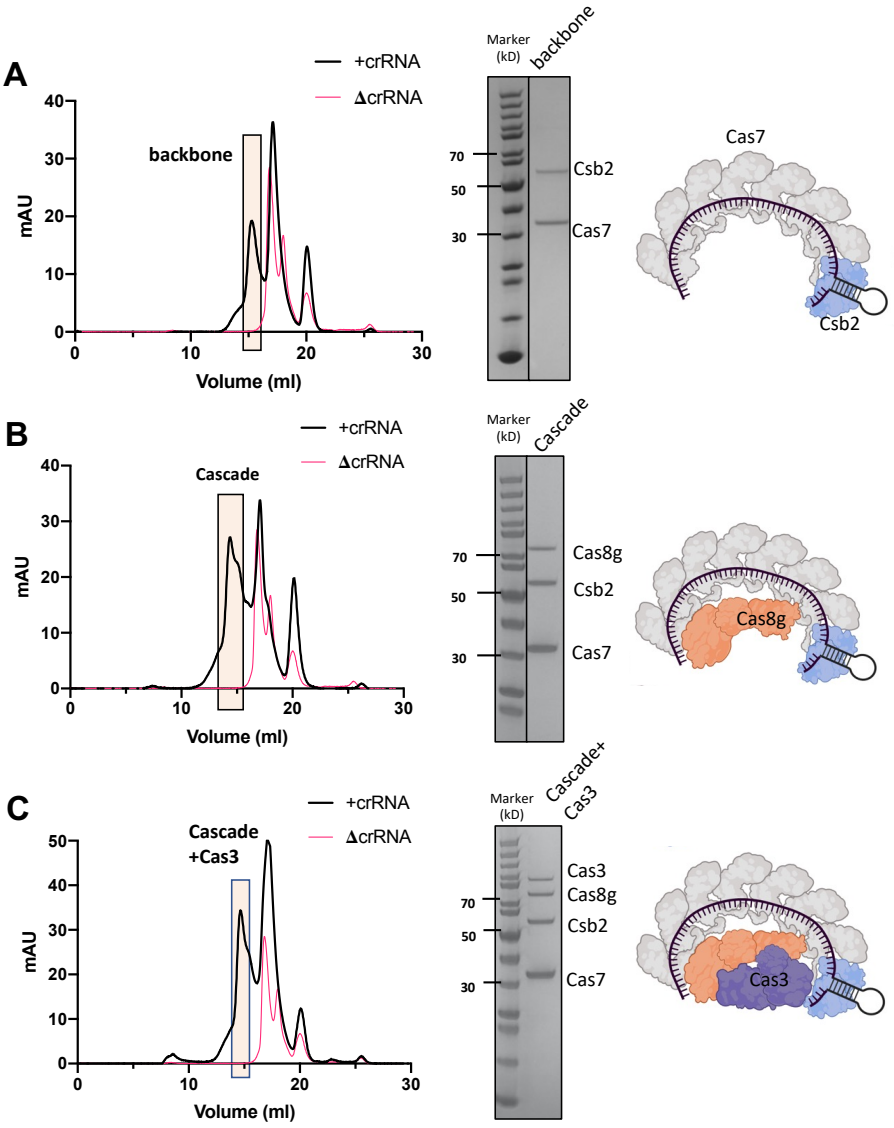


Figure 3.9. *In vitro* reconstruction of the type I-G complex.

Recombinant protein subunits were incubated with *in vitro* transcribed pre-crRNA and subjected to size exclusion chromatography. Each panel shows the resulting chromatography, SDS-PAGE analysis of the indicated fractions and schematic representation of the complex obtained (Schematic figures were made based on the structure data which will be discussed in next chapter). **(A)** Cas7 and Csb2 form a defined complex with crRNA. **(B)** Cas8g forms a stable Cascade complex with Cas7/Csb2/crRNA. **(C)** Cas3 forms a stable complex with the type I-G Cascade. In all cases, complex formation was dependent on the presence of crRNA.

When Cas8g is absent, Cas3 no long binds to the backbone (Figure 3.10), suggesting that Cas3 is associated to the backbone by the interaction with larger subunit Cas8g.

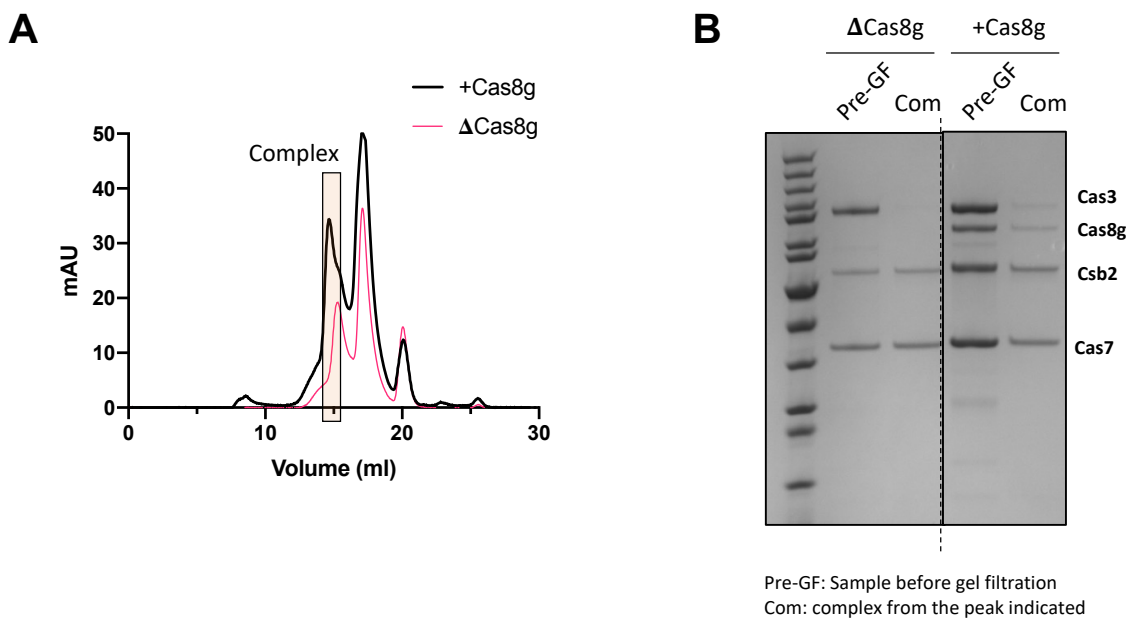


Figure 3.10. Cas3 interacts with Cas8g to be incorporated into the Cascade.

(A) Chromatography showing complex formation in the presence of Cas8g or in the absence of Cas8g. The rectangle indicated the fraction of the complex. **(B)** The complex from chromatography was submitted to SDS-PAGE.

We previously brought up one unsolved question concerning the functional role of Csb2 N-terminal domain. To address it, we performed the effector complex incubation with N- and C-terminal domain of Csb2. Neither N-terminal domain nor C-terminal domain alone generates the effector complex with Cas protein and crRNA (Figure 3.11). Intact Csb2 is required to form the complex, suggesting the Cas5-like N-terminal domain of Csb2 plays an important role in effector complex formation.

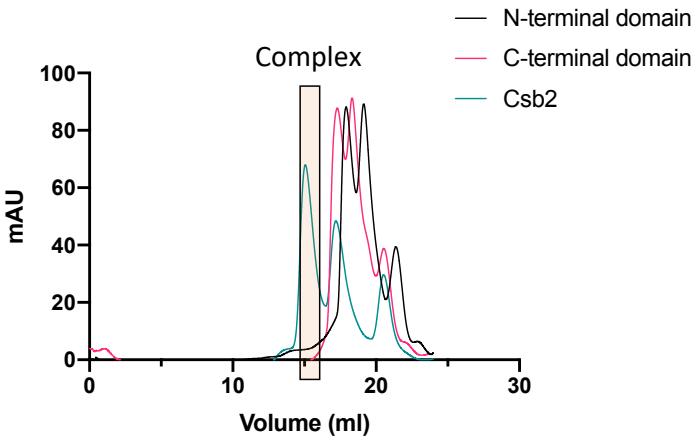


Figure 3.11. Two domains of Csb2 are required for complex formation.

Chromatography showing only completed csb2 formed the effector complex, neither N-terminal domain nor C-terminal domain alone formed the complex.

3.2.3.2 Cas3 pre-associated with Cascade is essential for DNA targeting

Once we obtained the effector complex of type I-G system, we proceeded to check the activity of type I-G RNP *in vitro*. We first focused on the effector complex targeting. The pre-crRNA for complex formation is a 643 nt RNA with spacers targeting Tetracycline resistance gene. The RNP with mature crRNA can target dsDNA with corresponding DNA sequence. Cascade was submitted to a plasmid binding assay with the target plasmid (pRAT) or the non-target control (pCDF) to investigate its binding with supercoiled (SC) dsDNA. The reaction was carried out in the absence of ATP, and the products were analysed on agarose gel. We observed a non-specific dsDNA nick in both target and non-target plasmid at high concentration of the complex, which might be attributed to nuclease contamination during the purification process. No binding was observed between Cascade and target plasmid. We then performed the same assay with Cascade and Cas3 complex. At the presence of Cas3, a

complete loss of free SC dsDNA was observed on the target plasmid, which was gel-shifted by the effector complex and no shift in the non-target plasmid assay, indicating the binding between the RNP and target SC dsDNA (Figure 3.12A). The plasmid binding assay was adopted from Westra, Edze R[235]. The binding of effector complex and supercoiled dsDNA is observed as a band shift on gel.

We strengthened the observation by introducing 80bp dsDNA for RNP targeting. The 80bp target dsDNA was either target strand or non-target strand labelled, following the electrophoretic mobility shift assay (EMSA) with type I-G effector complex, a faint shift was detected where Cascade (backbone and Cas8g) was present alone. But when Cas3 was incorporated, nearly all free dsDNA was gel-shifted by the effector complex (Figure 3.12B). Cas3 alone cannot bind to target dsDNA (Figure 3.12C&D). The observation in short dsDNA is consistent with that in plasmid assay where Cas3 pre-associated with Cascade is essential for dsDNA targeting.

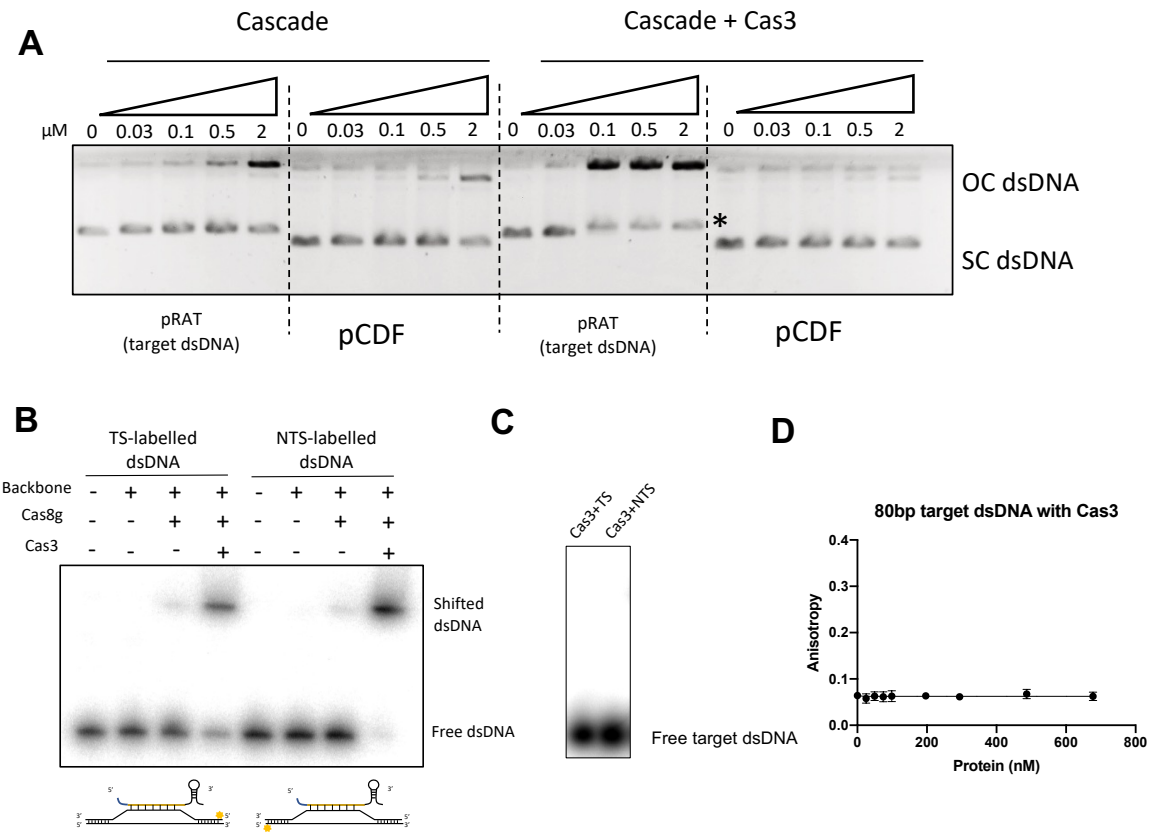


Figure 3.12. dsDNA targeting by effector complex.

(A) Target dsDNA pRAT and non-target dsDNA pCDF were incubated with type I-G complex Cascade or Cascade plus Cas3, following an overnight Agarose gel electrophoresis. OC (open circular) dsDNA, SC (supercoiled) dsDNA. * Shifted dsDNA. **(B)** Electrophoretic Mobility Shift Assay (EMSA) shows that type I-G Cascade only forms a stable complex with linear dsDNA targets in the presence of Cas3, 10 nM dsDNA was mixed with 0.8 μM effector complex. **(C)** Electrophoretic Mobility Shift Assay (EMSA) shows that Cas3 alone has no binding with dsDNA; TS, target strand labelled dsDNA; NTS, non-target strand labelled dsDNA. **(D)** Anisotropy showing Cas3 binding affinity with target dsDNA. Cas3 alone has no binding with target dsDNA. Data points and error bars represent the mean of five technical replicates and standard deviation.

3.2.3.3 Target dsDNA cleaved by type I-G effector complex

After targeting, type I-G effector complex is capable of cleaving target dsDNA. To test the cleavage ability of type I-G effector complex *in vitro*, we performed the same plasmid assay, but in the cleavage study, ATP was added into the reaction. Upon binding to type I-G effector complex (Cascade and Cas3), target free supercoiled dsDNA was gel-shifted, which had been observed in the binding assay without ATP. In addition to the binding, a progressive increase in target dsDNA cleavage was seen, giving rise to the background staining. In contrast, no gel-shifted band and background staining was observed for non-target plasmid (Figure 3.13A). To further confirm the gel-shifted band is due to the binding between effector complex and free SC dsDNA, the products were denatured by heating, and the shifted band disappeared, free SC dsDNA was released on the gel while the cleavage product can still be observed (Figure 3.13B).

ATP activates the type I-G cleavage, the key protein to this process is the Cas3. Cas3 is an ATP dependent helicase and ATP-independent nuclease. To confirm the cleavage of target dsDNA is generated by Cas3, two variants of Cas3 helicase domain mutation, K39A and D164A (Walker motif mutation, more details in Chapter 5), were expressed and purified. The helicase activity of K39A and D164A was abolished and no longer unwinds long dsDNA into ssDNA for nuclease domain cleavage. When the Cas3 variants were incorporated into Cascade, the effector complex could still bind to target SC dsDNA, but no smear of degradation was generated. More open circle (OC) dsDNA was seen, consistent with the ATP absent condition (Figure 3.13C).

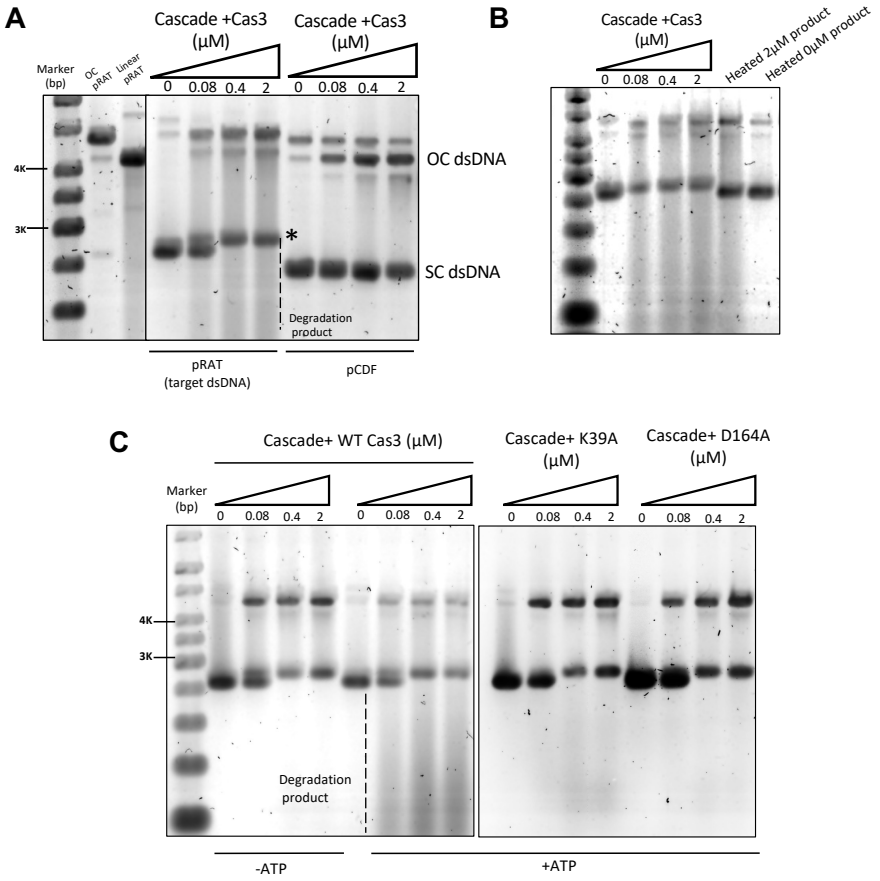


Figure 3.13. Supercoiled dsDNA targeting and degradation by effector complex.

(A) Supercoiled (SC) dsDNA plasmid cleavage and binding assay. Target plasmid pRAT and non-target plasmid pCDF were incubated with type I-G complex Cascade and Cas3 with ATP and analysed by gel electrophoresis. The target plasmid SC species was gel-shifted by Cascade (*), nicked to open circle (OC) form, and degraded to generate a background smear of DNA fragments. The non-target plasmid was partly nicked, but not bound or digested. pRAT obtained by nicking endonuclease digestion; linear pRAT gained by single restriction enzyme digestion. *Target SC dsDNA was bound by the effector complex. **(B)** Agarose gel electrophoresis shows that shifted band disappeared after product denaturing. **(C)** Target SC plasmid was incubated with Cascade and Cas3 in the absence (-) or presence (+) of ATP. ATP was required for the generation of the smear of degraded DNA species. Cas3 helicase domain mutant, Cas3 K39A and D164A aborted the target plasmid degradation.

We next checked the target dsDNA cleavage in more detail using the 80 bp short target dsDNA. The non-target strand (NTS) of target dsDNA was radioactively labelled first, the labelled dsDNA was then submitted for effector complex targeting. Products from the reaction were heat-denatured and loaded to a polyacrylamide-TBE gel. As expected, the cleavage was dependent on the presence of Cascade and Cas3. We generated a G+A ladder for mapping the cleavage site. In the absence of ATP, the initial cleavage site on the non-target strand is near the centre of R-loop (A8 position), defining the start position of Cas3 nuclease. While ATP was present, unwinding dsDNA, non-target strand was further cleaved by Cas3 nuclease, more cleavage sites were revealed away from the start point A8 in a 5' direction (Figure 3.14A&C). The target strand was not cleaved by type I-G effector complex *in vitro* (Figure 3.14B).

Overall, the type I-G effector complex was successfully reconstituted *in vitro*. Target dsDNA is specifically cleaved by type I-G effector complex.

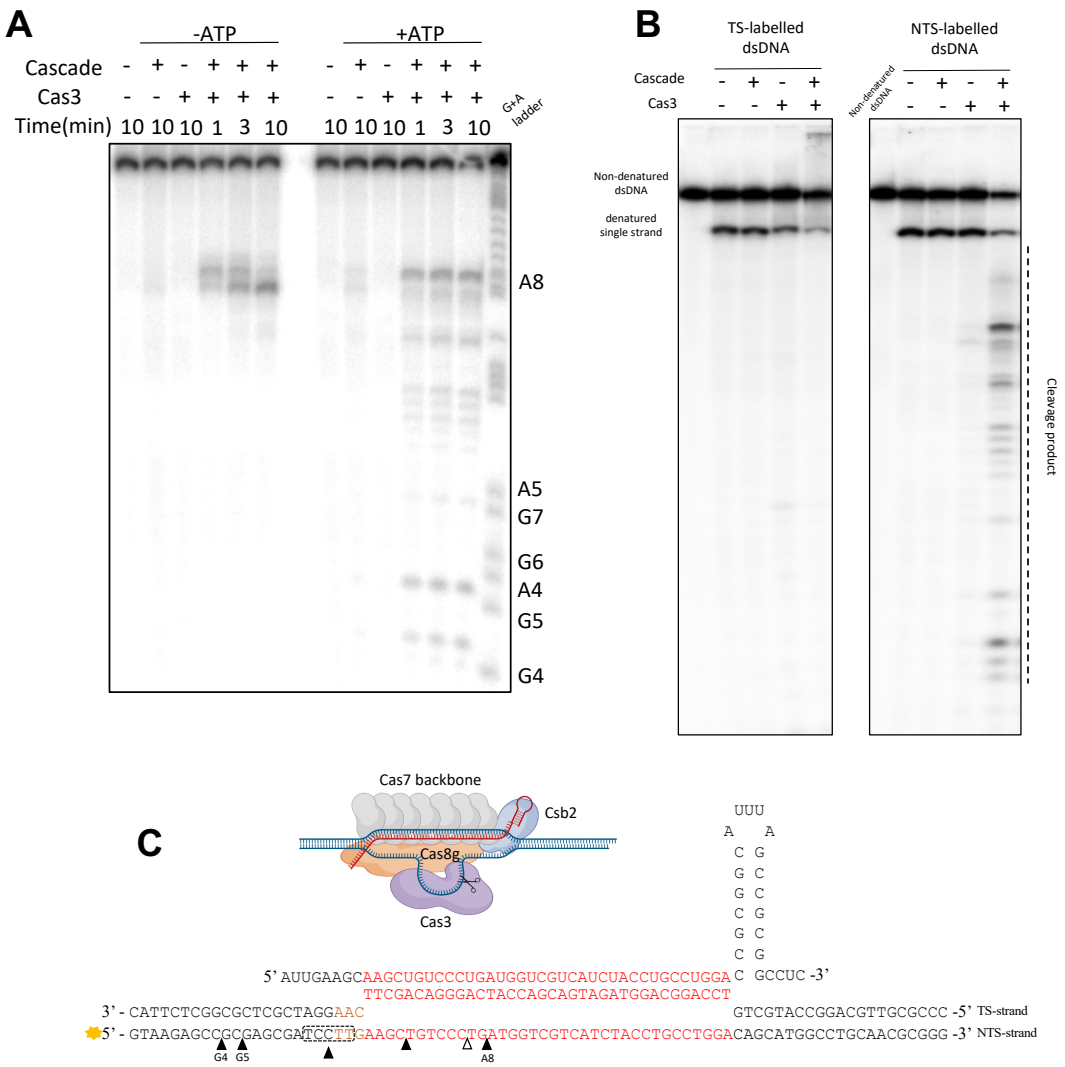


Figure 3.14. Mapping dsDNA cleavage

(A) 16.6 nM dsDNA target was incubated with 0.3 μM Cascade ± Cas3 in the presence and absence of ATP and analysed by denaturing gel electrophoresis. Without ATP, Cas3 cleaved the NTS in the centre of the R-loop (position A8). In the presence of ATP, Cas3 cleaves the NTS at sites 5' of the R-loop, consistent with the 3'-5' polarity of Cas3. **(B)** Target labelled or NTS-labelled dsDNA was incubated with Cascade and Cas3 in the presence of ATP, products separated on a denaturing polyacrylamide-TBE gel. **(C)** Schematic of the Cascade-target DNA complex, and mapping of cleavage sites observed for D. Black triangles show the cleavage sites when ATP is present. The five nucleotides boxed by dash lines are all cleavage sites.

3.2.4 In vivo reconstruction of type I-G effector complex

3.2.4.1 Invasive plasmid eradicated by type I-G CRISPR *in vivo*

After the reconstitution of type I-G effector complex *in vitro*, we then attempted to build type I-G system *in vivo*. We first constructed three vectors containing components of type I-G effector complex. Backbone genes of type I-G (*csb2, cas7* and *cas8g*) were cloned into pACE (Ampicillin resistance, Ap^R) under *lac* operon control. CRISPR array, repeat and spacer, was cloned into pCDF (Spectinomycin resistance, Sp^R), also *lac* operon control. Cas3 was incorporated into pRAT (Tetracycline resistance, Tet^R), under araBAD promoter control. pACE and pCDF were co-transformed into *E. coli C43* strain to provide a stable expression of Cascade targeting Tet^R gene since the spacer in pCDF was designed for Tet^R targeting. pRAT was then transformed into the Cascade expression cell line. Transformed cells were subsequently spread on three types of plates with different selecting conditions: I. Recipients, ampicillin and spectinomycin selection, indicating the initial recipient cell numbers. II. Transformants, ampicillin, spectinomycin and tetracycline selection, showing transformation efficiency, cells without Tet^R cannot survive. III, Induced, all antibiotics, lactose, and arabinose for full type I-G system induction (Figure 3.15A).

If the type I-G system is functionally activated *in vivo*, the target plasmid pRAT would be eradicated from the cells, losing the tetracycline resistance, hence cells cannot survive on the plates that contain tetracycline. In contrast to **Δ**Cas3 strain where pRAT was transformed without *cas3* gene, fully established type I-G system in *E. coli* significantly decreased the cell number on the induced plate. Even without induction, the small amount of type I-G system expressed from promoter leakiness could clear

the invasive target plasmid pRAT, leading to cell death on transformants plates (Figure 3.15B).

To further confirm the type I-G activity and exclude the possibility that the toxicity of type I-G caused the cell loss, we constructed a new spacer targeting the *lacZ* gene and incorporated into pCDF, following same process as above, Cascade with *lacZ* targeting was built in *E. coli* DH5α strain. But in this case, the cell strain was challenged by either non-target pRAT-Duet plasmid or target pRAT-lacZ (*lacZ* gene cloned into pRAT-Duet) plasmid. Cas3 was present in both conditions. Target strain suffered a huge loss in cell number while non-target strain showed adequate transformants on the plates (Figure 3.15C), indicating type I-G expressed *in vivo* and specifically targeting invasive plasmid, causing the loss of tetracycline resistance.

In general, type I-G can be built in the heterologous organism *E. coli* and specifically target dsDNA to eradicate the invasive plasmid.

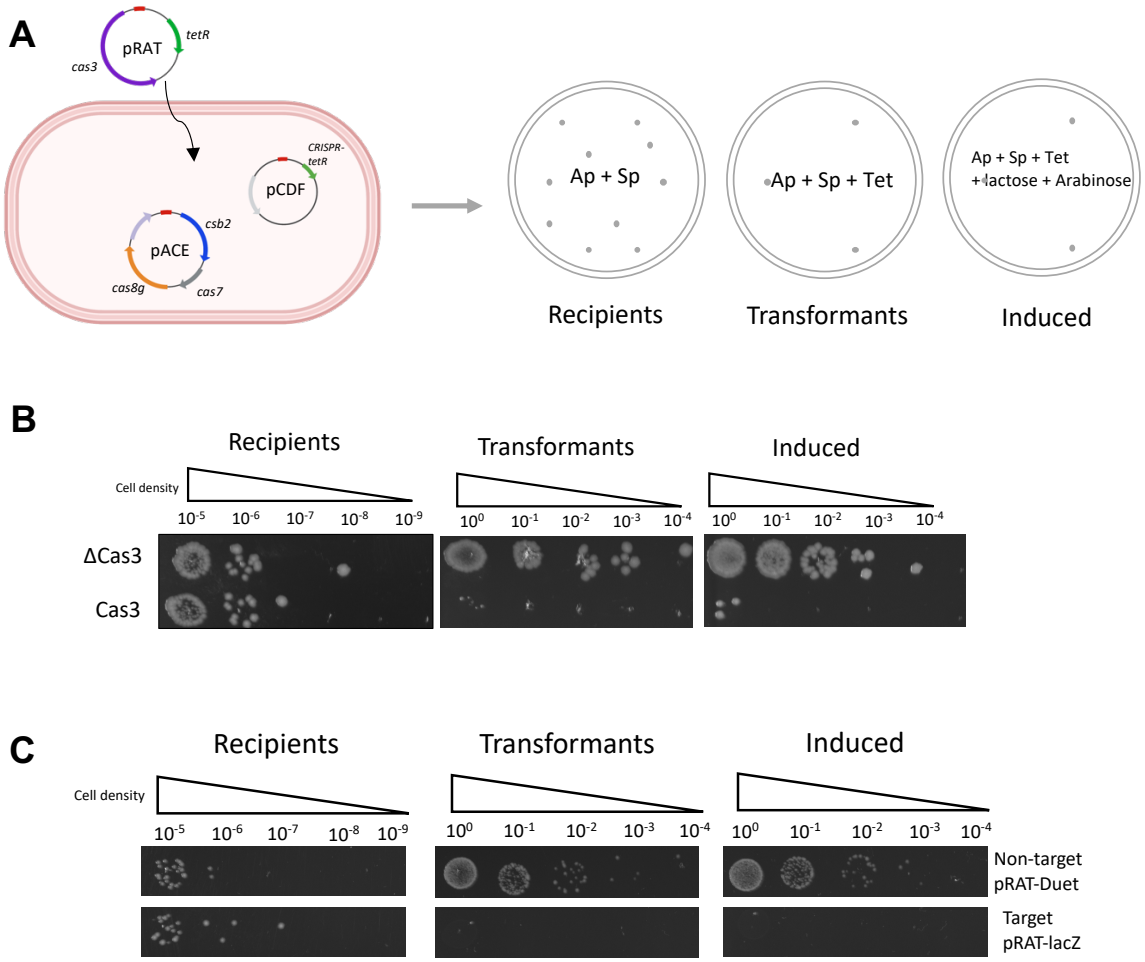


Figure 3.15. Plasmid challenge assay.

(A) A schematic diagram explaining the plasmid challenge assay. Competent cells harbouring type I-G system were challenged with target plasmid. Csb2, Cas7 and Cas8g gene were built in pACE vector (Ampicillin resistance); CRISPR array targeting the pRAT tetracycline resistance (TetR) gene was constructed in pCDF (Spectinomycin resistance); Cas3 gene in pRAT under arabinose promoter control. Recipients, Ampicillin and Spectinomycin in plates; Transformants, Ampicillin, Spectinomycin and tetracycline in plates; Induced, with all three antibiotics and lactose, arabinose for induction. **(B)** The cells on the plates in different condition. **Δ**Cas3, cells challenged with pRAT-Duet plasmid (Cas3 excluded). **(C)** Same plasmid challenge assay for target or non-target plasmid with Cas3 present. The cells on the plates in different condition.

3.2.4.2 Type I-G CRISPR protects cells from phage infection

Since *in vivo* reconstruction of type I-G eradicates invasive plasmid, we wondered if the type I-G could protect cells from phage infection. The phage immunity assay was conducted to answer this question.

We first co-transformed three aforementioned plasmids, pACE, pCDF and pRAT into *E. coli* C43 strain, and pCDF contains a new spacer that targets the temperate phage P1 late promoter activating (*lpa*) gene. The cells were then challenged by phage P1 (Figure 3.16A). Phage P1 infected the cells with three MOI (multiplicity of infection). At MOI 0.01 where phage infected cells with a low concentration, the growth curves showed that fully established type I-G system provides immunity against phage infection while the growth of **Δ**Cas3 strain was constrained by phage infection. At higher MOI (MOI=0.1 and MOI=1), type I-G consistently provides phage immunity and the cells lacking Cas3 suffers a nearly complete loss at early stage of infection. The cell growth recovers at late stage of infection, mainly due to phage P1 incorporated into host genome to start temperate infection (Figure 3.16B).

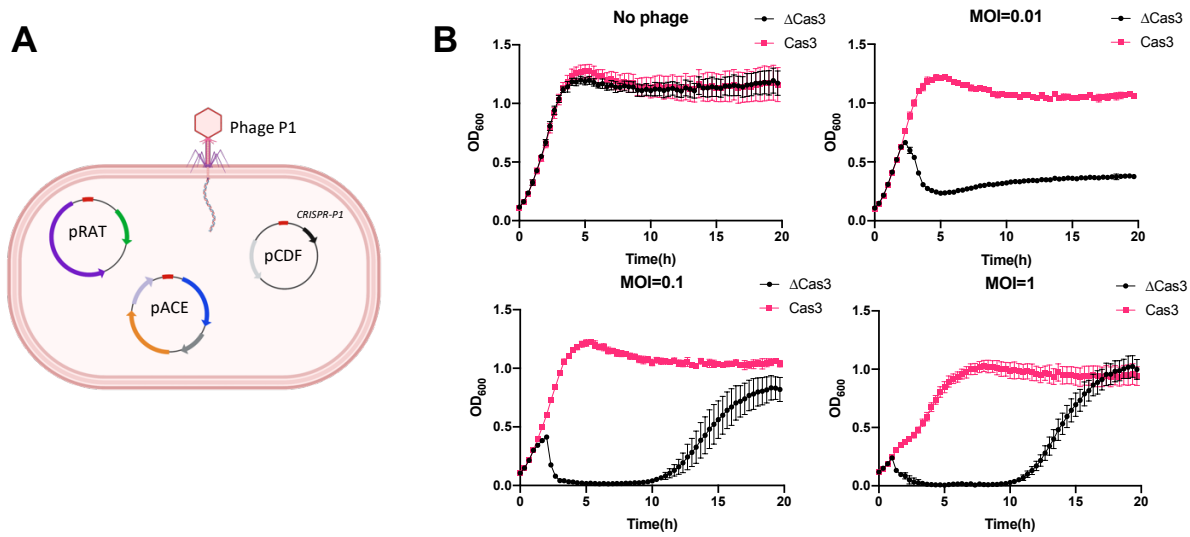


Figure 3.16. Phage challenge assay with induced type I-G system

(A) Phage immunity assay. Type I-G system established in *E. coil*, CRISPR array targeting phage P1 *lpa* gene. **(B)** Cell growth curve, no phage infection or phage infection (MOI=0.01, 0.1 and 1). **Δ**Cas3, cells lack of Cas3 gene. Data points represent the mean of six experimental replicates (two biological replicates and three technical replicates) with standard deviation shown.

Fully induced expression of type I-G in *E. coli* sufficiently protects cells from phage infection. We processed to test the type I-G interference in an uninduced condition where less type I-G was expressed. Similar to fully induced condition, at low MOI, type I-G system still provide sufficient interference to clear phage infection. But under high load of phage infection, it only partially secures cell growth (Figure 3.17).

Overall, type I-G is functional in the heterologous organism, providing clearance of invasive plasmid and phage immunity.

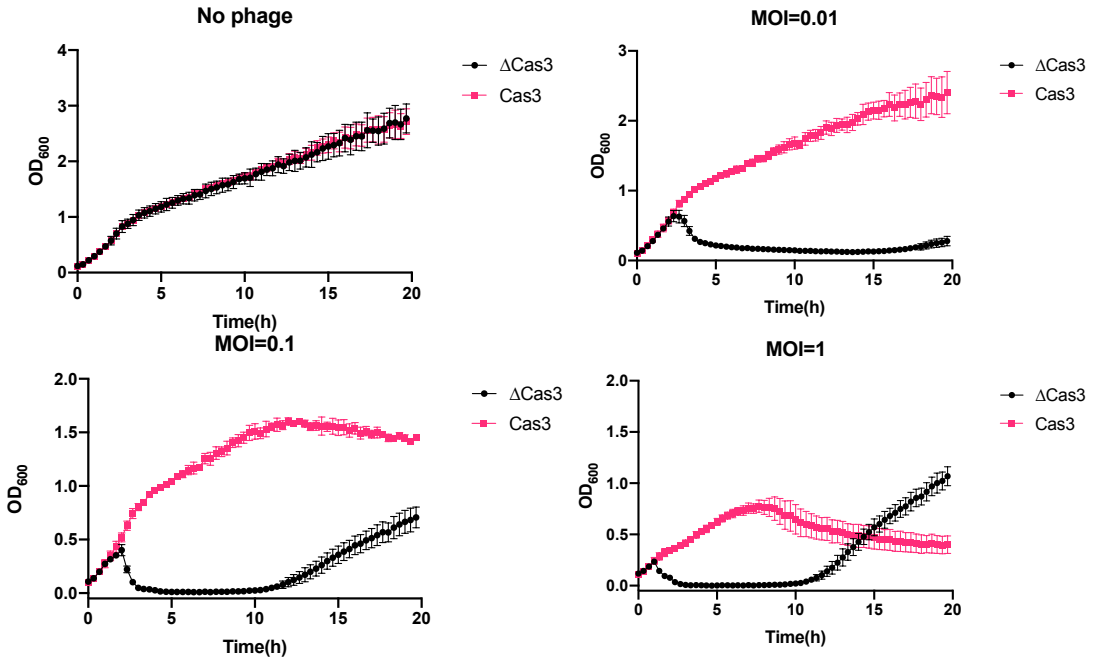


Figure 3.17. Phage challenge assay without induction

Cell growth curve without lactose and arabinose induction. No phage infection or phage infection (MOI=0.01, 0.1 and 1). **Δ**Cas3, cells lack of Cas3 gene. Data points represent the mean of six experimental replicates (two biological replicates and three technical replicates) with standard deviation shown.

3.3 Discussion

We have explored the mechanism of the type I-G expression and interference stages. In general, type I-G shares the features of common type I CRISPR systems, but it is divergent to other type I systems in terms of the details.

crRNA maturation of type I-G is executed by Csb2, the feature protein of type I-G. It generates a mature crRNA with an 8 nt 5'-handle and 3' hairpin, the common type I product, but instead of cutting the pre-crRNA at the bottom of hairpin like common type I[234], Csb2 cleaved at the 4 nt 3' end away from the base (Figure 3.3C), which is strikingly divergent. The extended 3' hairpin is probably necessary for Csb2 binding, the core protein that remains bound to the 3'hairpin after pre-crRNA cleavage.

When submitting Csb2 for long pre-crRNA (multiple repeats) cleavage (Figure 3.4), intermediate products were observed. A similar observation has been shown in type I-F system, where Cas6 of type I-F cleaves pre-crRNA, generation a series of product and the major products is the mature crRNA[236]. However, even at a high concentration of Csb2, the trace of mature crRNA product is still weakly shown on the gel. This might be attributed to the lack of Adenine nucleotides in the mature crRNA since the pre-crRNA is transcribed with [α-^{32}P]-ATP. More likely, the cleavage activity of Csb2 on long pre-crRNA is restricted by the complicated RNA structure. The pre-crRNA cleavage by Cas6 has high fidelity *in vivo* across type I CRISPR[234]. Despite no direct *in vivo* data showing the fidelity of pre-crRNA in type I-G, the reconstitution of type I-G *in vivo* successfully exerts target specific interference, suggesting the functional mature crRNA is generated *in vivo*.

Csb2, a bioinformatically predicted fusion of Cas5 and Cas6, was experimentally shown to harbour a Cas6-like C-terminal domain. The C-terminal domain alone is

sufficient to cleave pre-crRNA (Figure 3.8A), and even has a higher binding affinity with crRNA repeats compared to intact Csb2 (Figure 3.8B), but the yield of cleaved crRNA was slightly lower (Figure 3.8A). The Cas5-like N-terminal domain, however, cannot cleave pre-crRNA and is not associated with neither 5' -handle nor the 3'-hairpin (Figure 3.8). The most likely functional role of Cas5 is to structurally build the type I-G effector complex, its absence leads to a disruption of complex formation (Figure 3.11). But as we mentioned above, mature crRNA contains an abnormal 4 nt extension and intact Csb2 has a higher yield of mature crRNA, the Cas5-like N-terminal domain could still affect the crRNA maturation. The structure model from prediction shows that only first 150 residues match with Cas5 (Figure 3.6A), the uncommon end of N-terminal domain might be the reason why this Cas5-like domain is highly divergent to other Cas5 in type I systems.

Cas3 in type I-G system is pre-associated with Cascade before DNA targeting (Figure 3.9&3.12). A variant of type I-G system in *Bifidobacterium* has also shown that Cascade combined with Cas3 significantly improves DNA targeting[237]. This is a major divergence to the well-studied type I mechanism where Cas3 is recruited to Cascade upon binding to target DNA. A recent study on type I-A CRISPR system revealed that Cas3 of type I-A is also pre-binding to the Cascade, and Cas3 rigidified the PAM recognition subunit of type I-A Cascade, enabling target DNA binding[193]. The authors also proposed two model of type I system activation, the trans-recruitment of nuclease and the allosteric activation of nuclease. Cas3 of type I-G, like type I-A, can possibly be allosterically activated upon binding to target DNA. The difference between these two types of models may be attributed to the structure of Cas3, which we will discuss more in following chapters.

In the 80 bp dsDNA cleavage assay (Figure 3.14). The target strand (TS) of target dsDNA is not cleaved in the presence of ATP, which differs from the type I-E system, where TS of target dsDNA is cleaved when ATP is present and not cleaved when ATP is absent[204]. The two models of DNA degradation might be the explanation for this difference since the pre-associated Cas3 in type I-G probably interacts with PAM recognition subunit and hence lacks dynamic, while the Cas3 of type I-E is recruited to the NTS that is exposed by R-loop formation, where the same Cas3 has the potential to cleave the TS once the NTS is cleaved.

One thing should be noted is that the type I-G system we studied here is just a representative of type I-G systems. There is a diversity of type I-G variants, with a variable size in Cas8g (Figure 3.18). Cas8g in our study is the largest one out of three variants, also named as Cas8g1. Type I-G was characterized by the *csb2* gene, but those subtle differences in variants should not be ignored in future experimental study.

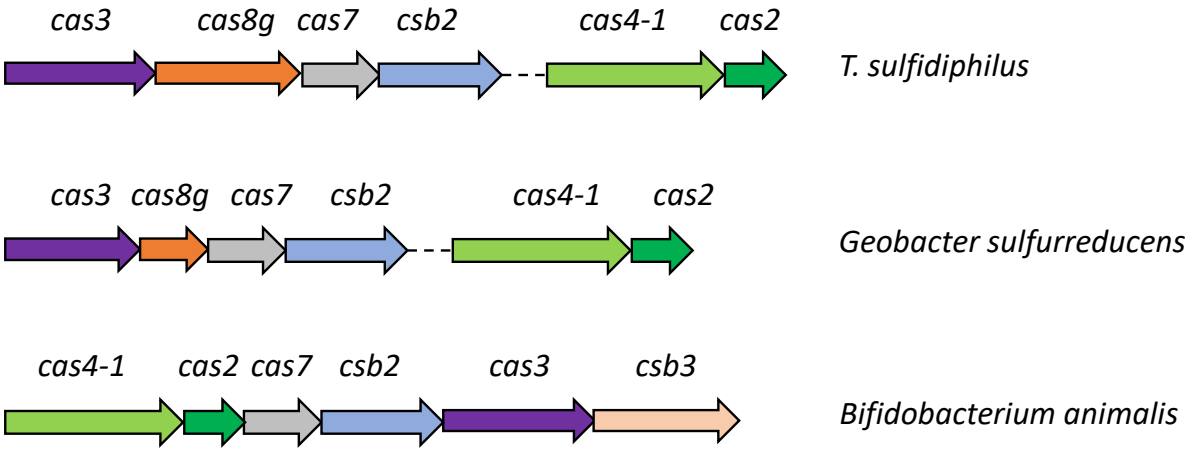


Figure 3.18. type I-G variants.

Adapted from Makarova *et al*[90]. Type I-G variants show variable large subunit Cas8g. Csb3 presumably is the large subunit of Cascade in *Bifidobacterium animalis*.

Chapter 4: Structure of type I-G effector complex

This chapter is adapted in part from the published manuscript: Structure and mechanism of the type I-G CRISPR effector[230].

4.1 Introduction

We discussed the structure of type I-E Cascade in the introduction section, showing the canonical structure of type I effector that comprises of a Cas7 backbone along the mature crRNA, forming a crescent shape. Cas6 remains associated with 3'-hairpin after pre-crRNA cleavage, Cas5 is positioned in 5'-handle, together with large subunit Cas8 and small subunit Cas11 in the centre of the groove (Figure 1.7C). The Cas7 backbone is conserved among type I subtypes while other Cas proteins vary in the complex construction.

To elucidate the structure of type I-G effector complex, we submitted *in vitro* expressed type I-G effector complex for single particle Cryo electron microscopy (cryo-EM). The Cascade complex obtained from *in vitro* incubation (Figure 3.9B) was submitted for cryo-EM. The complex consists of subunits Csb2, Cas7 and Cas8g with a 72 nt mature crRNA (Figure 4.1).

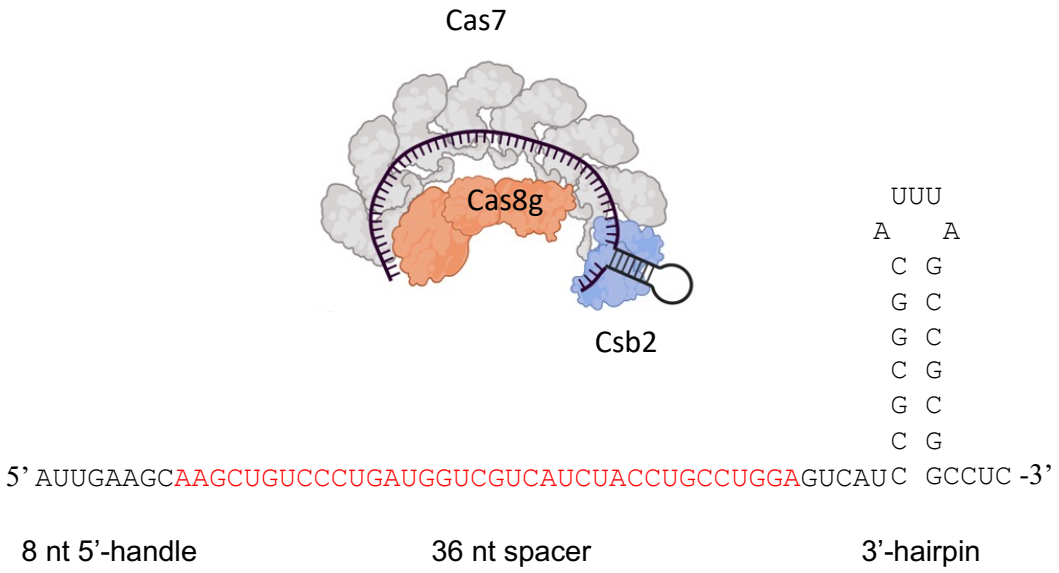


Figure 4.1. Cascade for cryo-EM.

A schematic of Cascade and the 72 nt mature crRNA, 36 nt spacer was in red.

The details of cryo-EM sample preparation, data collection and model building were carried out by our collaborator, Ramasubramanian Sundaramoorthy, which will not be discussed in this thesis. We will focus on the explanation of the structure and the comparison between type I-G Cascade and other type I complex in the following sections.

4.2 Result

4.2.1 Architecture of type I-G effector complex

4.2.1.1 Overview of type I-G effector complex

A first look at the type I-G Cascade showed distinctly distributed particles on the micrograph (Figure 4.2A). After 2D classification, a crescent shape of Cascade particles with prominent subunits was observed (Figure 4.2B). 3D model of type I-G Cascade further reveals the composition of the crescent shape, seven interlocking Cas7 subunits form the backbone of Cascade along the crRNA, Cas8g is positioned in the belly of the crescent (Figure 4.2C). Csb2 has high binding affinity for crRNA 3'-hairpin, but no clear density of Csb2 was observed in the area, while the complex for cryo-EM contains Csb2 stoichiometrically (Figure 3.9B), which suggests the intrinsic flexibility of this part of the complex, namely, the linkage of Csb2 to the rest of the complex is flexible. Same observation has been made in type I-A[193], type I-C[195] and type I-F[238] effector complex.

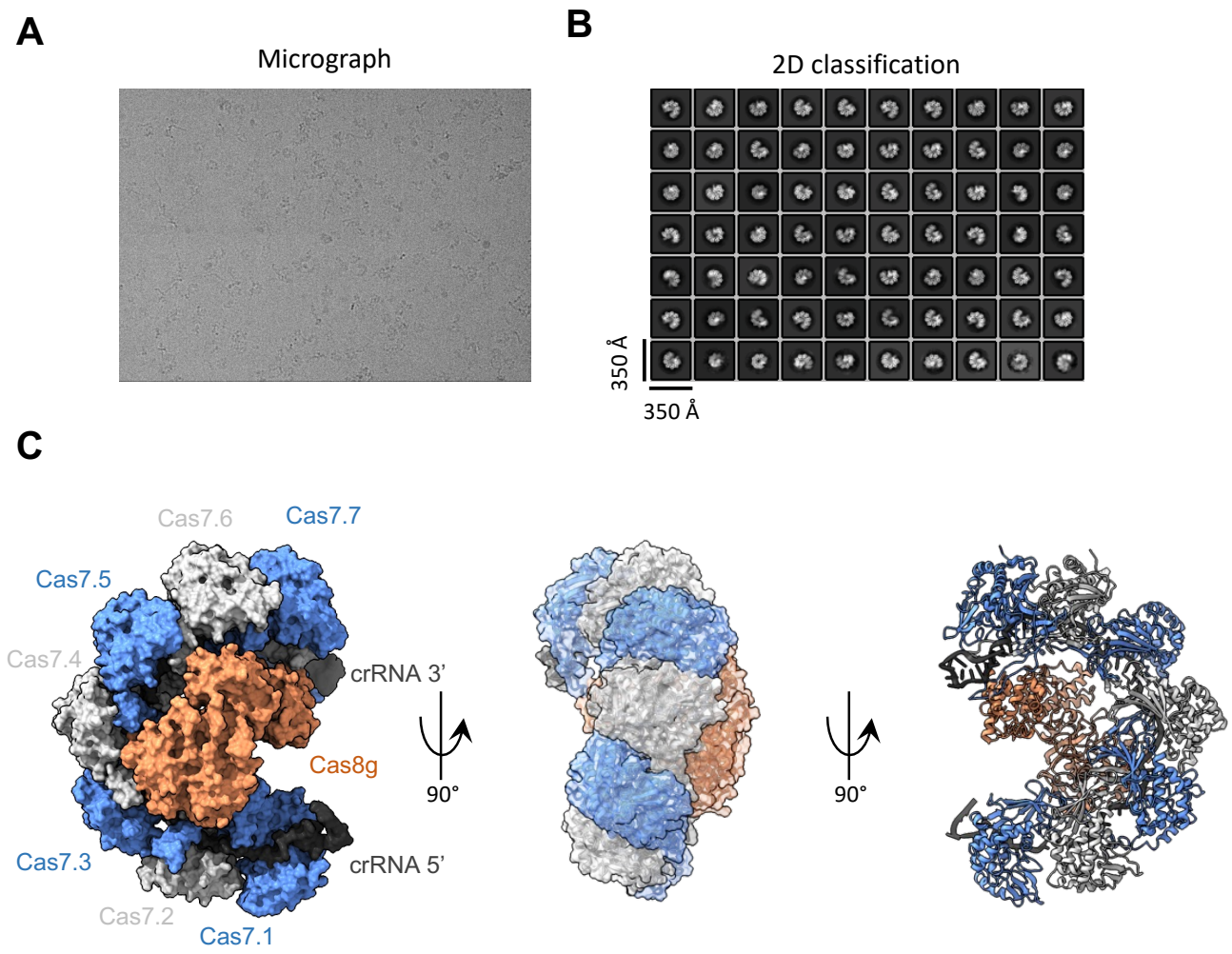


Figure 4.2. The overview of type I-G cascade.

(A) Representative micrograph of vitrified type I-G cascade complex. **(B)** Reference free 2D classification averages showing various projection images of type I-G cascade. **(C)** 180° rotated views of the Cryo-EM reconstructed maps of the Type I-G Cascade showing the arrangement of seven Cas7 subunits (Blue and grey staggered), the bound crRNA (black) and the large subunit Cas8g (orange). The volume corresponding to each subunit of Cas7, crRNA and Cas8g are segmented in ChimeraX and shown in surface representation. The refined structures of Cas7, crRNA and the large subunit Cas8g are placed within the Cryo-EM map and shown in cartoon representation. The seven Cas7 subunits adopt a crescent shaped architecture with the large subunit Cas8g at the belly.

4.2.1.2 Cas7-crRNA backbone organisation

The Cas7-crRNA backbone was refined to an overall resolution at 3.2Å. The five centre subunits have a higher resolution compared to the crRNA 5'-handle and 3'-hairpin. The crRNA interaction with Cas7 backbone is visualized (Figure 4.3A). We used the Alphafold2 predicted Cas7 structure as a starting model. Cas7 of type I-G exhibits a canonical central RAMP (Repeat Associated Mysterious protein) domain with extended β-hairpin, each Cas7 occupies 6 nucleotides of crRNA with a 5+1 nt pattern, where the sixth base flipped out in the opposite direction to the remaining five bases. 5'-handle is extended out and bent back while 3'-hairpin also exposes out the Cas7 backbone. The spacer sequence for targeting, complementary to the target DNA, is encompassed by Cas7 backbone (Figure 4.3B).

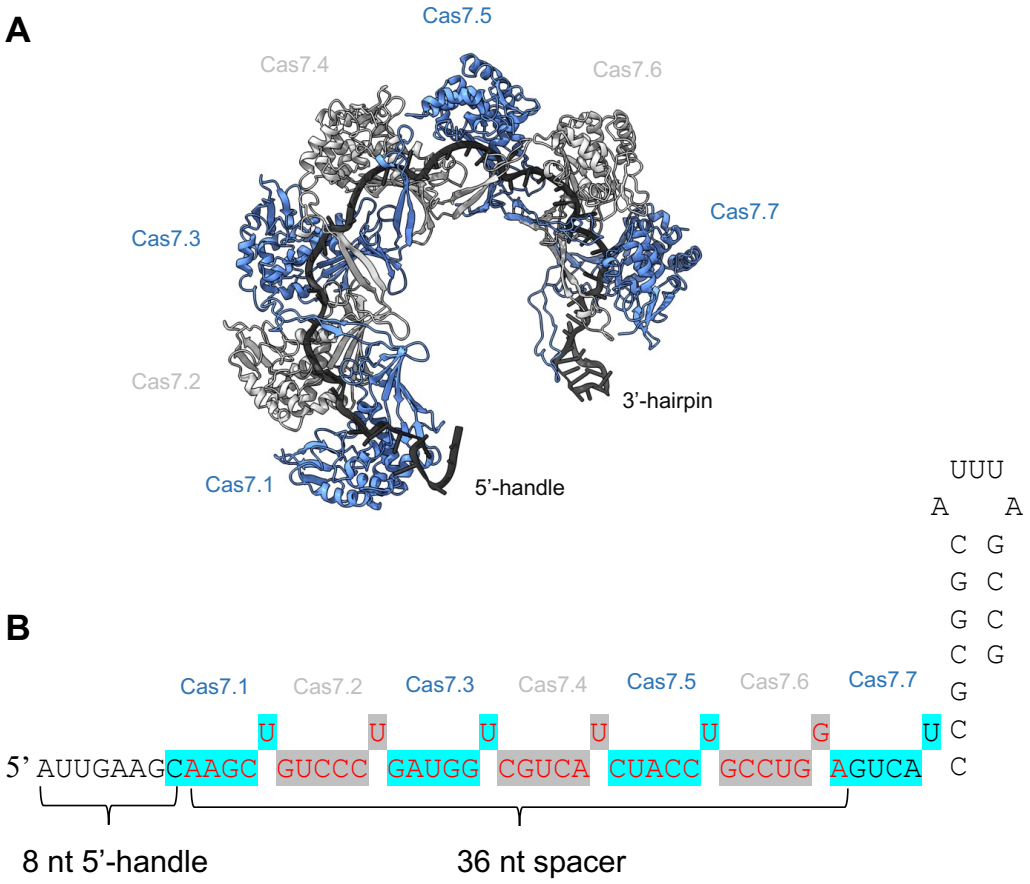


Figure 4.3. Organisation of Cas7 backbone and crRNA

(A) Refined structure of interconnected Cas7 subunits with bound crRNA shown in cartoon representation. The 5' end of the crRNA contains 8 bases of handle which was visible closer to the edge of the Cas7.1, and the 3' end of the crRNA constitutes 17 bases of the stem loop which was located immediately adjacent to the Cas7.7 subunit. **(B)** A schematic figure showing the 5+1 pattern and a flipped base in the crRNA. 36 nt spacer was in red.

The central region of Cas7 backbone (Cas7.3 and Cas7.4) is clearly visualized, crRNA 5+1 pattern is shown in density, where crRNA U25 and U31flipped at the opposite direction (Figure 4.4B). Some side chains of Cas7 could also be visualized in the density map (Figure 4.4C).

Overall, the structure of Cas7-crRNA backbone was elucidated at a high resolution (3.2Å), and it possesses a typical type I CRISPR organisation with interlocking Cas7 subunits spanning crRNA and a 5+1 nt pattern of crRNA positioning.

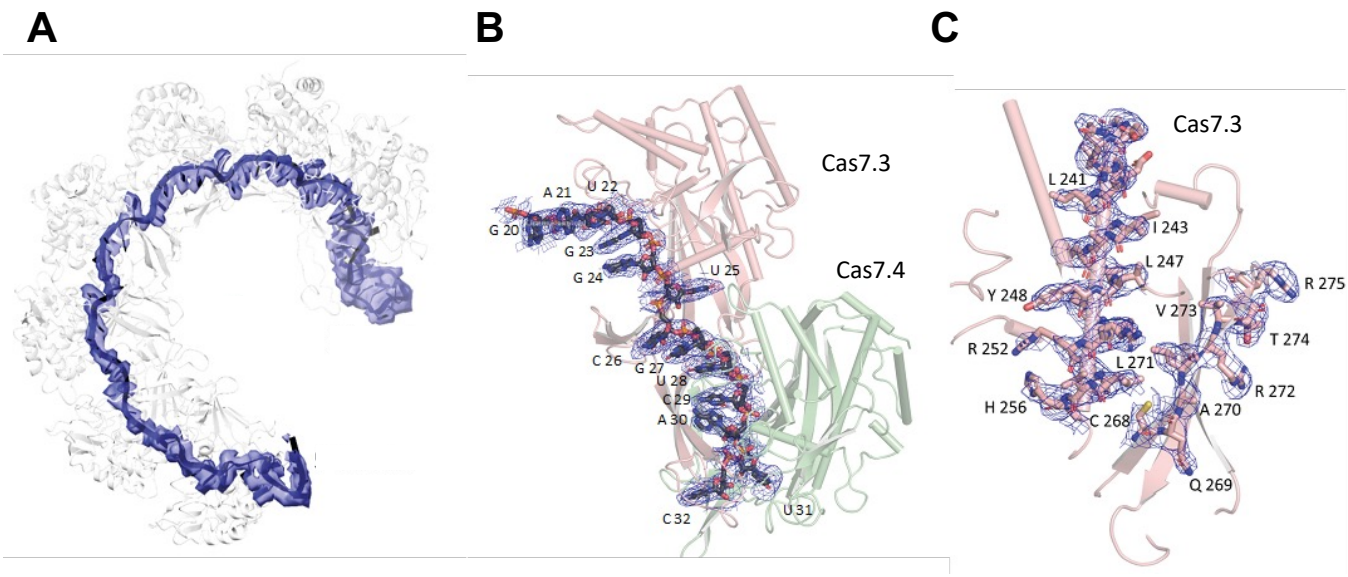


Figure 4.4. Architecture of Cas7 and crRNA

(A) Cryo-EM density for the crRNA is drawn in surface representation with the refined crRNA shown in cartoon representation. **(B)** A closer view of interaction of crRNA segment with the selected Cas7 subunits. The Cas7 subunits are shown in cartoon representation and the crRNA drawn as stick. The Cryo-EM density for the crRNA segment is shown in mesh. **(C)** Selected regions of Cas7 residues side chains are shown in stick and the corresponding Cryo-EM density drawn in mesh. Residues are numbered.

4.2.2 The large subunit, Cas8g

4.2.2.1 Cas8g, a “large and small” subunit

Cas8g locates in the belly of the crescent shape, the overall resolution of Cas8g is significantly lower than the Cas7-crRNA backbone, at 8.2Å. Cas8g, the large subunit

of type I-G system, has no detectable sequence to any other Cas8 protein. We used AlphaFold2 to predict its structure as a starting model. N-terminus of Cas8g was predicted as a α+β mixed domain while C-terminus of Cas8g is an α-helical domain (Figure 4.5A&45A). In the resolution scale we obtained, the N-terminal domain of Cas8g has an extensive interaction with the first two Cas7 subunit (Cas7.1 and Cas7.2), the specific residues for interaction cannot be elucidated due to the lack of resolution, but the crRNA 5'-handle is not involved in this interaction, which remains distant to Cas8g N-terminal domain (Figure 4.5B).

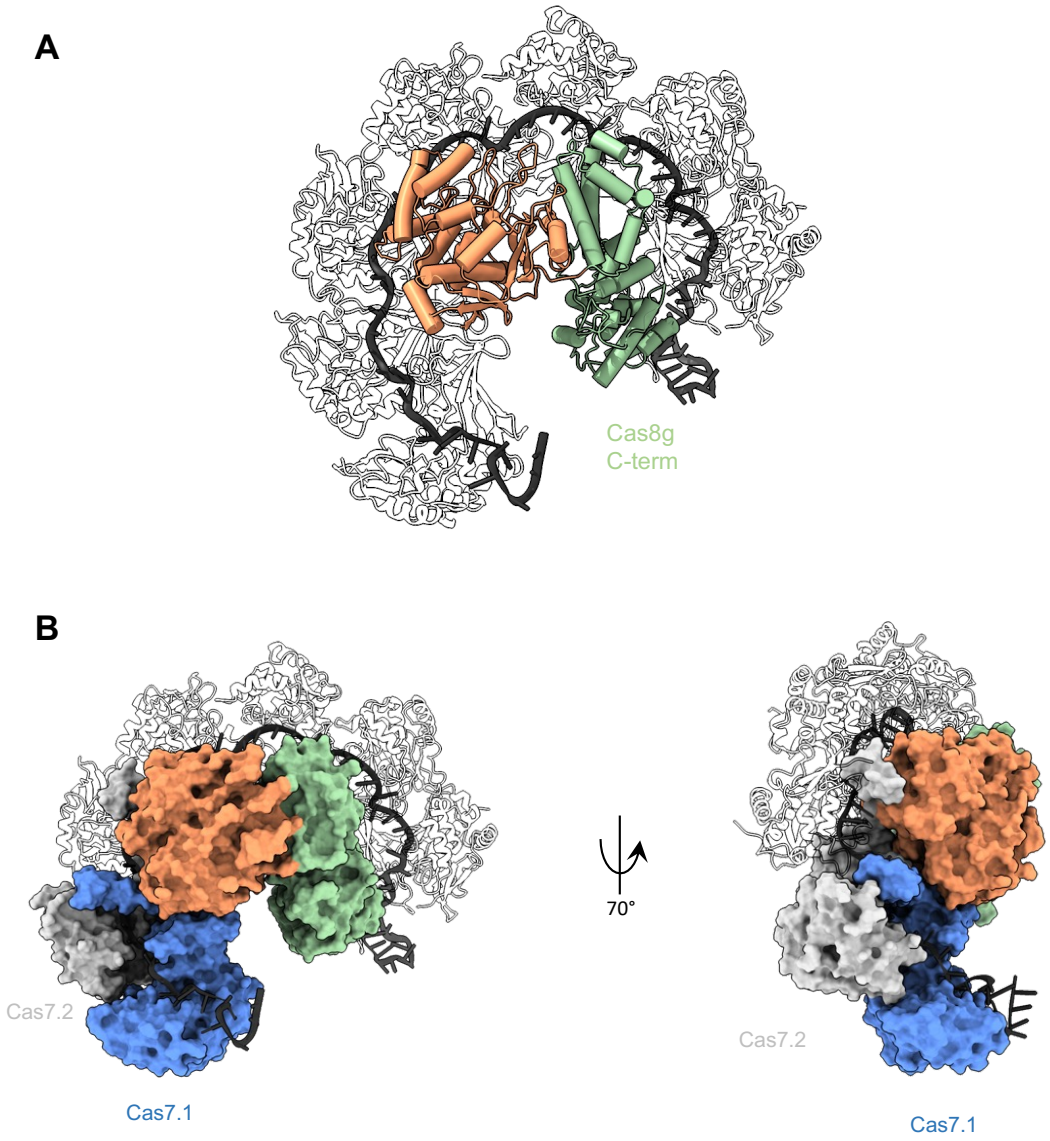


Figure 4.5. Large subunit Cas8g

(A) Cryo-EM refined structural model of the large subunit Cas8g shown in cartoon representation. The large subunit constitutes two distinct sub domains. The N-terminal domain of Cas8g (orange) adopts a mixed α and β fold. The C-terminal domain (light green) is reminiscent of Cas11 and was placed closer to the 3' end of the crRNA and the Cas7.7 subunit. **(B)** The N-terminal domain has extensive interaction with first two Cas7 (Cas7.1 and Cas7.2) subunits and the bound crRNA within this region. Due to lack of resolution, specific interacting residues could not be deciphered.

Since no significant hits for the Cas8g model in current protein data bank, we attempted to search the structural hits using Foldseek[239]. By comparing the Alphafold2 predicted structures in the complete Swissprot database, a significant hit for Cas8g N-terminal domain was detected. Cas8a2 subunit of type I-A system from *Methanocaldococcus jannaschii* and related archaea is predicted to match the N-terminal domain of Cas8g structurally (Figure 4.6B).

There were no structural hits for the C-terminal domain of Cas8g, but the α-helical construction indicates the resemblance of the small subunit Cas11 in Class I CRISPR effectors[240]. With a Cas11-like C-terminal domain and a Cas8-like N-terminal domain, Cas8g is highly divergent to other type I large subunit Cas8. However, we can still find clues in other type I systems that share the similarity to the arrangement of the "large and small" subunit. Cas8f protein from type I-F CRISPR also possesses two distinct domains with an α-helical rich domain positioned in the centre of the Cascade[198]. Type I-D effector complex requires a small subunit Cas11 that expressed from its large subunit gene *cas10d* to fully function[241]. Overall, the large subunit Cas8 in type I systems holds sequence and structural diversity, and Cas8g from type I-G has its unique arrangement in the structural organisation of effector complex.

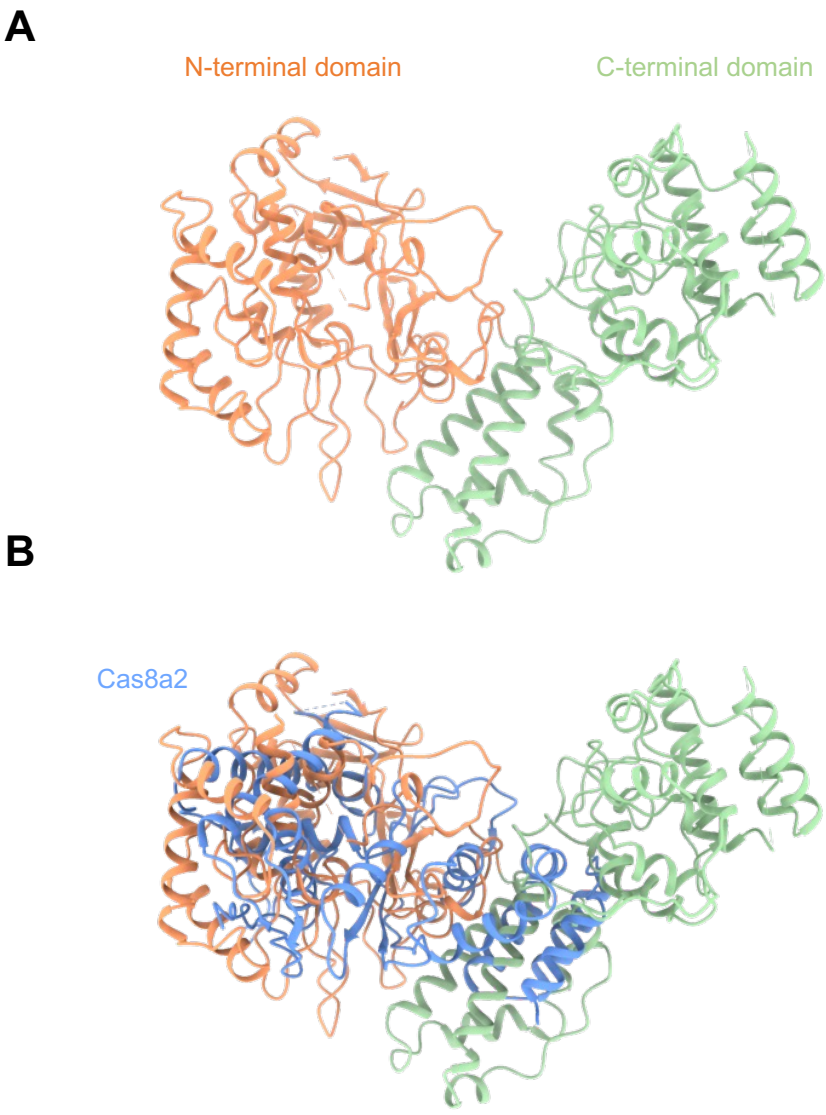


Figure 4.6. Structure alignment of Cas8g

(A) The N-terminal domain (orange) has a mixed α+β secondary structure while the C-terminal domain (green) is predicted for α helical bundle, similar to the composition of the Cas11 subunit in other effector complexes. **(B)** Structural overlay with the Alaphfold2 model of Cas8a2 (blue) from *Methanocaldococcus jannaschii*.

4.2.2.2 Mutations on Cas8g disrupt complex architecture

To further investigate the mysterious protein Cas8g, we designed site directed mutations of Cas8g to study the functional changes of the variants. Four sites were mutated based on the structure and sequence conservation, Cas8g sequence alignment data was showed in Appendix Figure 1. Three mutated sites locate in the N-terminal domain, including two surface area, R143 and R270-R271, arginine was mutated to glutamic acid for charge reversal. Inside the N-terminal domain, a loop was also mutated (ND176AA, 176 asparagine and 177 aspartic acid were mutated to alanine). The 662 and 663 arginines near the crRNA 3'-hairpin in the C-terminal domain of Cas8g were mutated to glutamic acid (Figure 4.7).

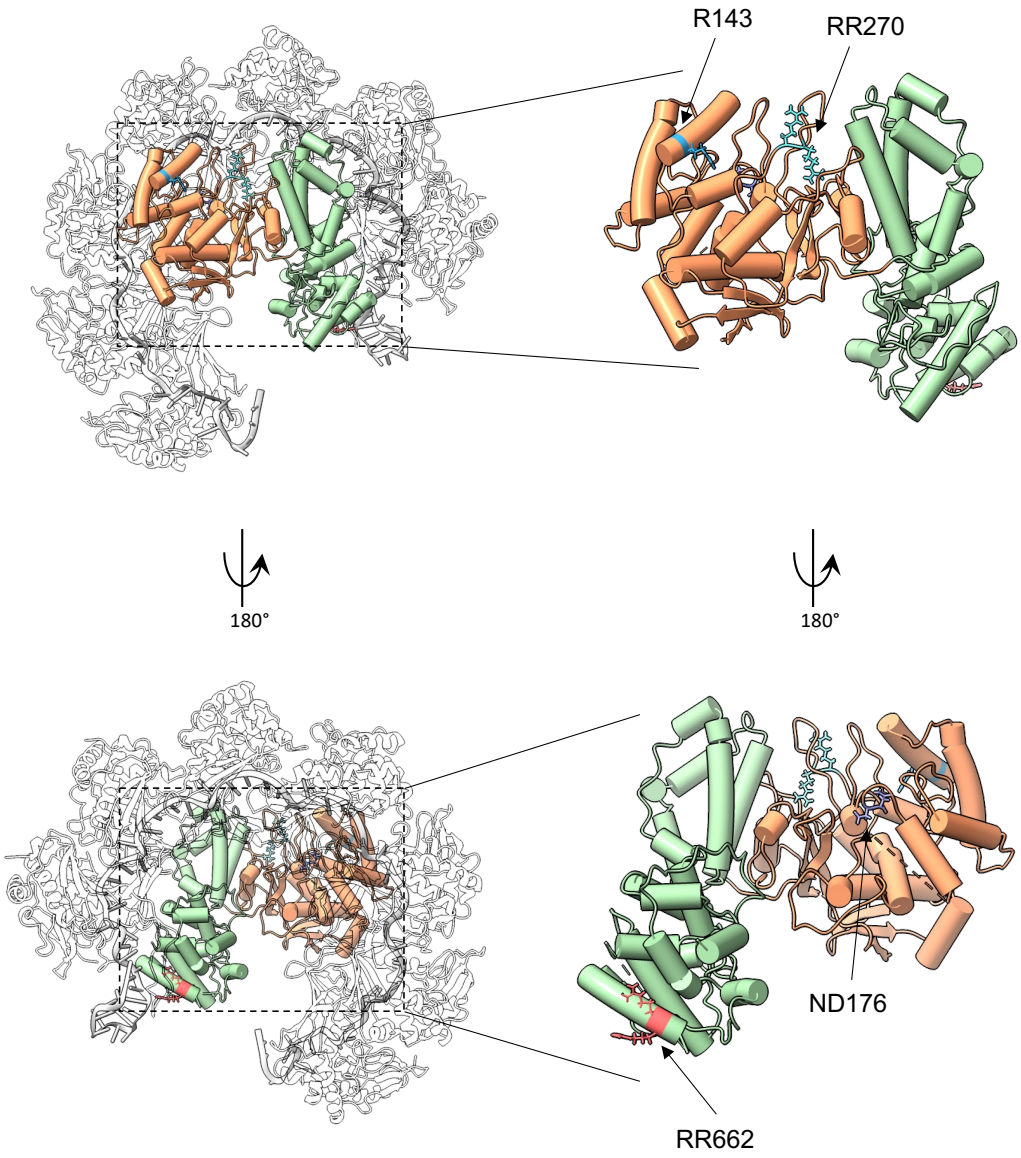


Figure 4.7. Mutations on Cas8g

Mutation sites on Cas8g showing in structural model. Two mutation R143E (143 arginine mutated to glutamic acid) and RR270EE (270 and 271 arginines mutated to glutamic acid) are on the surface of the N-terminal domain of Cas8g (orange), with one internal mutation ND176AA (176 asparagine and 177 aspartic acid mutated to alanine). RR662EE (662 and 663 arginines mutated to glutamic acid) is at the end of C-terminal domain (light green). Mutation sites were indicated by changed colour of atom representation and arrows. The Cas8g sequence alignment data is shown in Appendix Figure 1.

We constructed the mutated Cas8g sequence into expression vector and successfully expressed and purified the Cas8g mutants. Once we obtained the Cas8g variants, we first investigated the effect of Cas8g variants on complex formation by *in vitro* reconstruction of the type I-G effector. Intriguingly, Cas8g variants exerted different influence on the *in vitro* complex formation. After the incubation with Csb2, Cas7, Cas3 and the pre-crRNA, all samples that contained Cas8g variants eluted in complexes, an early peak showing on the chromatography (Figure 4.8A). But the composition of this eluted peak reveals the changes of the complex. Effector complex from ND176AA variant, the inner mutation of Cas8g N-terminal domain, still contained all 4 *cas* proteins like the wild-type Cas8g (Figure 3.9C), the architecture of type I-G effector is not disrupted. But the two surface mutations on N-terminal domain, R143E and RR270EE led to a loss of Cas3 in the complex composition, suggesting the interaction between Cas8g and Cas3 through the surface of Cas8g N-terminal domain. The complex formation of RR662EE variant, C-terminal domain mutation, results in a different composition where Cas8g and Cas3 are nearly missing, only weak association was detected on SDS-PAGE gel (Figure 4.8B). This might be attributed to the interaction between Cas8g and the backbone (Csb2 and Cas7) was disrupted, suggesting that Cas8g interacts with backbone, most probably Csb2, through its C-terminal domain.

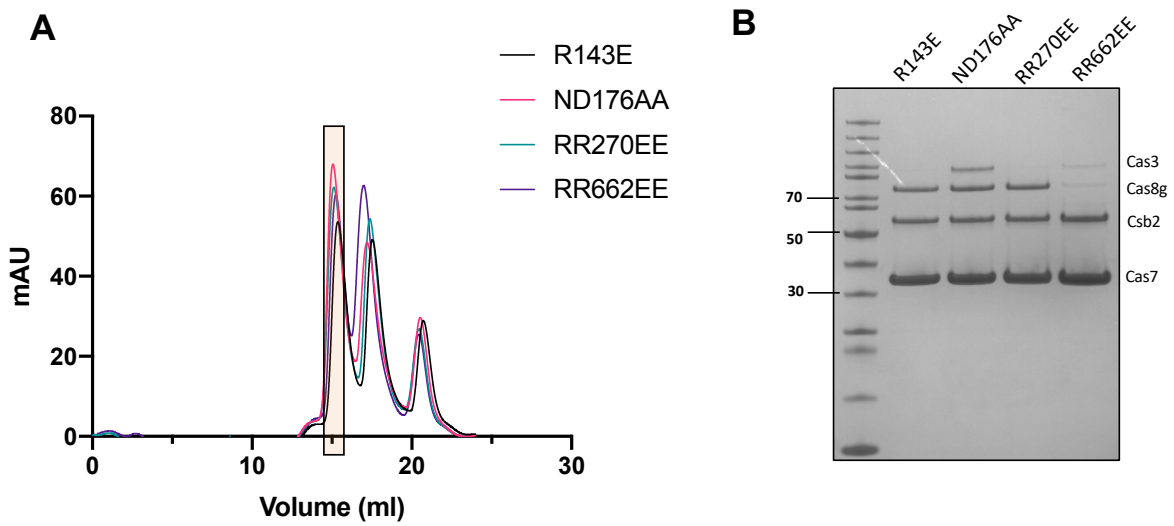


Figure 4.8. Complex formation disrupted by Cas8g mutants

(A) Cas8g mutants, Cas3, Csb2 and Cas7 were incubated with pre-crRNA and subjected to size exclusion chromatography. The resulting chromatograph was shown, and the boxed fraction was submitted for SDS-PAGE. **(B)** SDS-PAGE analysis of the indicated fractions.

We proceeded to check the effector complex target ability with the complex variants we obtained from *in vitro* incubation (Figure 4.9). Since Cas8g and Cas3 are missing in RR662EE mutated complex, it completely loses the ability for binding target dsDNA. However, the binding ability of RR270EE mutated complex was completely abolished even though Cas8g is present in the composition, not like the Cascade from wide-type Cas8g where a weak binding can still be detected, indicating the RR270 mutation is not only blocking the interaction of Cas3 but crucial for dsDNA targeting as well. R143E mutated complex had the same composition as the RR270EE complex (Cas3 missing), but it shows a binding affinity reminiscent of the wide-type Cascade instead of not binding at all, suggesting the change in R143 only disrupts Cas3 association and R143 is not involved in dsDNA targeting. Surprisingly, ND176AA mutated complex

which comprises of all 4 subunits, has a significant decrease in target dsDNA binding affinity compared to the wide-type Cascade-Cas3 complex. This inner mutation in Cas8g N-terminal domain might be involved in the process of DNA targeting, PAM recognition for example.

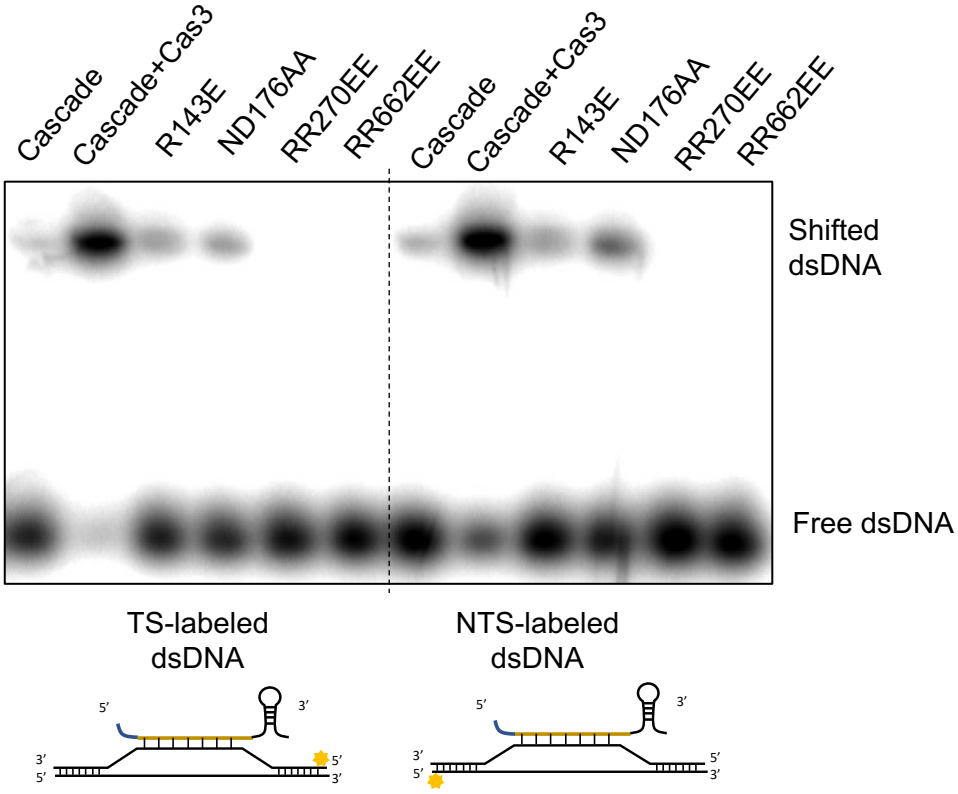


Figure 4.9. target dsDNA binding with Cas8g mutated complex

Electrophoretic Mobility Shift Assay (EMSA) shows the target dsDNA binding with Cas8g mutated complex. RR270EE and RR662EE mutated complex completely abolished the binding with target dsDNA, while R143E and ND176AA mutated complex showed weaker binding. 10 nM target dsDNA was incubated with 0.8 μM complex.

In vitro reconstruction of the Cas8g mutated complex reveals the potential sites in Cas8g that contribute to Cas3 interaction and dsDNA targeting. We then focused on

investigating the behaviour of these Cas8g variants *in vivo* with plasmid challenge assay and phage immunity assay.

R143E and RR270EE mutated complexes have same composition, but differ in target DNA binding *in vitro*, where RR270EE fully aborts the targeting while R143E still shows a weak targeting activity (Figure 4.9). *In vivo* expression of those two variants leads to a dramatic difference in plasmid challenge assay (Figure 4.10). RR270EE strain completely loses the ability to eradicate invasive plasmid even under fully induced condition, but R143E strain shows a strong interference on the target plasmid even without induction. The ND176AA mutated complex, all 4 components in composition but low targeting efficiency *in vitro*, exhibits a surprising outcome upon plasmid challenge, it also abolishes the target plasmid clearance *in vivo* (Figure 4.10). RR662EE mutated complex loses its DNA targeting ability *in vitro* but gains a high interference activity *in vivo* when the complex was fully induced (Figure 4.10). Those data suggest that the high expression of type I-G complex *in vivo* can recover the target DNA interference that might be altered by site directed mutation on Cas8g *in vitro*, R143E and RR662EE, for instance. These two site mutations might weaken the type I-G interference activity but not essential for the process. On the contrary, ND176AA and RR270EE mutations dramatically change the activity of type I-G interference, indicating the importance of the two sites on Cas8g for target interference. Compared mutated Cas8g complex data with Chapter 3 data, *in vitro* binding assay (Figure 3.12B) and *in vivo* plasmid challenge and phage immunity assay (Figure 3.15&3.16), we also notice that R143E mutated complex loses the ability to bind Cas3 *in vitro*, exhibiting the same behaviour as Cascade lacking Cas3, weakly binding with target dsDNA, however, *in vivo* assays shows that R143E mutated Cascade with Cas3 expression behaves like wide type Cascade with Cas3, fully interference *in vivo*. Wild-

type Cascade alone (ΔCas3, no Cas3 expression at all) cannot accomplish plasmid eradication and phage immunity. R143E mutated Cascade, despite the weak interaction with Cas3 *in vitro*, can bind Cas3 *in vivo* and activates interference, suggesting R143 of Cas8g might get involved in subunits interaction with Cas3 but not critical to abolish Cas3 binding to target dsDNA *in vivo*.

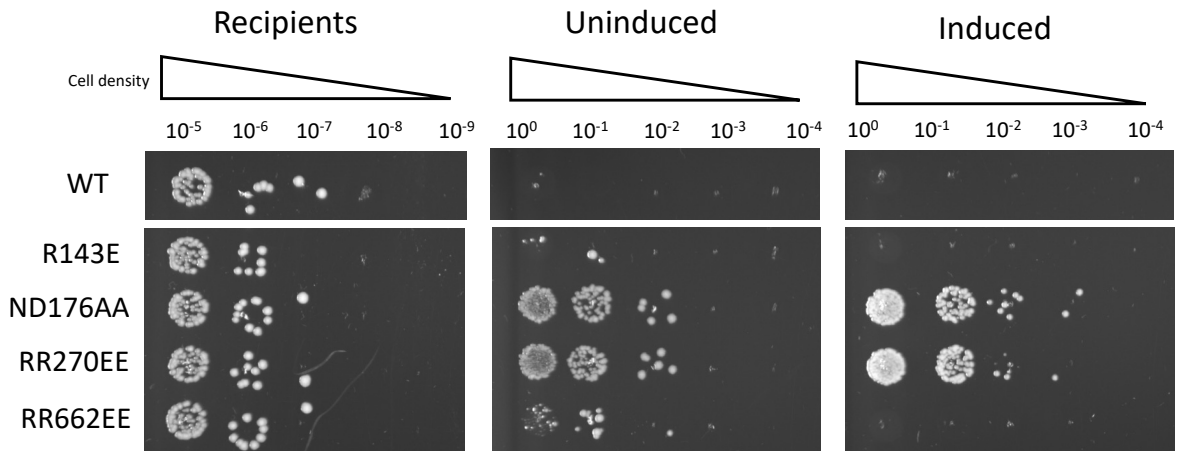


Figure 4.10. Plasmid challenge assay on Cas8g mutated complex

Cells on the plates in different condition. Recipients, Ampicillin and Spectinomycin in plates; Transformants, Ampicillin, Spectinomycin and tetracycline in plates; Induced, with all three antibiotics and lactose, arabinose for induction.

Phage immunity assay on Cas8g mutated complexes showed a consistent outcome to the plasmid challenge assay (Figure 4.11). Under fully induced condition, R143E and RR662EE mutated complexes effectively protected cells from phage infection, but ND176AA and RR270EE mutated complexes no longer held immunity to phage infection.

In summary, the mutations on Cas8g reveal its crucial role in type I-G CRISPR interference and potential sites of interaction with other Cas proteins and target DNA.

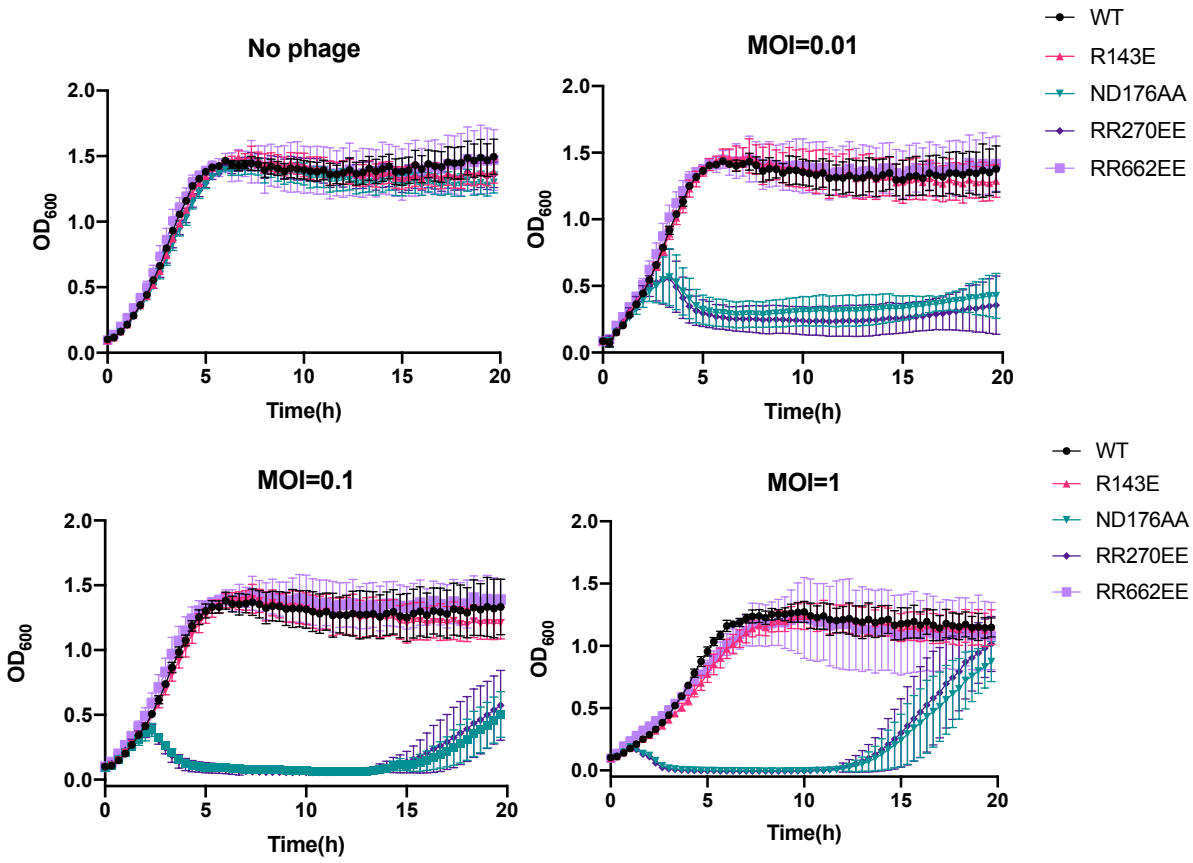


Figure 4.11. Phage challenge assay with Cas8g mutated complex

Phage immunity assay. Cas8g mutated type I-G system established in *E. coil*, CRISPR array targeting phage P1 *lpa* gene. Cell growth curve, no phage infection or phage infection (MOI=0.01, 0.1 and 1). Data points represent the mean of six experimental replicates (two biological replicates and three technical replicates) with standard deviation shown.

4.3 Discussion

In this chapter, we showed the structure of type I-G Cascade complex obtained from cryo-EM, a crescent shape Cas7-crRNA backbone with Cas8g located in the belly.

Type I-G crRNA in the backbone shares a canonical 5+1 nt pattern in RNA arrangement, but the curvature is different from other subtypes, reminiscent of type I-

F (Figure 4.12). 7 Cas7 subunits span the crRNA, each occupies 6 nt of crRNA. Different from the backbone of type I-D, I-E and I-F, where only 6 Cas7 form the backbone. Type I-A, I-C and I-G all require 7 Cas7 for complex formation. The length of crRNA determines the number of Cas7 subunit on the complex.

Browsing the structure of type I effector complexes (Figure 4.12), the Cas5 subunit is consistently present at the 5' end of crRNA, intensively interacting with the large subunit Cas8, and involved in PAM recognition. In the type I-G structure, however, the Cas5 subunit is absent. The signature gene of type I-G system, Csb2, a fusion of Cas5 and Cas6 has been discussed in chapter 3. This fusion probably leads to the missing Cas5 subunit in the structure, since Csb2 has high binding affinity to crRNA 3'-hairpin instead of the 5'-handle (Figure 3.7). The N-terminal domain of Cas8g is, on the other hand, distant from the 5' end of the crRNA. Together with the absence of Cas5, showing the most divergent point of type I-G effector from other subtypes. This also opens the question how the PAM is recognised by type I-G effector. The likely explanation is that the structure of Cas8g is dynamic in the belly and has a good chance of conformational change upon binding target dsDNA. Cas3 is pre-associated with Cascade in type I-G, like Cas3 in type I-A. In fact, Cas3 of type I-A binds to Cas8a fixes the Cas8a N-terminal domain, enabling PAM recognition by Cas8a N-terminal domain[193]. Cas3 of type I-G is associated to the Cascade by the interaction with Cas8g (Figure 3.10). It could also reduce the dynamic of Cas8g in type I-G system, leading to reliable PAM recognition. As we have seen in the biochemical data, Cas3 pre-associated with Cascade significantly increases DNA binding (Figure 3.12).

Cascade alone binds to target dsDNA at a low binding affinity (Figure 3.12B&4.9), and the addition of Cas3 significantly improves the binding affinity. Cas3 is pre-associated with Cascade by interacting with Cas8g (Figure 3.10), the interaction between Cas3

and Cascade prior to the dsDNA binding likely leads to a conformation change where the PAM recognition subunit would rigidify for a better dsDNA targeting. Type I-A effector complex has been demonstrated to be in line with this model, where the Cascade of type I-A has a weak binding with target dsDNA, and the addition of Cas3' and Cas3'' enables a better structural change for PAM recognition[193]. However, the binding between Cascade alone and target dsDNA is not neglectable, namely, even without Cas3, Cascade of type I-G and type I-A can still bind to target dsDNA, assumably, Cas3 would also be recruited upon dsDNA recognition. The pre-association model and the recognition-recruitment model might co-exist in the type I-G and type I-A systems. This can explain why R143E mutated Cascade still activated *in vivo* (Figure 4.10&4.11) while it is not pre-associated with Cas3 *in vitro* (Figure 4.8), with the binding between R143E mutated Cascade and target DNA *in vivo*, may occur at a low frequency, Cas3 can be recruited to the R143E mutated Cascade, despite disrupted interaction between R143E Cas8g and Cas3. This could lead to an interesting assumption that the search of target dsDNA evolves from the recognition-recruitment model solely to the co-existence of the pre-associated model. The driving force of this evolution is likely to be the pressure of anti-CRISPRs. In this scenario, anti-CRISPR proteins would block the target dsDNA binding by interacting with Cascade or abolish the Cas3 activity to depress CRISPR immunity, and the bacteria would counterattack the anti-CRISPR proteins by building a pre-associated effector complex to dismiss the influence from anti-CRISPR proteins. Evidence that supports this assumption has been observed: anti-CRISPR protein AcrIF1 of type I-F system prevents target DNA binding by binding to Cascade[67], AcrIF11 modifies PAM recognition subunit to prevent target dsDNA recognition[242], AcrIE1 of type I-E system binds to Cas3 to abolish Cas3 nuclease activity[243]. Those interruptions from anti-

CRISPR proteins could be overcome by building a pre-associated Cascade-Cas3 complex.

Another feature of type I effector complex is the small subunit, Cas11 that constantly locates in the belly of the Cascade. Cas11 subunit of type I-B, type I-C and type I-D systems is encoded within the gene of larger subunit[241], small subunit Cas11 is required for Cascade formation, however, in type I-G system, there is no small subunit expression from the large subunit *cas8g* gene, instead the small subunit seems fuse to the large subunit Cas8. The C-terminal domain of Cas8g has an α-helical rich structure and is located in the belly, in line with the canonical Cas11. This could be a reduction of expression burden and make type I-G more compact in terms of subunits numbers.

The curvature of the type I-G effector complex structure is close to type I-E and type I-F systems and distinct from type I-C and type I-D that has a similar curvature to type III CRISPR systems[195, 241]. Type I-D has shown type I and type III features on the biochemical level since it carries both Cas3 and Cas10, the signature proteins of type I and type III systems[196]. The structure of type I-D also suggests that type I-D system is a halfway house between type I and type III systems on the evolutionary level.

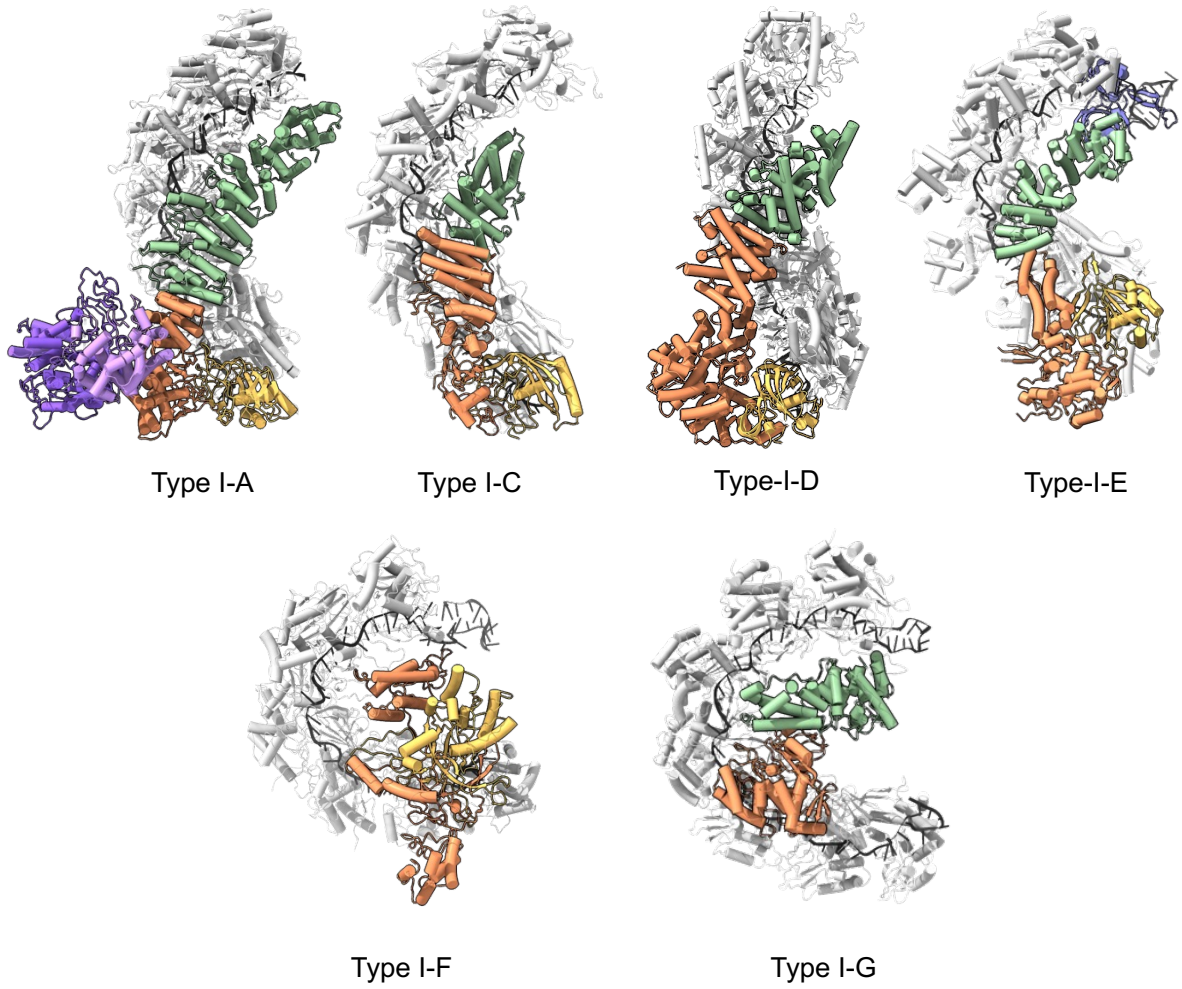


Figure 4.12. Comparison of type I CRISPR effector structures

The crRNA is shown in black and Cas7 backbone in light grey. Cas5 is shown in yellow, Cas6 in blue, Cas8 in orange, Cas11 in green. For type I-G, Cas8 N-terminal domain is orange and C-terminal domain green. For type I-A, the Cas3 HD (pink) and helicase (violet) proteins are also shown. For type I-D, Cas8 is replaced with a Cas10 subunit, also in orange, and bound target RNA is also present in the structure. Structures shown are: type I-A[193] , type I-C[195], type I-D[244], type I-E[245], type I-F[238] and type I-G (this study).

Chapter 5: Genome editing in prokaryotes by type I-G system

This chapter is adapted in part from the published manuscript: Repurposing the atypical Type I-G CRISPR system for bacterial genome engineering[246].

5.1 Introduction

The type I-G CRISPR system specifically targets dsDNA and leads to DNA degradation. Like other CRISPR systems, this feature can be repurposed for genome editing. But in contrast to the paradigm of type II Cas9 editing, type I CRISPR constantly generates long-range DNA break if no desired repair templates provided. Very recently, a type I-G variant in *Bifidobacteria* has been repurposed for self-genome editing[237], a showcase for exploiting endogenous CRISPR system, demonstrating the potential for introducing type I-G system to heterologous organisms on genome editing.

Since we successfully built *T. sulfidiphilus* type I-G system in the heterologous organism *E. coli*, we attempted to redirect the type I-G system for *E. coli* genome editing. To achieve genome editing, we redesigned the type I-G expressing vector. Four *cas* genes (*cas3, cas8g, cas7 and csb2*) of type I-G were cloned into pM2 vector under arabinose inducible promoter control. The CRISPR array was cloned into pSPACER for generating pre-crRNA (Figure 5.1 A&B). pM2 vector was transformed into *E. coli* MG1655 strain first, pSPACER was then introduced to build the type I-G effector complex. To visualize genome editing outcomes, *lacZ* gene on *E. coli* genome was targeted. Cells with intact *lacZ* gene remain blue on the X-gal plates while *lacZ* edited cells appear to be white. Spreading the cells on the X-gal plates after

transformation results in both blue and white colonies, white colonies indicate that *lacZ* gene was successfully edited on its genome (Figure 5.1C&D).

With this rationale, we then investigated the type I-G editing in detail.

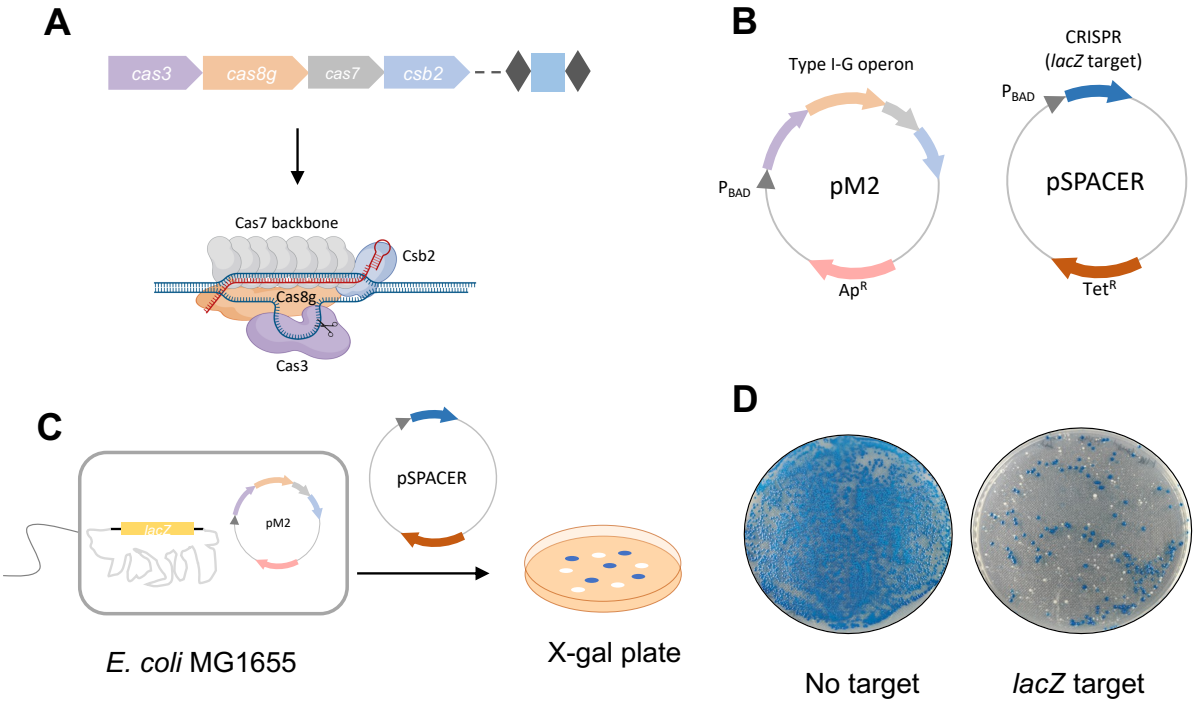


Figure 5.1. Target editing by type I-G

(A) An overview of type I-G CRISPR operon from *Thioalkalivibrio sulfidiphilus* HL-EbGr7. Cas proteins (7 x Cas7, Csb2, Cas8g and Cas3) together with crRNA form the effector complex and target dsDNA. **(B)** pM2 vector contains type I-G operon (*cas3*, *cas8g*, *cas7* and *csb2*) under arabinose promoter control; pSPACER vector, with CRISPR repeat and spacer (target *lacZ*) under arabinose promoter control. **(C)** A schematic of building type I-G system in *E.coli*; pM2 was transformed into *E. coli* MG1655 strain *(lacZ* intact) first, and pSPACER was introduced to activate type I-G *lacZ* targeting, cells were spread on the X-gal plates for blue-white screening. **(D)** A representation of blue-white screening assay with no target control and *lacZ* target spacer.

5.2 Results

5.2.1 Type I-G for genome targeting

5.2.1.1 Type I-G genome targeting decreases cell survivability

When repurposing the type I-G for *lacZ* genome targeting, a significant decrease in cell number was observed with the generation of white colonies (Figure 5.1D). Targeting the genome results in around a 2 orders of magnitude decrease in transformation efficiency compared to empty vector control, EV, with no target spacer (Figure 5.2A). Even without induction, cell loss could still be observed, and the transformation efficiency was comparable to the arabinose induced condition, suggesting a low level of type I-G expression from promoter leakiness was sufficient to target the genome. The cell survivability decrease has been observed in endogenous genome targeting from other type I systems[247, 248], heterologous Cas9 targeting on bacterial genome also hugely decreased cell number[249]. The cell loss is most likely attributed to double-strand DNA breaks (DSB) and/or deletions of a flanked essential gene generated by genome targeting. Bacteria lack efficient DNA repair to recover DNA damage, hence, causing cell death upon genome targeting. But cells can overcome the DNA damage and survive at a low frequency. On blue-white screening, a small number of white colonies was observed on the plates for the *lacZ* targeting while no white colonies appeared on the EV control plates (Figure 5.2B). *lacZ* gene on bacterial genome was successfully targeted by type I-G system.

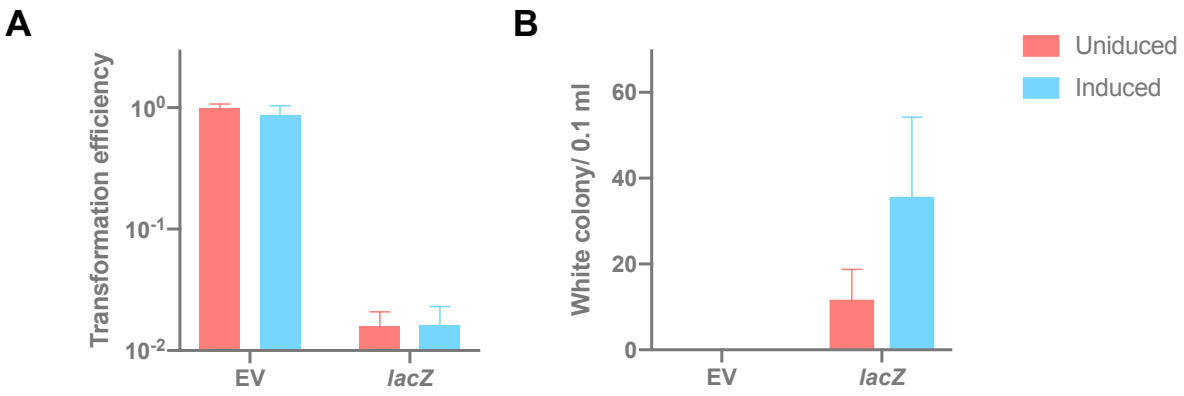


Figure 5.2. Type I-G editing on lacZ.

(A) (B) Transformation efficiency and white colony number out of transformants on the plate after transformation of the empty vector control or *lacZ* target spacer to wildtype Cas3 strain with L-arabinose induction (blue) or without induction (red); Transformation efficiency was calculated as the number of transformants divided by the number of transformants for original plasmid without target (Empty vector control); Values and error bars represent the mean of three biological replicates and standard deviation.

5.2.1.2 Type I-G genome targeting creating bi-directional long-range deletion

Since type I-G system targets *lacZ* on genome, we proceeded to investigate the targeting outcomes. A set of primers flanking the *lacZ* gene on the genome was designed to detect the deletion range of *lacZ* targeting (Figure 5.3A). 18 white colonies from genome targeting were submitted to tiling PCR with the set of primers from 55 kb downstream up to 20 kb upstream of *lacZ*. The 20 kb upstream region was intact across all white colonies since the essential gene *hemB* is adjacent. Colonies where *hemB* is deleted cannot survive. But a upstream deletion smaller than 20 kb could be observed in most of white colonies. On the other hand, the downstream region of *lacZ* gene was broadly deleted, and the deletion extended to at least 55 kb downstream (Figure 5.3B).

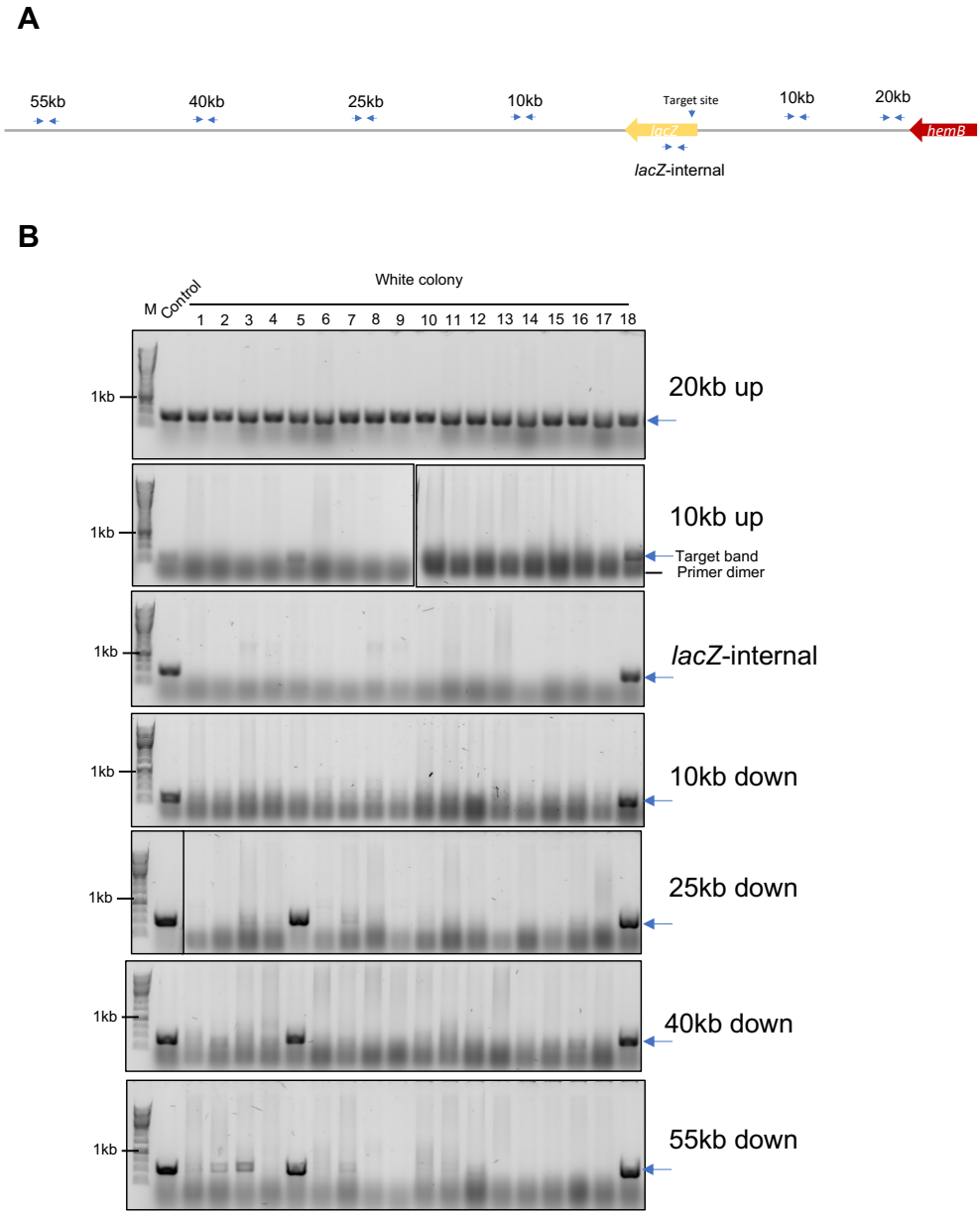


Figure 5.3. Tiling PCR of type I-G targeting *lacZ*.

(A) An overview of location of tiling PCR primers; A pair of small blue arrow represents a set of PCR primers. **(B)** Tiling PCR product from different primers was submitted for electrophoresis on a 0.8% agarose gel. 18 white colonies and 1 blue colony (Control) were assayed. M, marker. Blue arrows indicate positive PCR products.

As shown in Figure 5.4A, a deletion map of *lacZ* targeting by the type I-G effector was generated with tiling PCR results. Out of 18 tested white colonies, 17 survivors had a target *lacZ* deletion. 13 out of 17 had a deletion at least as far as 55 kb downstream of *lacZ*, 3 survivors yielded a deletion of 45 kb to 55 kb downstream, and 1 survivor experienced a shorter 10 kb to 25 kb downstream deletion. The *lacZ* gene of one white colony (WT-18) remained intact with no detected deletion around the target site, suggesting a spontaneous *lacZ* mutation. For upstream deletion, 16 colonies showed 10-20 kb deletions and 1 had a shorter (≤ 10 kb) deletion (Figure 5.4B). Overall, type I-G targeting cell genome generates bi-directional long-range deletion flanked the target site.

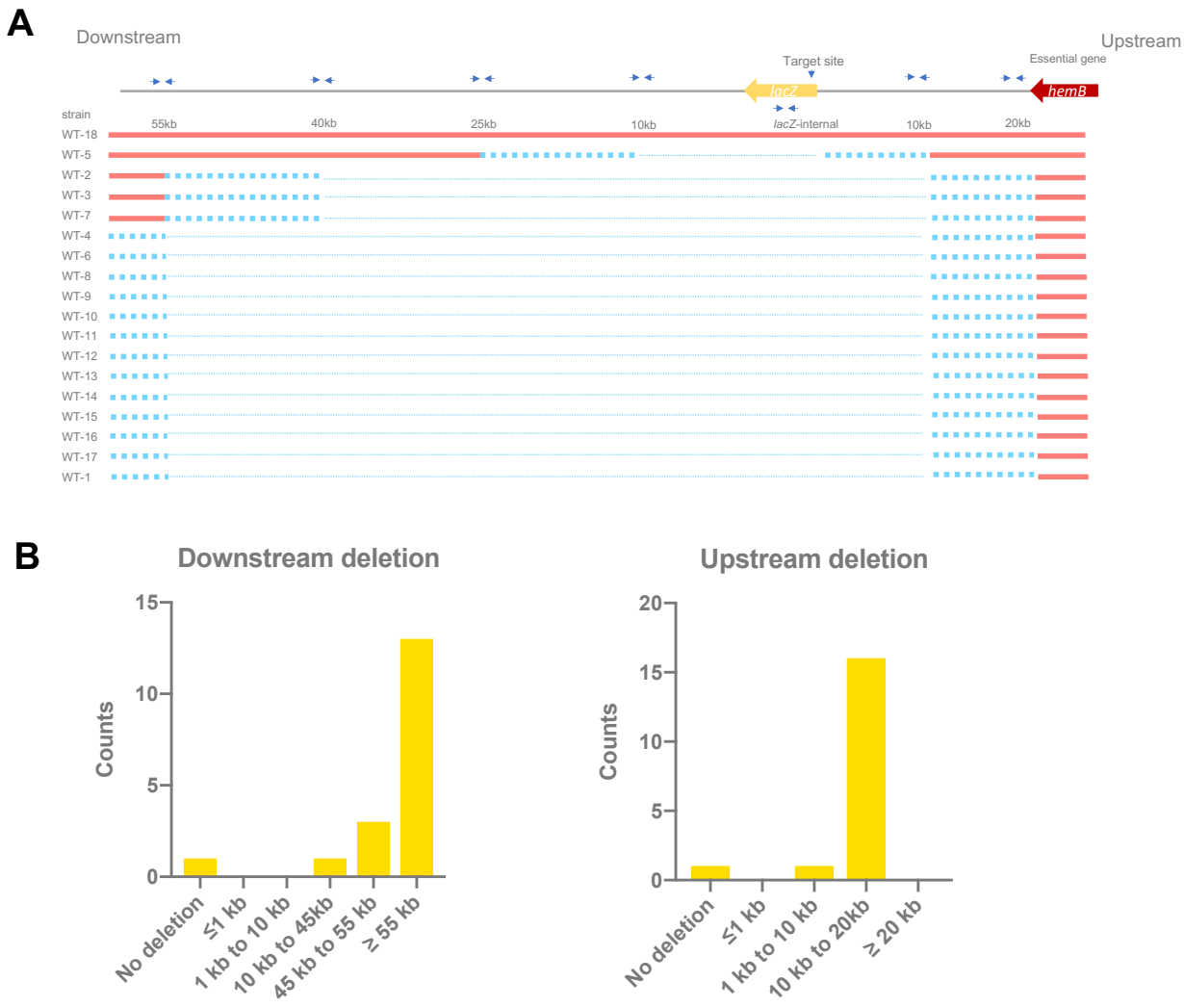


Figure 5.4. Deletion map of *lacZ*.

(A) A deletion map showing the outcome of *lacZ* targeting by type I-G; 18 white colonies generated by *lacZ* targeting were submitted for tiling PCR to determine the deletion range; Pairs of small blue arrows represent tiling PCR primers; Red lines indicate the intact sequence on genome; Blue dot lines indicates possible deleted sequence; Thin blue dash lines indicate confirmed deleted sequence. **(B)** Editing outcome of *lacZ* targeting by wildtype type I-G; 18 white colonies generated by *lacZ* targeting were submitted for tiling PCR to determine the deletion range; Counts of white colonies are plotted against deletion range.

To further reveal the deletion outcome, we proceeded to target *lacZ* adjacent genes. If the type I-G consistently yields the long-range bi-directional deletion on target site, white colonies can still be detected even *lacZ* is not the directed target. *yahK*, a gene located in 20 kb downstream of *lacZ*, and *frmA*, 13 kb upstream of *lacZ*, were chosen for targeting (Figure 5.5A). Cell survivability was consistent across different targets, all undergoing a significant cell loss (Figure 5.5B). Targeting *yahK* resulted in a comparable white colony number to the *lacZ* targeting. However, targeting *frmA* dramatically lowered the number (Figure 5.5C). The proximity to the essential gene *hemB* is probably causing the low number. The adjacent gene targeting confirms the deletion patten of type I-G.

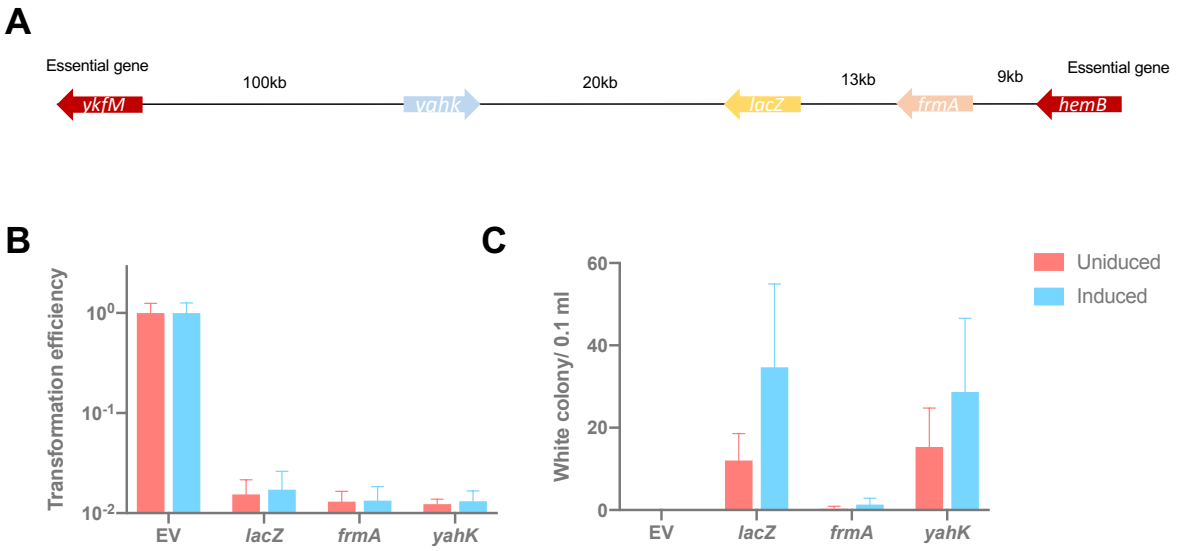


Figure 5.5. *yahK* and *frmA* target editing.

(A) *yahK* and *frmA* location on *E. coli* genome. **(B) (C)** Transformation efficiency and white colony number out of transformants on the plate after transformation of the empty vector, *lacZ* target spacer, *frmA* target spacer or *yahK* target spacer respectively with L-arabinose induction (blue) or without induction (red); Transformation efficiency was calculated as the number of transformants divided by the number of transformants for original plasmid without target (Empty vector control); Values and error bars represent the mean of three biological replicates and standard deviation.

5.2.2 The effect of type I-G Cas3 variants on genome targeting

The type I-G system generates bi-directional long-range deletions. The boundary of cleavage varies and is not controllable, therefore we sought to generate more controlled editing with the type I-G system. Altering the type I-G system to yield short range deletion is a potential approach to achieve controllable editing. Cas3, the key protein of type I-G target degradation, attracted our attention, as modifying Cas3 activity could change the pattern of type I-G genome degradation.

5.2.2.1 Functional changes of Cas3 variants

We first scrutinized Cas3 at the sequence level. Compared to the representative Cas3 from type I-E system in *Thermobifida fusca*, type I-G Cas3 comprises of the canonical HD nuclease domain and helicase domain (RecA family motor), but the domain organisation is different, the HD nuclease domain is located in the C-terminus instead of N-terminus (Figure 5.6). Based on structure prediction and sequence alignment, we designed Cas3 site directed mutations to modify Cas3 activity. Since Cas3 helicase domain unwinds target dsDNA and provides ssDNA for nuclease degradation, abolishing Cas3 helicase activity is highly likely to generate shorter deletions upon targeting. Subsequently, two sites in the helicase domain, K39 of Walker A motif and D164 of Walker B motif, were mutated to alanine to disrupt Cas3 helicase activity. Mutation on HD nuclease domain predicted to completely abort Cas3 cleavage activity, a mutation at the nuclease active site D625 was also constructed for investigation.

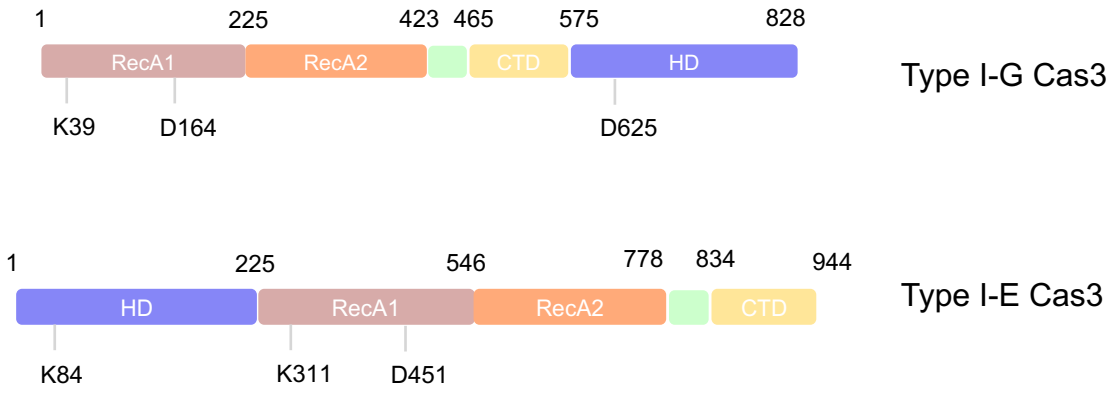


Figure 5.6. Comparison of type I-G and type I-E Cas3.

Domain organisation and active site residues of *Thermobifida fusca* (type I-E) and *T. sulfidiphilus* (type I-G) Cas3. Each has two RecA-family motor domains containing active site residues that define the Walker A and Walker B boxes of the helicase motor. The HD nuclease domain is present at the N-terminus of *T. fusca* Cas3 but at the C-terminus of *T. sulfidiphilus* Cas3. Type I-G Cas3 mutated residues were indicated, the corresponding residues in type I-E Cas3 also labelled.

We previously expressed and purified Cas3 helicase deficient variants (Cas3 K39A and D164A) and incorporated into Cascade for *in vitro* assay (Chapter 3, Figure 3.13). These two variants bind to target DNA but no longer degrade target DNA *in vitro*. We then introduced the two variants and the new helicase deficient variant (D625A) to *E. coli* to study the functional changes.

We first performed plasmid challenge assay on the Cas3 variants. The two helicase mutated variants K39A and D164A lost the ability to eradicate the invasive plasmid when uninduced (transformants), but the ability was recovered when expression of type I-G was fully induced (Figure 5.7A). A phage immunity assay revealed a

consistent result where only fully induced Cas3 K39A and D164A variants were capable of protecting cells from phage infection. Indeed, the D164A strain grew to a higher cell density than the wild-type strain across a range of MOIs, and the cell density of the K39A strain was even higher. On the other hand, these two variants showed compromised immunity when expression was uninduced compared to wild-type Cas3 (Figure 5.7B). The same plasmid challenge and phage immunity assay for wild-type Cas3 has been performed in Figure 3.15, 3.16 and 3.17.

We previously observed in an *in vitro* assay that in the absence of ATP (and thus helicase activity), target DNA was only cleaved at the site of Cas3 loading (Figure 3.14A). Together with *in vivo* data, it suggests that when Cas3 helicase activity was disrupted, it cannot translocate on DNA, but can still cut the target DNA. The result is that type I-G CRISPR defence is weakened but not abolished, and a high expression level of helicase deficient effector can block the plasmid and phage replication *in vivo*.

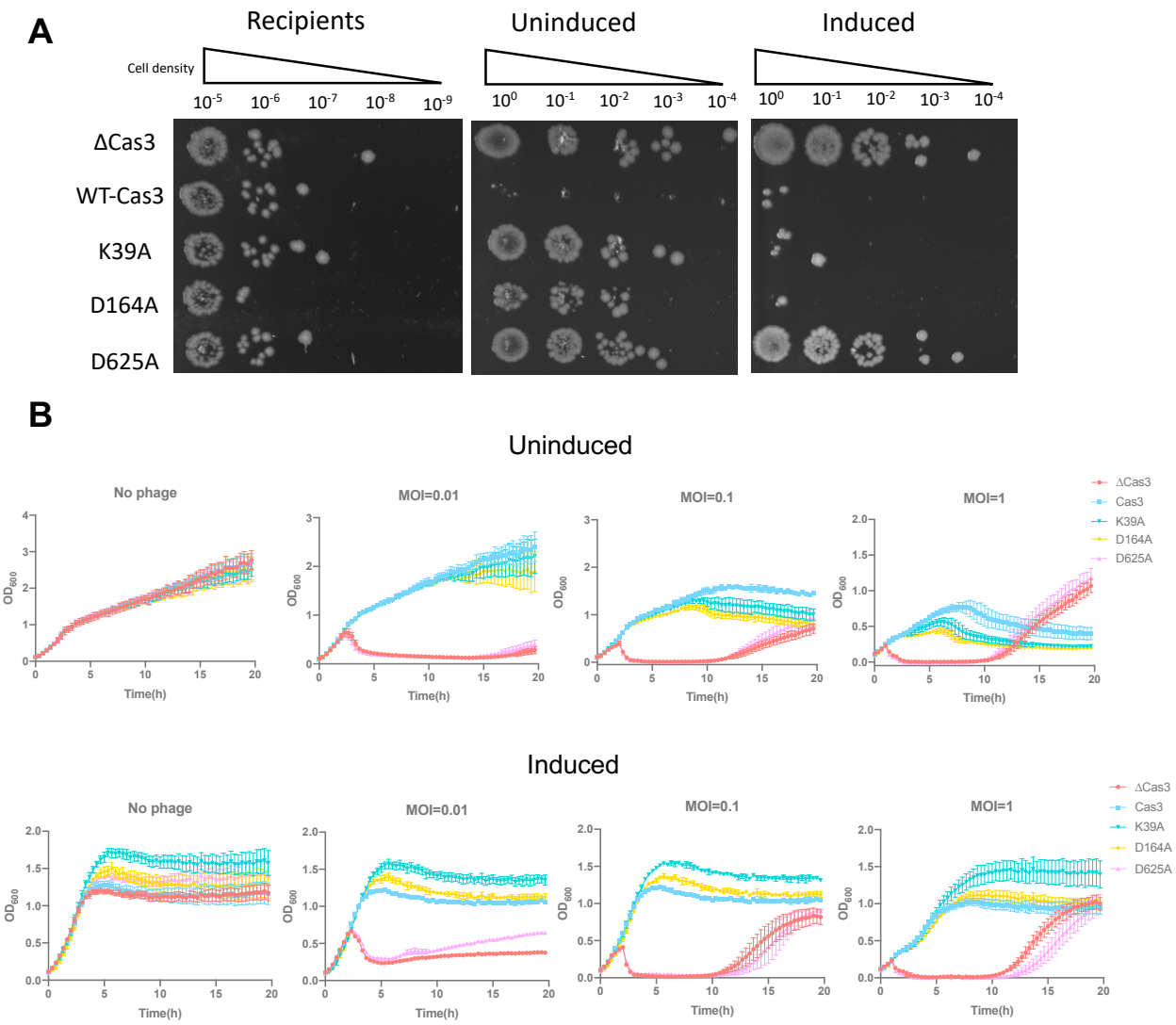


Figure 5.7. Plasmid and Phage P1 Challenge with wild-type and variant Cas3.

(A) Cells on the plates in different condition. Recipients, Ampicillin and Spectinomycin in plates; Uninduced, Ampicillin, Spectinomycin and Tetracycline in plates; Induced, with all three antibiotics and lactose, arabinose for induction of the type I-G CRISPR system. The Cas3 variants **Δ**Cas3, K39A, D164A and D625A were also studied. **(B)** Cell growth curve under phage P1 challenge (MOI=0, 0.01, 0.1 and 1). **Δ**Cas3 cells lack the *cas3* gene. Data points represent the mean of six experimental replicates (two biological replicates and three technical replicates) with standard deviation shown.

The phenotype of HD nuclease domain mutant D625A was indistinguishable from the ΔCas3 control (Figure 5.7A&B). It is reasonable to conclude that the disruption on Cas3 nuclease active site completely aborts type I-G interference. However, when we expressed and purified Cas3 D625A *in vitro*, we noticed that Cas3 D625A eluted from a size exclusion column as an aggregate (Figure 5.8). These data suggest that mutations disrupting active site of the Cas3 nuclease domain may disrupt protein folding or stability rather than just inactivating the nuclease domain.

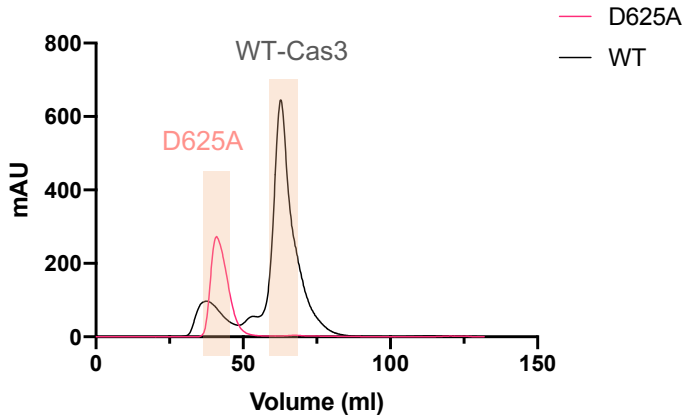


Figure 5.8. Cas3 D625A variant is aggregated.

Size Exclusion chromatography of WT Cas3 and Cas3 D625A. Cas3 D625A eluted right after the void volume, suggesting an aggregated state.

5.2.2.2 Cas3 helicase deficient variant altering genome editing outcome

Cas3 helicase deficient variants functionally target dsDNA with a compromised activity. Presumably, repurposing the variant for genome targeting would result in a different outcome from wild-type editing. *lacZ* gene targeting with Cas3 K39A variant was performed as for wild-type Cas3. The first thing we noticed is that cell survivability no longer decreased when the expression was uninduced, and when fully induced, cell

loss was still observed (Figure 5.9A). White colonies also appeared on induced plates at a low frequency (Figure 5.9B). Despite no significant change in numbers compared with wild-type editing, the introduction of Cas3 K39A provided an inducible way to activate type I-G genome editing.

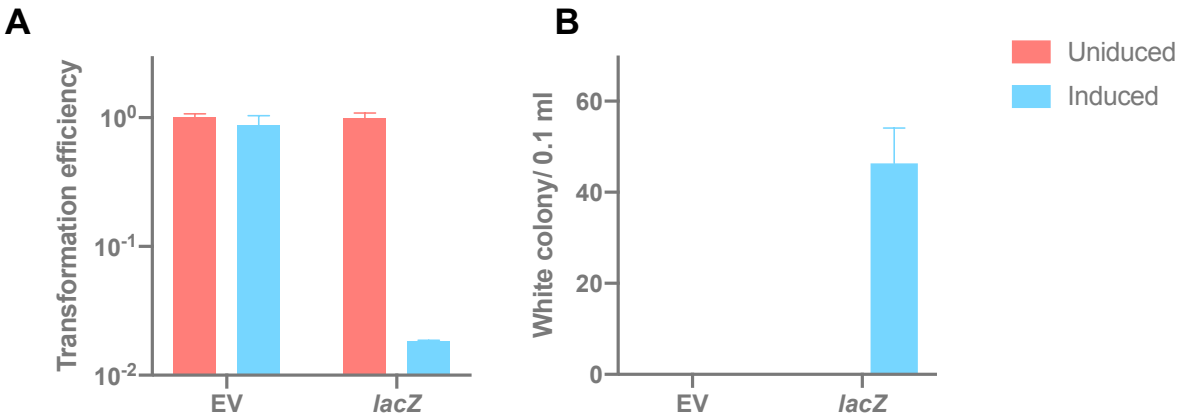


Figure 5.9. K39A Cas3 on *lacZ* targeting.

(A) (B) Transformation efficiency and white colony number out of transformants on the plate after transformation of the empty vector control or *lacZ* target spacer to K39A Cas3 strain with L-arabinose induction (blue) or without induction (red); Transformation efficiency was calculated as the number of transformants divided by the number of transformants for original plasmid without target (Empty vector control); Values and error bars represent the mean of three biological replicates and standard deviation.

We proceeded to examine the targeting outcomes of the K39A variant by tiling PCR. 18 white colonies were assayed with the primer set shown in Figure 5.10A. Stark differences from wild-type editing were observed. Out of 18 white colony survivors analysed by tiling PCR, 14 gave a PCR product when using an internal *lacZ* primer, 4 had a localised *lacZ* deletion, and only 1 of them showed long-range deletion to 25 kb downstream (Figure 5.10B). A pair of primers that cover the target area was then used to obtain further detail of deletions. Intriguingly, the products from target area PCR differed in size, indicating a small deletion (Figure 5.10C).

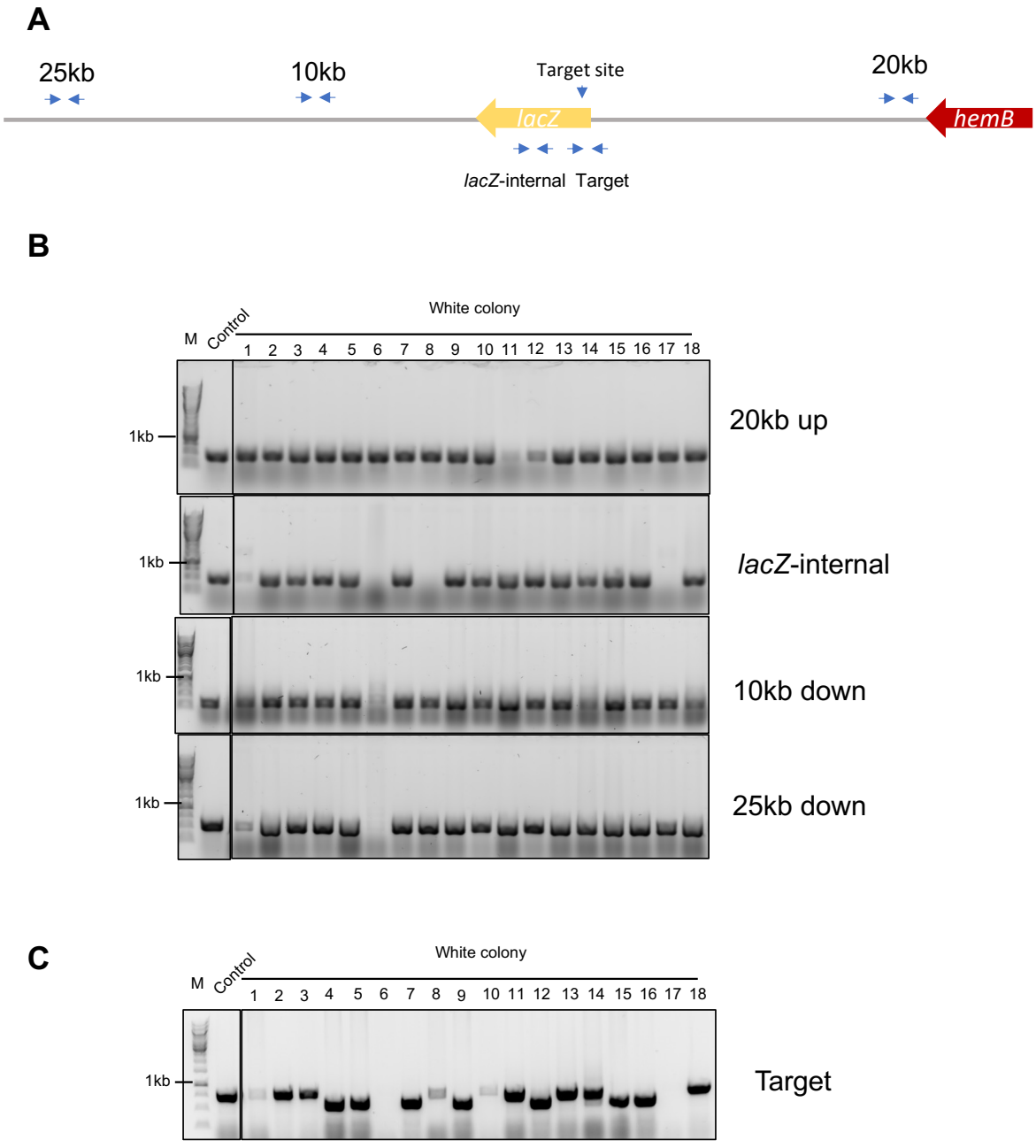


Figure 5.10. Tiling PCR of type I-G K39A targeting *lacZ*.

(A) An overview of location of tiling PCR primers; A pair of small blue arrows represents a set of PCR primers. **(B)** Tiling PCR products from different primers were submitted for electrophoresis on a 0.8% agarose gel. 18 white colonies and 1 blue colony (Control) were assayed. M, marker. **(C)** PCR product with primers that cover the target area was submitted for electrophoresis on a 0.8% agarose gel.

Visible bands on agarose gel from target area PCR (Figure 5.10C) were extracted and submitted for sequencing. A deletion map was generated based on tiling PCR and sequence data (Figure 5.11), showing a pattern of small deletion (Figure 5.11B).

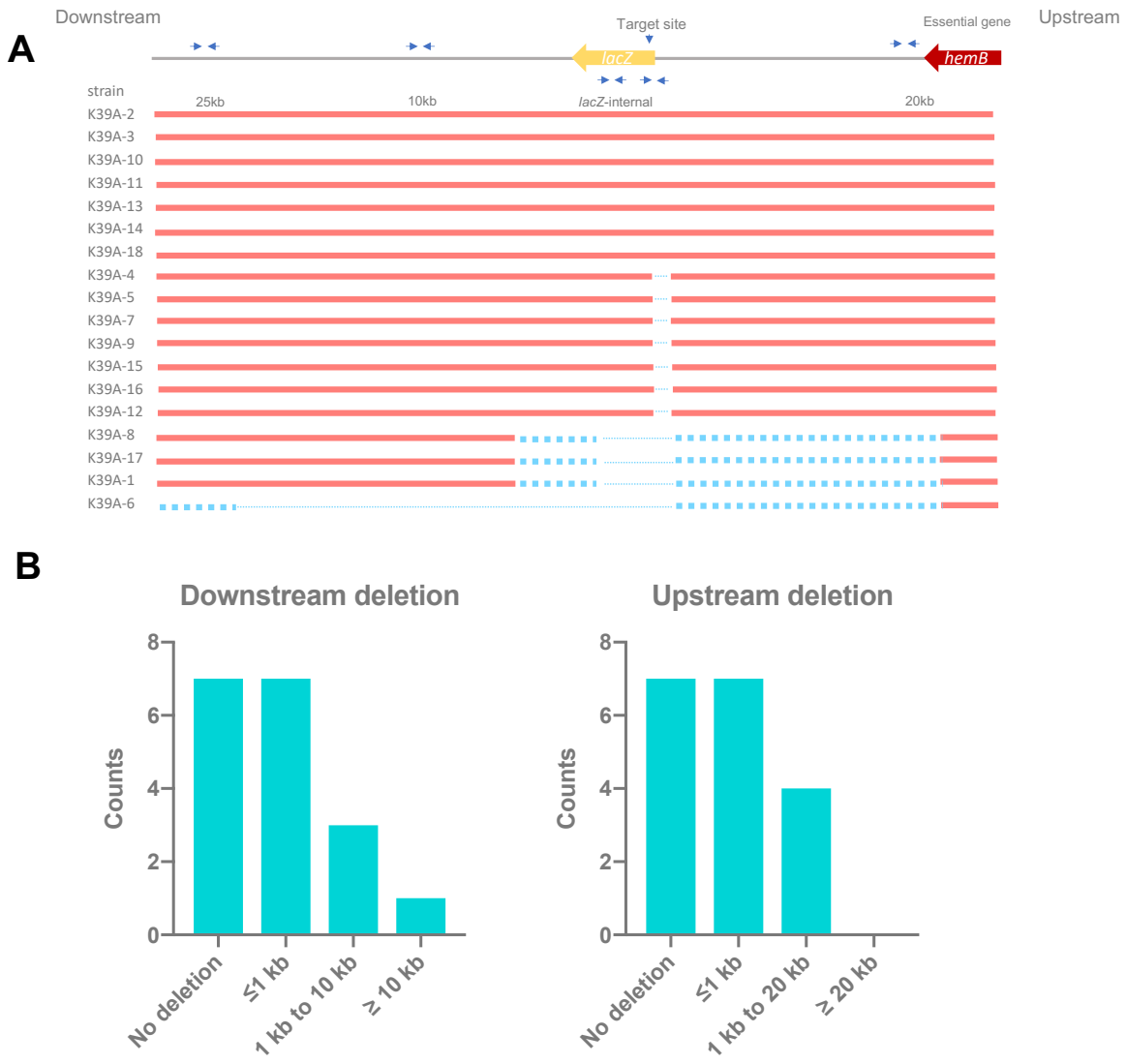


Figure 5.11. The deletion map of K39A Cas3 editing.

(A) A deletion map showing the outcome of *lacZ* targeting by Cas3 K39A mutated type I-G; 18 white colonies generated by *lacZ* targeting were submitted for tiling PCR to determine the deletion range; Pairs of small blue arrows represent tiling PCR primers; Red lines indicate the intact sequence on genome; Blue dot lines indicates possible deleted sequence; Thin blue dash lines indicate confirmed deleted sequence. **(B)** Editing outcome of *lacZ* targeting by Cas3 K39A mutated type I-G; 18 white colonies generated by *lacZ* targeting were submitted for tiling PCR to determine the deletion range; Counts of white colonies are plotted against deletion range.

5.2.2.3 Various deletion outcomes caused by microhomology

Genome targeting with Cas3 K39A variant generates small deletion. We sequenced 13 PCR products from target area amplification (Figure 5.10C). Out of 13 sequenced PCR bands, 1 contained a point mutation, 7 had a precise 180 bp deletion flanking the target site and the other 5 had no obvious edit (Figure 5.12). Investigation of the sequences revealed the presence of 11 bp homology area flanking the deleted region. This suggests that the DNA break introduced by Cas3 was repaired by limited strand resection and repair by microhomology-mediated end joining (MMEJ).

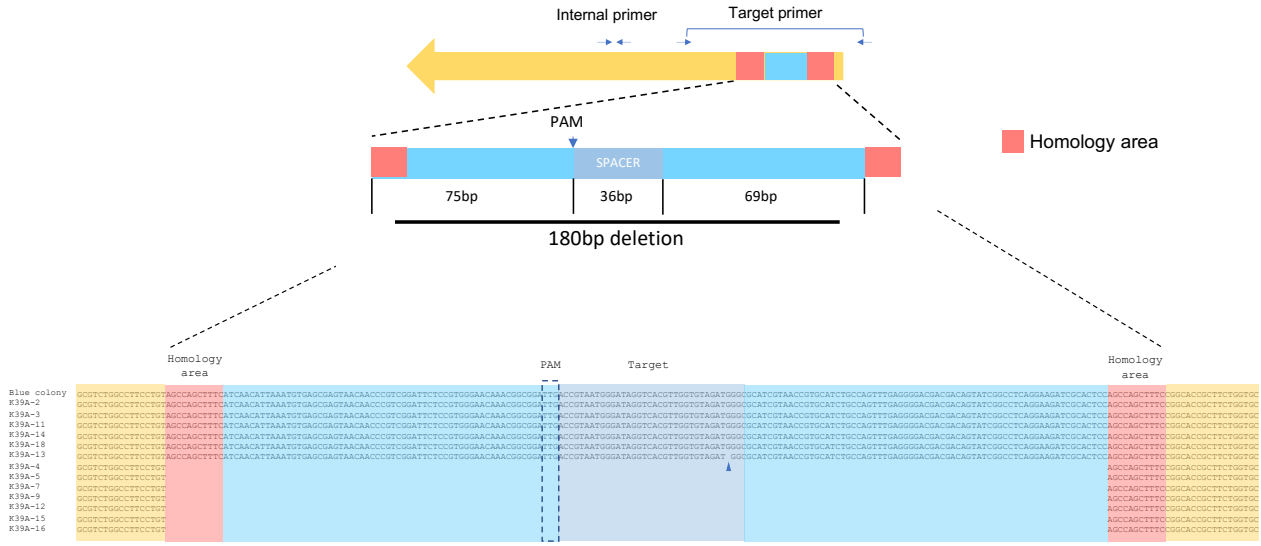


Figure 5.12. Type I-G editing on *lacZ*.

A schematic of 180 bp deletion by Cas3 K39A targeting *lacZ*; Blue, deleted sequence; Red, homology area; PAM, protospacer adjacent motif; SPACER, spacer sequence for *lacZ* targeting; Primers used for amplification of target area were indicated by arrows. Sequence analysis of the 180 bp deletion by Cascade with Cas3 K39A; Blue arrow, a point mutation was detected; Direct repeats in the flanking sequence are shown in red.

To further investigate this phenomenon, we targeted a second site in the *lacZ* gene using a different crRNA and analysed it as before. DNA sequencing showed various editing outcomes: intact target site (8 out of 17 tested colonies), point mutation (2/17), 24 bp short deletion (5/17), 324 bp long deletion (1/17) and a 11 bp insertion (1/17) (Figure 5.13). Microhomology areas ranging from 2 to 8 bp were apparent flanking the deleted regions. Overall, genome targeting via Cas3 variant K39A generated unique editing outcomes compared with the wild-type editing.

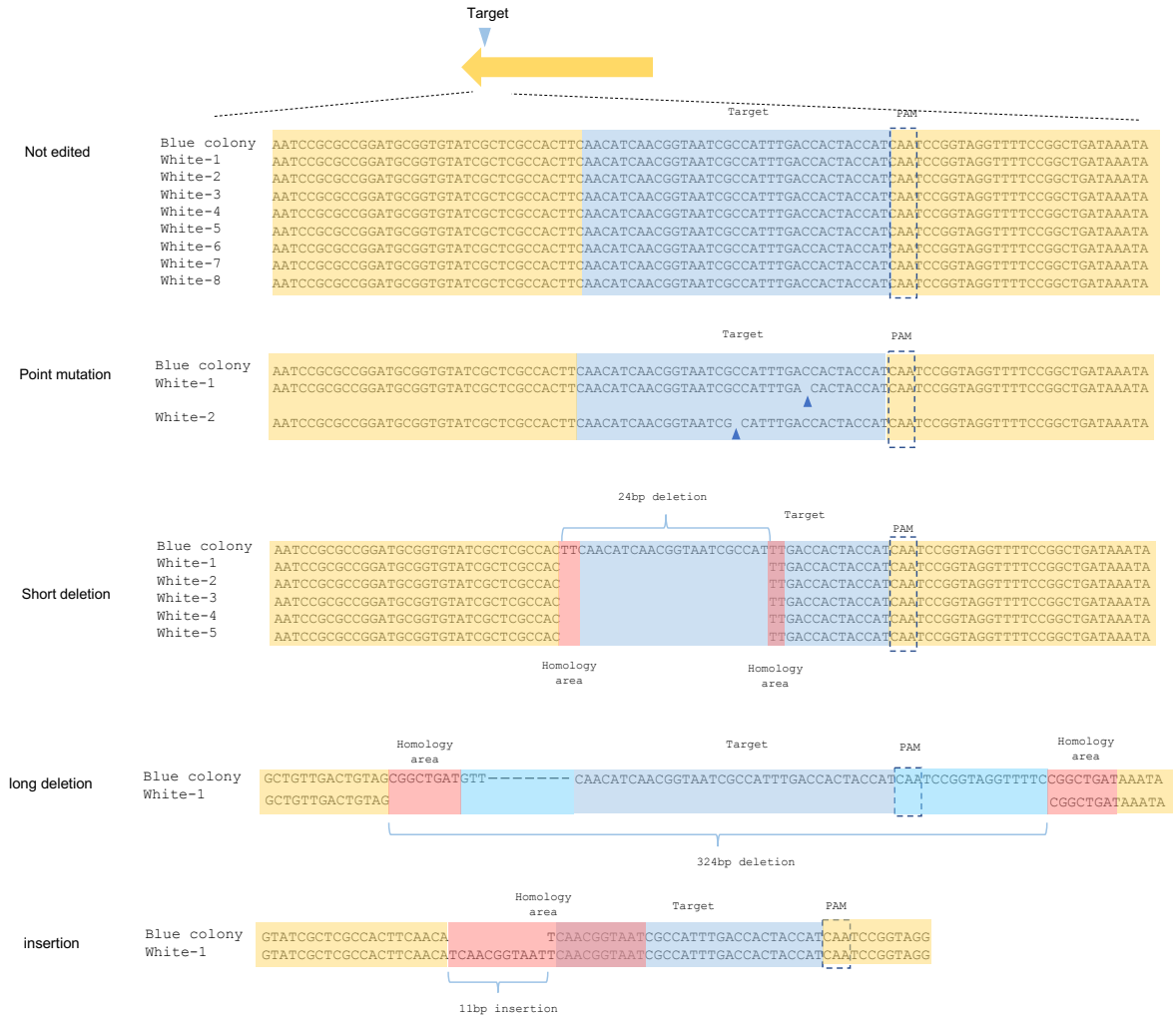


Figure 5.13. K39A target deletion on alternative *lacZ* site.

A different target site on *lacZ* gene produces various editing outcomes; Blue arrow, point mutation site. Homology in red.

5.2.3 Desirable genome editing by type I-G CRISPR

5.2.3.1 Introducing a HDR template increased cell survivability and editing efficiency

Type I-G genome editing generates DNA break that is repaired by either homologous directed repair (HDR) or MMEJ in bacteria. The HDR pathway is not effectively

activated if no templates are available for repair, and MMEJ becomes the main pathway to repair genome, as we observed in type I-G editing. MMEJ is an “error-prone” repair pathway and the outcome of repair is not controllable, mainly depending on the microhomology flanking the target area. We attempted to obtain desirable editing with type I-G system. To achieve this, HDR templates were introduced to activate the HDR pathway.

We first cloned two homologous arms as HDR templates into the pSPACER vector, generating the new vector pHR, which contains a CRISPR array and a donor HDR template. The two homologous arms were 600 bp in length, flanking the *lacZ* gene on the genome (Figure 5.14A). If HDR was successfully activated upon *lacZ* targeting, the region between the two homologous arms on the genome would be excised.

pHR was transformed into the *E. coli* MG1655 harbouring pM2, and blue-white screening was performed as before (Figure 5.14B). The introduction of HDR templates did not improve the cell survivability in wild-type type I-G targeting. However, more survivors were observed in Cas3 K39A variant targeting (Figure 5.14C). A higher number of white colonies were also apparent from blue-white screening (Figure 5.14C). The improvement of cell survivability and target editing efficiency in Cas3 K39A editing suggested that HDR was activated, resulting in shorter DNA deletions. But in wild-type type I-G editing, the profoundly damaged DNA might not be recovered by HDR, hence, no significant survivability change was observed.

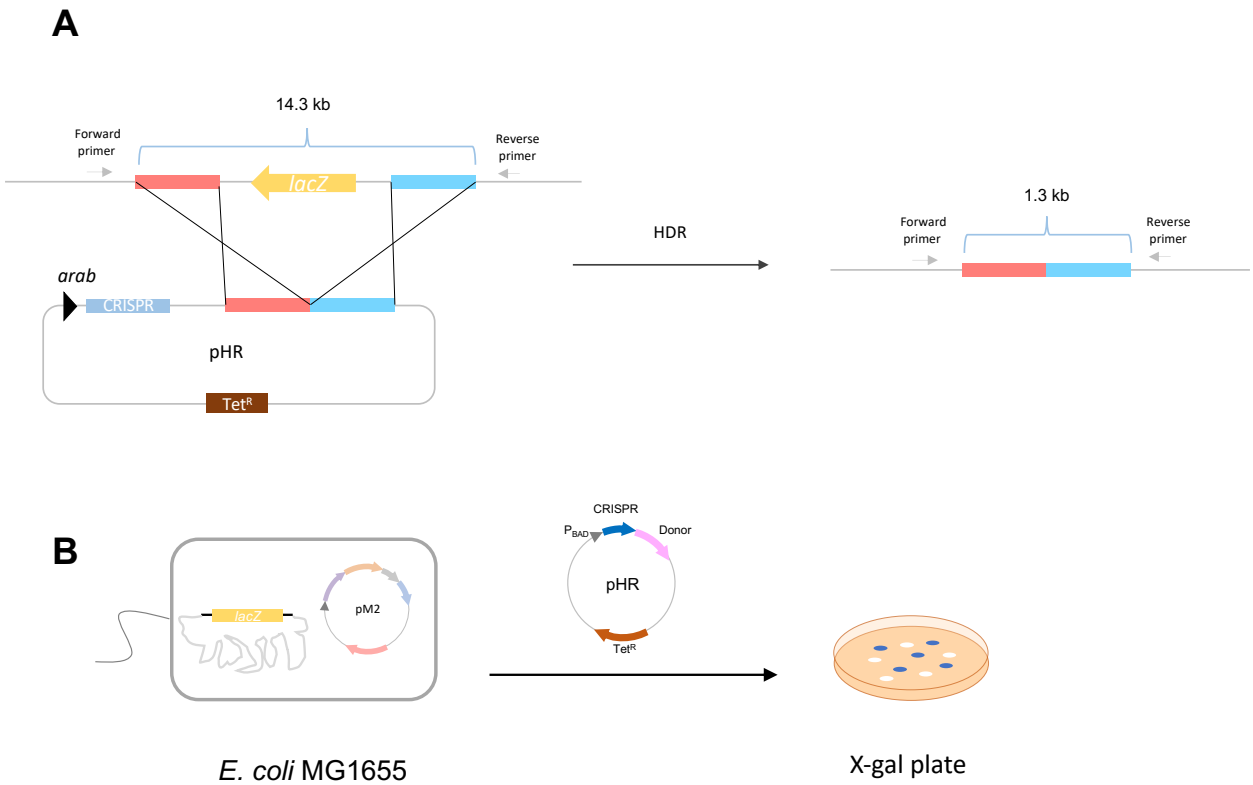


Figure legend next page

Figure 5.14. Homology-directed Repair.

(A) A schematic of homologous recombination; Homologous arms in red and blue; Primers for desired HDR verification was shown in grey arrow. **(B)** A schematic of homology directed repair; pM2 was transformed into *E. coli* MG1655 strain *(lacZ* intact) first, and pHR with donor sequence and *lacZ* target spacer was introduced to activate type I-G *lacZ* targeting and desired HDR, cells were spread on the X-gal plates for blue-white screening. **(C, D)** Transformation efficiency and white colony number on the plate after transformation of pHR into WT *cas3* strain or K39A *cas3* mutant strain with L-arabinose induction (blue) or without induction (red); Transformation efficiency was calculated as the number of transformants divided by the number of transformants for original plasmid without target (Empty vector control). Values and error bars represent the mean of three biological replicates and standard deviation; * $p<0.05$, paired T-test.

5.2.3.2 Desirable genome editing achieved by HDR

More white colonies were observed from blue-white screening of Cas3 K39A targeting, indicating more target editing on *lacZ*. We proceeded to examine the editing outcome of these white colony survivors.

We assayed white colonies by PCR with the primer set shown in Figure 5.14A. The primer set will amplify the *lacZ* region (covering the homologous arms) on genome. If HDR successfully repairs DNA damage with the donor HDR template, PCR amplification with the primer set will generate a 1.3 kb product (Figure 5.14A). 8 white colonies from wild-type type I-G targeting and 30 white colonies from Cas3 K39A targeting were tested, only 1 out 8 wild-type targeting events generated a 1.3 kb PCR product, while 28 out of 30 colonies from Cas3 K39A targeting gave the desired

product (Figure 5.15). Provision of HDR templates thus activates the HDR pathway and significantly increases the target editing efficiency.

Overall, the introduction of HDR templates in type I-G genome editing improves the editing efficiency and generates desirable editing, particularly when combined with the helicase deficient variant of Cas3.

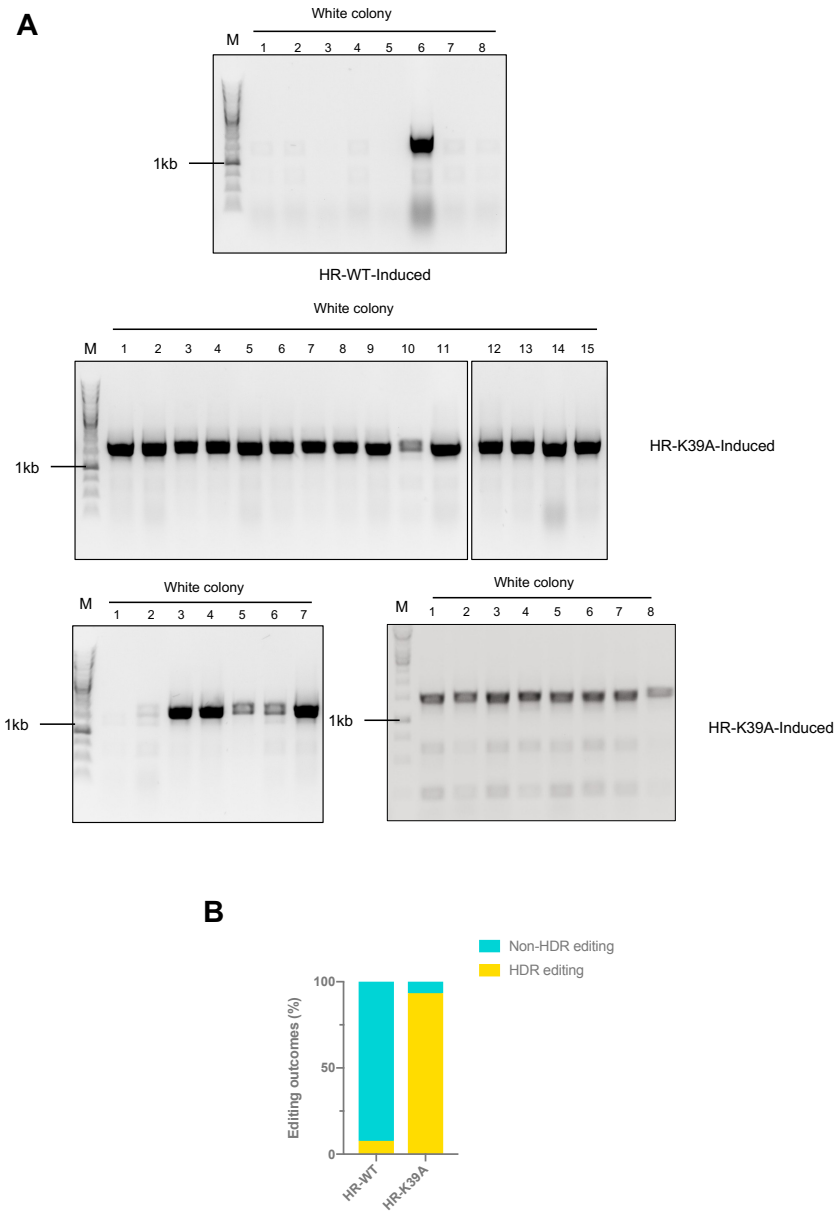


Figure 5.15 Desirable HDR editing.

(A) PCR product with verification primers were submitted for electrophoresis on a 0.8% agarose gel; 1.3kb product was detected, showing desired HDR with donor templates. **(B)** Editing outcomes with donor templates; Percentage of white colonies with desired HDR editing (yellow), without HDR editing (green); 30 individual white colonies from HR-K39A, 13 from HR-WT were assayed.

5.3 Discussion

As we discussed in chapter 3, type I-G Cas3 is pre-associated with Cascade rather than being recruited on target DNA binding. This pre-binding between Cas3 and Cascade has been shown in type I-A CRISPR where the authors proposed an allosteric activation mode, in contrast with the common trans-recruitment mode[193]. The type I-G Cas3 shares common features with canonical Cas3 (a SF2-helicase domain and a HD-nuclease domain), but at the protein sequence level, the HD nuclease domain is in C-terminus instead of N-terminus (Figure 5.6), while type I-A Cas3 is split to two individual genes, Cas3' (Helicase) followed by Cas3'' (HD nuclease)[86]. This difference in protein sequence of Cas3 may be one of the factors that leads to the emergence of two different Cascade-Cas3 activation modes.

Type I CRISPR degrades target dsDNA and generates either unidirectional[212-215, 217, 218] or bi-directional[193, 209, 216] deletion of the targetted genome. In our study, Type I-G Cascade-Cas3 specifically created bi-directional long-range deletion in *E. coli*. Previous research revealed Cas3 unidirectionally translocates on target DNA in vitro[103, 250]. Our data also shows type I-G Cascade degrades short dsDNA unidirectionally in vitro (Figure 3.14). Cas3 of type I-G probably only possesses the ability of unidirectional degradation, but in vivo degradation machinery might contribute to target deletion when DNA repair is suppressed in prokaryotes. Similar bidirectional degradation has been observed for type I-C CRISPR in prokaryotic organisms[209], but when type I-C from another organism was repurposed to apply in eukaryotes, it demonstrates unidirectional degradation[218], indicating that editing outcome depends on both the type of CRISPR system and the organisms being studied.

When we introduced the helicase deficient variant Cas3 K39A for genome editing, a 180 bp deletion or intact genome was observed on the target site (Figure 5.12). A further change of target site generated various editing outcomes that are attributed to “error-prone” repair (Figure 5.13). This alternative end joining repair, or microhomology mediated end joining, based on microhomology and had enhanced efficiency when HDR was abolished[251]. Meanwhile, providing a HDR template improved the cell survivability and produced desired editing by homologous recombination (Figure 5.14). Some white colonies still possessed an intact *lacZ* gene, suggesting interference inhibits *lacZ* expression without creating genome deletions. Recently, a Cas3 helicase variant of the *Zymomonas mobilis* type I-F system has been applied endogenously to carry out genome editing with high efficiency[252]. In that study, crRNA was used to target each strand of the target, resulting in “dual nicking” of the target gene. Our data suggest that targeting genes with a single guide RNA can still lead to efficient gene disruption and gene replacement, simplifying the procedure.

The comparison between Class I and Class II CRISPR in leveraging them for genome engineering is shown in Box 1. Class II CRISPR systems, especially Cas9, have been intensively studied and developed since the discovery of CRISPR due to its simplicity, however, the current genome engineering is still limited and cannot cover all the needs of modern biological research despite the large number of efforts being invested. Low editing efficiency, off-target effects and the difficulty of delivery have impeded the application of CRISPR systems. Novel editors, such as base editors and primer editors based on Cas9 (described in Chapter 1), have been developed to tackle those problems, but the challenges still exist. Discovery of novel editing systems plays a crucial role in conquering those challenges. The application of Class I systems has

not drawn much attention compared with Class II systems. It is a treasure vault for discovering novel gene editing tools.

Type I-G has the expertise in generating long-range deletion upon genome targeting. The multiple subunits have the potential to be modified to improve gene editing. The delivery of CRISPR systems is of importance to successful genome engineering, however, delivering multi-subunits is more challenging. Type I-G system is a relatively simple, 4-gene CRISPR system. It is more compact compared with other subtypes of type I systems, which could be an advantage in terms of delivery.

In conclusion, we repurposed type I-G for genome targeting in the heterologous organism *E. coli*. It opens the opportunity for further application of type I-G in heterologous prokaryotic systems or even eukaryotic systems.

Box 1 The comparison between Class I and Class II CRISPR in genome engineering

	Class 1 (Type I and Type III)	Class2 (Type II, V and VI)
Prevalence	Widely spread in prokaryotic organisms. Endogenous CRISPR can be repurposed for self-genome engineering.	Restrictedly spread in specific organisms. CRISPR can be introduced into heterologous organisms for prokaryotic genome editing, but compatibility could be an issue that hinders engineering.
PAM	A wide range of PAM sequence since a variety of Class I systems.	Restricted PAM sequence especially for Cas9. PAMless Cas can be a method to circumvent the PAM limitation, but the trade-off is the targeting specificity.
Targeting outcome	Type I systems generate long-range deletion on the genome. Type III targets specific RNA transcript and activates unspecific RNA degradation.	Type II (Cas9) and type V (Cas12) generate indels on genome targeting. Type VI (Cas13) targets RNA and triggers off-target cleavage of host RNAs.
Accessibility of delivery	Multiple Cas proteins, packaging and delivery of CRISPR to target organisms or tissues is challenging.	Single Cas protein has more flexibility of packaging and delivery than the multi-subunits systems.

Chapter 6: Conclusions and future direction

6.1 Conclusion

This thesis focuses on type I-G CRISPR system, the only subtype has not been explored in type I CRISPR family. We first expressed and purified Cas proteins of type I-G system from *Thioalkalivibrio sulfidiphilus*, investigating the expression stage of type I-G CRISPR. The unique Cas protein, Csb2, was shown to cleave pre-crRNA into mature crRNA, generating a canonical 5'-8 nt-handle and 3'-haipin, the cleavage site is 4 nt downstream of the base of hairpin instead of the exact base of hairpin, which is divergent from other type I systems. Further study revealed its two-domain organisation, a fusion of Cas5 and Cas6. Csb2 has a higher binding affinity for 3'-hairpin of mature crRNA than that for 5'-8 nt-handle. This is another major divergence from type I CRISPR, where Cas5 is constantly associated with the 5'- handle. The functional role of Cas5 domain in Csb2 requires further investigation.

We proceeded to investigate the interference stage of the type I-G system. By incubating Cas proteins with pre-crRNA, type I-G effector complex was successfully reconstituted *in vitro*. The Cascade of type I-G comprises of Csb2, Cas7 and large subunit Cas8g. In the process of constructing the effector complex, we found that Cas3 is pre-associated with Cascade rather than being recruited to Cascade upon target DNA binding. The pre-binding of Cas3 is of importance to the target DNA recognition. The reconstructed type I-G effector complex specifically targets and degrades DNA *in vitro*. Cas3 and Cascade is pre-associated before target DNA binding, but the recruitment of Cas3 upon dsDNA binding model may co-exists, which need further investigation.

Type I-G was then introduced to *E. coli* for *in vivo* reconstruction. Type I-G system expressed in the heterologous organism and eradicated invasive plasmid upon phage challenge assay. It also provides sufficient immunity against phage infection.

All together, the data in chapter 3 revealed the type I-G expression and interference stage on a biochemical level for the first time.

In chapter 4, we studied the structure of type I-G effector complex. The overview of type I-G Cascade is a crescent shape consisting of Cas7-crRNA backbone and Cas8g located in the belly of the backbone. Cas7-crRNA backbone shows a canonical organisation where each Cas7 occupies 6 nt of crRNA, and the crRNA is of a 5+1 pattern that one base is flipped out in the opposite direction to the other 5 bases.

Cas8g, the large subunit, however, is more complicated. There are no homologous hits in current protein database. By exploring the Alphafold predicted structure database, Cas8a2 from type I-A was shown to share structure similarity to the N-terminal domain of Cas8g. The large subunit in type I CRISPR systems is typically important for PAM recognition. The conformation of the N-terminal domain appears dynamic in type I-G, and the association with Cas3 probably stabilizes the conformation, allowing PAM recognition. The C-terminal domain of Cas8g, on the other hand, is rich in α-helix, a feature of the (absent) small subunit Cas11. As it is likely positioned in the centre of backbone, like Cas11 in other type I CRISPR, we concluded that small subunit of type I-G is fused to the large subunit Cas8g. Site directed mutations of conserved residues in Cas8g disrupt the architecture of type I-G effector complex, revealing two potential sites, RR270 and ND176, on Cas8g of Cas protein interaction and PAM recognition, but the more detailed mechanism of PAM

recognition will require a high-resolution structure of type I-G effector binding target dsDNA.

A schematic model is shown in Figure 6.1.

CRISPR systems can be repurposed for genome editing. We applied the type I-G system in bacterial genome editing in chapter 5. We first observed a significant decrease in cell number when redirecting type I-G system to target *E. coli* genome. Type I-G targeting genome DNA generates long-range bi-directional deletion on the genome, which cannot be efficiently repaired by bacteria, subsequently, lowering the cell survivability.

We attempted to reduce the toxicity and create desirable editing. By modifying Cas3, we observed a compromised interference in a helicase deficient Cas3 variant, Cas3 K39A. It was introduced for genome editing and altered the deletion outcome from long-range deletion to small deletion. Microhomology mediated end joining is the major pathway to repair the DNA break and hence generates various editing outcomes.

We further introduced HDR templates to activate HDR, combining with Cas3 K39A variant, the cell survivability and editing efficiency are improved and desirable genome editing is achieved by manipulating the HDR templates.

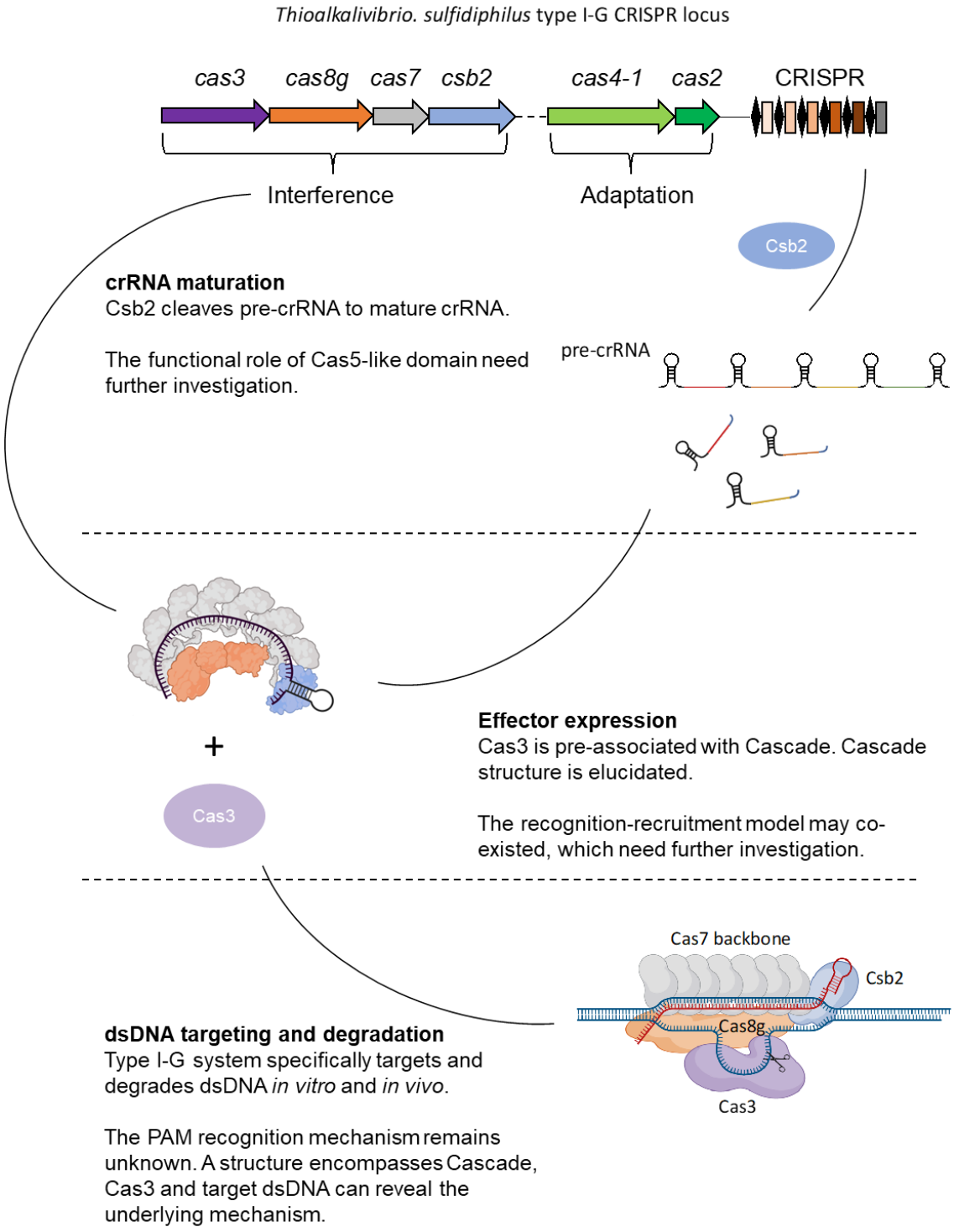


Figure 6.1 Type I-G expression and interference

A schematic figure showing the elucidated mechanism of type I-G in this thesis and the unsolved puzzles need to be addressed in the future study.

6.2 Future direction

We obtained the structure of type I-G Cascade (Cas7-crRNA backbone and Cas8g), but the PAM recognition mechanism remains unclear. Cas3 proved to be crucial for target DNA binding, and it is reasonable to assume Cas3 plays a role in PAM recognition. Obtaining the structure of Cas3 and target DNA with Cascade is the most direct way to elucidate the details of DNA targeting mechanism. One major future aim is to reconstruct the complete effector complex, including Cascade, Cas3 and target dsDNA, submitting for structure analysis. This work will improve our understanding of the type I-G CRISPR system, including the functional role of the Cas5-like domain in Csb2.

Understanding the interaction between subunits of type I-G is helpful for elucidating the type I-G mechanisms before obtaining a high-resolution structure. Protein cross-link following mass-spectrometry is a useful approach to investigate the interaction. Furthermore, protein pulldown assay could also dissect the underlying interaction.

To determine the co-existence of two Cas3 binding modesl, pre-associated or recognition-recruitment, Cas8g variant, R143E for instance, could be used to build Cascade, supplementing with Cas3 and submitting for DNA targeting assay.

Another direction focuses on the application of type I-G system. Type I-G has exhibited its genome editing ability in *E. coli*. Although repurposing endogenous CRISPR system for genome editing is the ideal approach, there are 60% of bacteria without CRISPR loci[253]. The type I-G system is a relatively compact 4 genes system, which might benefit its application in other prokaryotic systems. Obviously, type II CRISPR systems (Cas9), possesses simplicity in terms of repurposing for genome editing, but

as the most prevalent CRISPR system, type I system can cover a much broader range of PAM, and may gain a better fitness when introducing to other prokaryotic organisms.

Furthermore, type I-G can potentially be introduced to eukaryotic organism for gene editing. Compared with the well-established Cas9 editing, type I CRISPR requires the Cascade-Cas3 effector complex to achieve target editing. The more complicated effector means a more precise targeting. Off-target effect is suppressed since Cascade-Cas3 require a large conformational change, the sequence match is more strict compared to type II systems on targeting[214].

Type I CRISPR also holds advantages for long-range genome deletion. Applications such as exon skipping, complete removal of viral sequences or transposons, and efficient analysis of long non-coding regions can be achieved with less crRNA compared with type II editing. But editing efficiency and the way to deliver large effector is the main hurdle to apply type I-G system[254].

One novelty of applying type I systems is the potential to modify the multiple subunits to achieve various editing outcomes, as we did with the Cas3 protein. Modification on the helicase domain of Cas3 has generated different editing outcomes. It is promising to generate more modifications on type I system for gene engineering.

References

[1] Lehnherr, H., Maguin, E., Jafri, S., and Yarmolinsky, M. B. (1993) Plasmid addiction genes of bacteriophage P1: doc, which causes cell death on curing of prophage, and phd, which prevents host death when prophage is retained, *Journal of molecular biology 233*, 414-428.

[2] Takeuchi, O., and Akira, S. (2010) Pattern Recognition Receptors and Inflammation, *Cell 140*, 805-820.

[3] Demaria, O., Cornen, S., Daëron, M., Morel, Y., Medzhitov, R., and Vivier, E. (2019) Harnessing innate immunity in cancer therapy, *Nature 574*, 45-56.

[4] Bonilla, F. A., and Oettgen, H. C. (2010) Adaptive immunity, *Journal of Allergy and Clinical Immunology 125*, S33-S40.

[5] Altfeld, M., Fadda, L., Frleta, D., and Bhardwaj, N. (2011) DCs and NK cells: critical effectors in the immune response to HIV-1, *Nature Reviews Immunology 11*, 176-186.

[6] Dy, R. L., Richter, C., Salmond, G. P., and Fineran, P. C. (2014) Remarkable mechanisms in microbes to resist phage infections, *Annual review of virology 1*, 307-331.

[7] Rostøl, J. T., and Marraffini, L. (2019) (Ph) ighting phages: how bacteria resist their parasites, *Cell host & microbe 25*, 184-194.

[8] Bernheim, A., and Sorek, R. (2020) The pan-immune system of bacteria: antiviral defence as a community resource, *Nat Rev Microbiol 18*, 113-119.

[9] 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, *Nature microbiology 3*, 26-31.

[10] 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, *Journal of bacteriology 200*, e00792-00717.

[11] 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, *Nature microbiology 3*, 47-52.

[12] 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 microbiology letters 168*, 119-125.

[13] Clement, J.-M., Lepouce, E., Marchal, C., and Hofnung, M. (1983) Genetic study of a membrane protein: DNA sequence alterations due to 17 lamB point mutations affecting adsorption of phage lambda, *The EMBO journal 2*, 77-80.

[14] Ko, C. C., and Hatfull, G. F. (2018) Mycobacteriophage Fruitloop gp52 inactivates Wag31 (DivIVA) to prevent heterotypic superinfection, *Molecular microbiology 108*, 443-460.

[15] Tal, N., and Sorek, R. (2022) SnapShot: Bacterial immunity, *Cell 185*, 578-578.e571.

[16] Bertani, G., and Weigle, J. (1953) Host controlled variation in bacterial viruses, *Journal of bacteriology 65*, 113.

[17] Oliveira, P. H., Touchon, M., and Rocha, E. P. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts, *Nucleic acids research 42*, 10618-10631.

[18] Tock, M. R., and Dryden, D. T. (2005) The biology of restriction and anti-restriction, *Current opinion in microbiology 8*, 466-472.

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

[20] Wang, L., Chen, S., Xu, T., Taghizadeh, K., Wishnok, J. S., Zhou, X., You, D., Deng, Z., and Dedon, P. C. (2007) Phosphorothioation of DNA in bacteria by dnd genes, *Nature chemical biology 3*, 709-710.
[21] Ketting, R. F. (2011) The many faces of RNAi, *Developmental cell 20*, 148-161.
[22] Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells, *Nature 404*, 293-296.
[23] Shabalina, S. A., and Koonin, E. V. (2008) Origins and evolution of eukaryotic RNA interference, *Trends in ecology & evolution 23*, 578-587.
[24] 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, *Biology direct 4*, 29.
[25] Swarts, D. C., Jore, M. M., Westra, E. R., Zhu, Y., Janssen, J. H., Snijders, A. P., Wang, Y., Patel, D. J., Berenguer, J., and Brouns, S. J. (2014) DNA-guided DNA interference by a prokaryotic Argonaute, *Nature 507*, 258-261.
[26] Swarts, D. C., Szczepaniak, M., Sheng, G., Chandradoss, S. D., Zhu, Y., Timmers, E. M., Zhang, Y., Zhao, H., Lou, J., and Wang, Y. (2017) Autonomous generation and loading of DNA guides by bacterial Argonaute, *Molecular cell 65*, 985-998. e986.
[27] Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K., and Aravin, A. A. (2013) Bacterial argonaute samples the transcriptome to identify foreign DNA, *Molecular cell 51*, 594-605.
[28] Koopal, B., Potocnik, A., Mutte, S. K., Aparicio-Maldonado, C., Lindhoud, S., Vervoort, J. J. M., Brouns, S. J. J., and Swarts, D. C. (2022) Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA, *Cell 185*, 1471-1486.e1419.
[29] Duckworth, D., Glenn, J., and McCorquodale, D. (1981) Inhibition of bacteriophage replication by extrachromosomal genetic elements, *Microbiological Reviews 45*, 52-71.
[30] Lopatina, A., Tal, N., and Sorek, R. (2020) Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy, *Annual Review of Virology 7*, 371-384.
[31] Keen, E. C. (2015) A century of phage research: Bacteriophages and the shaping of modern biology, *BioEssays 37*, 6-9.
[32] 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, *Journal of Biological Chemistry 275*, 23219-23226.
[33] Kaufmann, G. (2000) Anticodon nucleases, *Trends in biochemical sciences 25*, 70-74.
[34] 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, *Journal of bacteriology 189*, 1417-1425.
[35] Harms, A., Brodersen, D. E., Mitarai, N., and Gerdes, K. (2018) Toxins, targets, and triggers: an overview of toxin-antitoxin biology, *Molecular cell 70*, 768-784.
[36] Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., and Salmond, G. P. (2009) The phage abortive infection system, ToxIN, functions as a protein–RNA toxin–antitoxin pair, *Proceedings of the National Academy of Sciences 106*, 894-899.
[37] Yee, T., Furuichi, T., Inouye, S., and Inouye, M. (1984) Multicopy single-stranded DNA isolated from a gram-negative bacterium, Myxococcus xanthus, *Cell 38*, 203-209.
[38] Simon, A. J., Ellington, A. D., and Finkelstein, I. J. (2019) Retrons and their applications in genome engineering, *Nucleic acids research 47*, 11007-11019.
[39] Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., and Sorek, R. (2020) Bacterial Retrons Function In Anti-Phage Defense, *Cell 183*, 1551-1561.e1512.

[40] Johnson, A. G., Wein, T., Mayer, M. L., Duncan-Lowey, B., Yirmiya, E., Oppenheimer-Shaanan, Y., Amitai, G., Sorek, R., and Kranzusch, P. J. (2022) Bacterial gasdermins reveal an ancient mechanism of cell death, *Science 375*, 221-225.
[41] Kayagaki, N., Stowe, I. B., Lee, B. L., O'Rourke, K., Anderson, K., Warming, S., Cuellar, T., Haley, B., Roose-Girma, M., Phung, Q. T., Liu, P. S., Lill, J. R., Li, H., Wu, J., Kummerfeld, S., Zhang, J., Lee, W. P., Snipas, S. J., Salvesen, G. S., Morris, L. X., Fitzgerald, L., Zhang, Y., Bertram, E. M., Goodnow, C. C., and Dixit, V. M. (2015) Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling, *Nature 526*, 666-671.
[42] Shi, J., Zhao, Y., Wang, K., Shi, X., Wang, Y., Huang, H., Zhuang, Y., Cai, T., Wang, F., and Shao, F. (2015) Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death, *Nature 526*, 660-665.
[43] Ding, J., Wang, K., Liu, W., She, Y., Sun, Q., Shi, J., Sun, H., Wang, D.-C., and Shao, F. (2016) Pore-forming activity and structural autoinhibition of the gasdermin family, *Nature 535*, 111-116.
[44] Whiteley, A. T., Eaglesham, J. B., de Oliveira Mann, C. C., Morehouse, B. R., Lowey, B., Nieminen, E. A., Danilchanka, O., King, D. S., Lee, A. S. Y., Mekalanos, J. J., and Kranzusch, P. J. (2019) Bacterial cGAS-like enzymes synthesize diverse nucleotide signals, *Nature 567*, 194-199.
[45] Cohen, D., Melamed, S., Millman, A., Shulman, G., Oppenheimer-Shaanan, Y., Kacen, A., Doron, S., Amitai, G., and Sorek, R. (2019) Cyclic GMP–AMP signalling protects bacteria against viral infection, *Nature 574*, 691-695.
[46] Tal, N., Morehouse, B. R., Millman, A., Stokar-Avihail, A., Avraham, C., Fedorenko, T., Yirmiya, E., Herbst, E., Brandis, A., Mehlman, T., Oppenheimer-Shaanan, Y., Keszei, A. F. A., Shao, S., Amitai, G., Kranzusch, P. J., and Sorek, R. (2021) Cyclic CMP and cyclic UMP mediate bacterial immunity against phages, *Cell 184*, 5728-5739.e5716.
[47] Ofir, G., Herbst, E., Baroz, M., Cohen, D., Millman, A., Doron, S., Tal, N., Malheiro, D. B. A., Malitsky, S., Amitai, G., and Sorek, R. (2021) Antiviral activity of bacterial TIR domains via immune signalling molecules, *Nature 600*, 116-120.
[48] Ablasser, A., Goldeck, M., Cavlar, T., Deimling, T., Witte, G., Röhl, I., Hopfner, K.-P., Ludwig, J., and Hornung, V. (2013) cGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING, *Nature 498*, 380-384.
[49] Hogrel, G., Guild, A., Graham, S., Rickman, H., Grüschow, S., Bertrand, Q., Spagnolo, L., and White, M. F. (2022) Cyclic nucleotide-induced helical structure activates a TIR immune effector, *Nature 608*, 808-812.
[50] Lau, R. K., Ye, Q., Birkholz, E. A., Berg, K. R., Patel, L., Mathews, I. T., Watrous, J. D., Ego, K., Whiteley, A. T., and Lowey, B. (2020) Structure and mechanism of a cyclic trinucleotide-activated bacterial endonuclease mediating bacteriophage immunity, *Molecular cell 77*, 723-733. e726.
[51] Fitzgerald, K. A., and Kagan, J. C. (2020) Toll-like receptors and the control of immunity, *Cell 180*, 1044-1066.
[52] Bernheim, A., Millman, A., Ofir, G., Meitav, G., Avraham, C., Shomar, H., Rosenberg, M. M., Tal, N., Melamed, S., Amitai, G., and Sorek, R. (2021) Prokaryotic viperins produce diverse antiviral molecules, *Nature 589*, 120-124.
[53] Gizzi, A. S., Grove, T. L., Arnold, J. J., Jose, J., Jangra, R. K., Garforth, S. J., Du, Q., Cahill, S. M., Dulyaninova, N. G., Love, J. D., Chandran, K., Bresnick, A. R., Cameron, C. E., and Almo, S. C. (2018) A naturally occurring antiviral ribonucleotide encoded by the human genome, *Nature 558*, 610-614.

[54] Kronheim, S., Daniel-Ivad, M., Duan, Z., Hwang, S., Wong, A. I., Mantel, I., Nodwell, J. R., and Maxwell, K. L. (2018) A chemical defence against phage infection, *Nature 564*, 283-286.
[55] Kever, L., Hardy, A., Luthe, T., Hünnefeld, M., Gätgens, C., Milke, L., Wiechert, J., Wittmann, J., Moraru, C., Marienhagen, J., and Frunzke, J. (2022) Aminoglycoside Antibiotics Inhibit Phage Infection by Blocking an Early Step of the Infection Cycle, *mBio 13*, e00783-00722.
[56] Tal, N., Millman, A., Stokar-Avihail, A., Fedorenko, T., Leavitt, A., Melamed, S., Yirmiya, E., Avraham, C., Brandis, A., Mehlman, T., Amitai, G., and Sorek, R. (2022) Bacteria deplete deoxynucleotides to defend against bacteriophage infection, *Nature Microbiology 7*, 1200-1209.
[57] 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.
[58] Makarova, K. S., Wolf, Y. I., Snir, S., and Koonin, E. V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems, *Journal of bacteriology 193*, 6039-6056.
[59] Koonin, E. V., Makarova, K. S., and Wolf, Y. I. (2017) Evolutionary genomics of defense systems in archaea and bacteria, *Annual review of microbiology 71*, 233-261.
[60] 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.
[61] Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C., and Laub, M. T. (2022) A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome, *Nature Microbiology 7*, 1568-1579.
[62] Pfeifer, E., Sousa, J. M., Touchon, M., and Rocha, E. P. C. (2022) When bacteria are phage playgrounds: interactions between viruses, cells, and mobile genetic elements, *Current Opinion in Microbiology 70*, 102230.
[63] Penadés, J. R., and Christie, G. E. (2015) The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites, *Annual Review of Virology 2*, 181-201.
[64] Tormo-Más, M. Á., Mir, I., Shrestha, A., Tallent, S. M., Campoy, S., Lasa, Í., Barbé, J., Novick, R. P., Christie, G. E., and Penadés, J. R. (2010) Moonlighting bacteriophage proteins derepress staphylococcal pathogenicity islands, *Nature 465*, 779-782.
[65] Fillol-Salom, A., Rostøl, J. T., Ojiogu, A. D., Chen, J., Douce, G., Humphrey, S., and Penadés, J. R. (2022) Bacteriophages benefit from mobilizing pathogenicity islands encoding immune systems against competitors, *Cell 185*, 3248-3262.e3220.
[66] Borges, A. L., Davidson, A. R., and Bondy-Denomy, J. (2017) The discovery, mechanisms, and evolutionary impact of anti-CRISPRs, *Annual review of virology 4*, 37-59.
[67] Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K. L., and Davidson, A. R. (2015) Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins, *Nature 526*, 136-139.
[68] Samuel, B., and Burstein, D. (2023) A diverse repertoire of anti-defense systems is encoded in the leading region of plasmids, *bioRxiv*, 2023.2002. 2015.528439.
[69] Martínez Arbas, S., Narayanasamy, S., Herold, M., Lebrun, L. A., Hoopmann, M. R., Li, S., Lam, T. J., Kunath, B. J., Hicks, N. D., and Liu, C. M. (2021) Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics, *Nature microbiology 6*, 123-135.
[70] Srikant, S., Guegler, C. K., and Laub, M. T. (2022) The evolution of a counter-defense mechanism in a virus constrains its host range, *eLife 11*, e79549.

[71] Williams, M. C., Reker, A. E., Margolis, S. R., Liao, J., Wiedmann, M., Rojas, E. R., and Meeske, A. J. (2023) Restriction endonuclease cleavage of phage DNA enables resuscitation from Cas13-induced bacterial dormancy, *Nature Microbiology*.

[72] Brockhurst, M. A., Koskella, B., and Zhang, Q.-G. (2021) Bacteria-phage antagonistic coevolution and the implications for phage therapy, *Bacteriophages: biology, technology, therapy*, 231-251.

[73] Mojica, F. J., Juez, G., and Rodriguez‐Valera, F. (1993) Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites, *Molecular microbiology 9*, 613-621.

[74] 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, *Journal of bacteriology 169*, 5429-5433.

[75] Mojica, F., Ferrer, C., Juez, G., and Rodríguez‐Valera, F. (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning, *Molecular microbiology 17*, 85-93.

[76] Mojica, F. J., Díez‐Villaseñor, C., Soria, E., and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria, *Molecular microbiology 36*, 244-246.

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

[78] Mojica, F. J., and Garrett, R. A. (2013) Discovery and seminal developments in the CRISPR field, In *CRISPR-Cas Systems*, pp 1-31, Springer.

[79] Mojica, F. J., García-Martínez, J., and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements, *Journal of molecular evolution 60*, 174-182.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

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

[85] Marraffini, L. A. (2015) CRISPR-Cas immunity in prokaryotes, *Nature 526*, 55-61.

[86] 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., Horvath, P., Moineau, S., Mojica, F. J., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R., and Koonin, E. V. (2015) An updated evolutionary classification of CRISPR-Cas systems, *Nat Rev Microbiol 13*, 722-736.

[87] Jansen, R., Embden, J. D. A. v., Gaastra, W., and Schouls, L. M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes, *Molecular Microbiology 43*, 1565-1575.
[88] Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., and Koonin, E. V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis, *Nucleic acids research 30*, 482-496.
[89] Haft, D. H., Selengut, J., Mongodin, E. F., and Nelson, K. E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, *PLoS computational biology 1*.
[90] 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., van der Oost, J., and Koonin, E. V. (2011) Evolution and classification of the CRISPR-Cas systems, *Nat Rev Microbiol 9*, 467-477.
[91] 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., Moineau, S., Mojica, F. J. M., Scott, D., Shah, S. A., Siksnys, V., Terns, M. P., Venclovas, Č., White, M. F., Yakunin, A. F., Yan, W., Zhang, F., Garrett, R. A., Backofen, R., van der Oost, J., Barrangou, R., and Koonin, E. V. (2020) Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants, *Nature Reviews Microbiology 18*, 67-83.
[92] Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2022) Evolutionary Classification of CRISPR-Cas Systems, In *Crispr*, pp 13-38.
[93] 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.
[94] 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.
[95] Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system, *Microbiology 155*, 733-740.
[96] 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.
[97] 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.
[98] Lee, H., Zhou, Y., Taylor, D. W., and Sashital, D. G. (2018) Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays, *Molecular cell 70*, 48-59. e45.
[99] 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, *Molecular cell 70*, 814-824. e816.
[100] 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.
[101] Fineran, P. C., Gerritzen, M. J., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S. A., Staals, R. H., and Brouns, S. J. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation, *Proceedings of the National Academy of Sciences 111*, E1629-E1638.
[102] 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 IF CRISPR-Cas

system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer, *Nucleic acids research 42*, 8516-8526.
[103] Redding, S., Sternberg, Samuel H., Marshall, M., Gibb, B., Bhat, P., Guegler, Chantal K., Wiedenheft, B., Doudna, Jennifer A., and Greene, Eric C. (2015) Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System, *Cell 163*, 854-865.
[104] 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., Ke, A., and Finkelstein, I. J. (2018) Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex, *Cell 175*, 934-946.e915.
[105] 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. e516.
[106] McGinn, J., and Marraffini, L. A. (2019) Molecular mechanisms of CRISPR–Cas spacer acquisition, *Nature Reviews Microbiology 17*, 7-12.
[107] 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.
[108] Xiao, Y., Ng, S., Nam, K. H., and Ke, A. (2017) How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration, *Nature 550*, 137-141.
[109] Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J. 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.
[110] Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M., and Terns, M. P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex, *Cell 139*, 945-956.
[111] 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, *Journal of bacteriology 196*, 310-317.
[112] Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., and Albers, S.-V. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity, *Molecular cell 45*, 303-313.
[113] 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.
[114] Sokolowski, R. D., Graham, S., and White, M. F. (2014) Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system, *Nucleic acids research 42*, 6532-6541.
[115] 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, *Proceedings of the National Academy of Sciences 108*, 21218-21222.
[116] Hatoum-Aslan, A., Samai, P., Maniv, I., Jiang, W., and Marraffini, L. A. (2013) A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs, *Journal of Biological Chemistry 288*, 27888-27897.
[117] Walker, F. C., Chou-Zheng, L., Dunkle, J. A., and Hatoum-Aslan, A. (2017) Molecular determinants for CRISPR RNA maturation in the Cas10–Csm complex and roles for non-Cas nucleases, *Nucleic acids research 45*, 2112-2123.
[118] Özcan, A., Pausch, P., Linden, A., Wulf, A., Schühle, 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, *Nature microbiology 4*, 89-96.

[119] Makarova, K. S., Aravind, L., Wolf, Y. I., and Koonin, E. V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems, *Biology direct 6*, 38.
[120] van der Oost, J. (2022) Molecular Mechanisms of Type I CRISPR -Cas Systems, In *Crispr*, pp 39-52.
[121] Marraffini, L. A. (2022) Mechanism of Type III CRISPR - Cas Immunity, In *Crispr*, pp 71-84.
[122] 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.
[123] Chylinski, K., Le Rhun, A., and Charpentier, E. (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems, *RNA biology 10*, 726-737.
[124] Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus, *RNA biology 10*, 841-851.
[125] Karvelis, T., and Siksnys, V. (2022) Molecular Mechanisms of Type II CRISPR - Cas Systems, In *Crispr*, pp 53-69.
[126] 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., and Regev, A. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system, *Cell 163*, 759-771.
[127] Yan, W. X., Hunnewell, P., Alfonse, L. E., Carte, J. M., Keston-Smith, E., Sothiselvam, S., Garrity, A. J., Chong, S., Makarova, K. S., and Koonin, E. V. (2019) Functionally diverse type V CRISPR-Cas systems, *Science 363*, 88-91.
[128] Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., and Charpentier, E. (2016) The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA, *Nature 532*, 517-521.
[129] Harrington, L. B., Ma, E., Chen, J. S., Witte, I. P., Gertz, D., Paez-Espino, D., Al-Shayeb, B., Kyrpides, N. C., Burstein, D., and Banfield, J. F. (2020) A scoutRNA is required for some type V CRISPR-Cas systems, *Molecular cell 79*, 416-424. e415.
[130] Beckett, M. Q., Ramachandran, A., and Bailey, S. (2022) Type V CRISPR-Cas Systems, In *Crispr*, pp 85-97.
[131] 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.
[132] Abudayyeh, O. O., and Gootenberg, J. S. (2022) CRISPR-Cas13: Biology, Mechanism, and Applications of RNA-Guided, RNA-Targeting CRISPR Systems, In *Crispr*, pp 99-120.
[133] Xiao, Y., Luo, M., Hayes, R. P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017) Structure basis for directional R-loop formation and substrate handover mechanisms in type I CRISPR-Cas system, *Cell 170*, 48-60. e11.
[134] 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, *Proceedings of the National Academy of Sciences 108*, 10098-10103.
[135] Wiedenheft, B., van Duijn, E., Bultema, J. B., Waghmare, S. P., Zhou, K., Barendregt, A., Westphal, W., Heck, A. J., Boekema, E. J., and Dickman, M. J. (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions, *Proceedings of the National Academy of Sciences 108*, 10092-10097.

[136] Jung, C., Hawkins, J. A., Jones, S. K., Xiao, Y., Rybarski, J. R., Dillard, K. E., Hussmann, J., Saifuddin, F. A., Savran, C. A., and Ellington, A. D. (2017) Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips, *Cell 170*, 35-47. e13.
[137] Huo, Y., Nam, K. H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M. D., Zhou, S., Rajashankar, K., Kurinov, I., Zhang, R., and Ke, A. (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation, *Nature Structural & Molecular Biology 21*, 771-777.
[138] 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.
[139] 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 & development 30*, 447-459.
[140] Estrella, M. A., Kuo, F.-T., and Bailey, S. (2016) RNA-activated DNA cleavage by the Type III-B CRISPR–Cas effector complex, *Genes & development 30*, 460-470.
[141] Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č., and Siksnys, V. (2016) Spatiotemporal control of type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition, *Molecular cell 62*, 295-306.
[142] Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017) A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems, *Science 357*, 605-609.
[143] 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.
[144] Staals, R. H., Zhu, Y., Taylor, D. W., Kornfeld, J. E., Sharma, K., Barendregt, A., Koehorst, J. J., Vlot, M., Neupane, N., and Varossieau, K. (2014) RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus, *Molecular cell 56*, 518-530.
[145] Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, Č., 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, *Molecular cell 56*, 506-517.
[146] Athukoralage, J. S., Rouillon, C., Graham, S., Grüschow, S., and White, M. F. (2018) Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate, *Nature 562*, 277-280.
[147] Wang, L., Mo, C. Y., Wasserman, M. R., Rostøl, J. T., Marraffini, L. A., and Liu, S. (2019) Dynamics of Cas10 govern discrimination between self and non-self in type III CRISPR-Cas immunity, *Molecular cell 73*, 278-290. e274.
[148] Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R. A., Randau, L., Sørensen, S. J., and Shah, S. A. (2020) Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids, *Nucleic acids research 48*, 2000-2012.
[149] Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., and Lin, S. (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation, *Science 343*, 1247997.
[150] Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., and Doudna, J. A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage, *Science 351*, 867-871.
[151] 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.

[152] Liu, J.-J., Orlova, N., Oakes, B. L., Ma, E., Spinner, H. B., Baney, K. L., Chuck, J., Tan, D., Knott, G. J., and Harrington, L. B. (2019) CasX enzymes comprise a distinct family of RNA-guided genome editors, *Nature 566*, 218-223.
[153] Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C. A., Li, Z., Cress, B. F., Knott, G. J., Jacobsen, S. E., Banfield, J. F., and Doudna, J. A. (2020) CRISPR-CasΦ from huge phages is a hypercompact genome editor, *Science 369*, 333-337.
[154] Yang, H., Gao, P., Rajashankar, K. R., and Patel, D. J. (2016) PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease, *Cell 167*, 1814-1828. e1812.
[155] 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.
[156] 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.
[157] Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B., Shmakov, S., Makarova, K. S., Semenova, E., and Minakhin, L. (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector, *Science 353*, aaf5573.
[158] 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.
[159] Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., and Marraffini, L. A. (2013) Multiplex genome engineering using CRISPR/Cas systems, *Science 339*, 819-823.
[160] Chang, H. H., Pannunzio, N. R., Adachi, N., and Lieber, M. R. (2017) Non-homologous DNA end joining and alternative pathways to double-strand break repair, *Nature reviews Molecular cell biology 18*, 495-506.
[161] Bétermier, M., Bertrand, P., and Lopez, B. S. (2014) Is non-homologous end-joining really an inherently error-prone process?, *PLoS genetics 10*, e1004086.
[162] Waters, C. A., Strande, N. T., Wyatt, D. W., Pryor, J. M., and Ramsden, D. A. (2014) Nonhomologous end joining: a good solution for bad ends, *DNA repair 17*, 39-51.
[163] Mateos-Gomez, P. A., Kent, T., Deng, S. K., McDevitt, S., Kashkina, E., Hoang, T. M., Pomerantz, R. T., and Sfeir, A. (2017) The helicase domain of Polθ counteracts RPA to promote alt-NHEJ, *Nature structural & molecular biology 24*, 1116-1123.
[164] Sharma, S., Javadekar, S., Pandey, M., Srivastava, M., Kumari, R., and Raghavan, S. (2015) Homology and enzymatic requirements of microhomology-dependent alternative end joining, *Cell death & disease 6*, e1697-e1697.
[165] Symington, L. S., and Gautier, J. (2011) Double-strand break end resection and repair pathway choice, *Annual review of genetics 45*, 247-271.
[166] Zhang, H., Tombline, G., and Weber, B. L. (1998) BRCA1, BRCA2, and DNA damage response: collision or collusion?, *Cell 92*, 433-436.
[167] Heyer, W.-D., Ehmsen, K. T., and Liu, J. (2010) Regulation of homologous recombination in eukaryotes, *Annual review of genetics 44*, 113-139.
[168] Lotfy, P., and Hsu, P. D. (2022) Genome Editing with CRISPR-Cas Systems, In *Crispr*, pp 163-193.
[169] Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012) Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria, *Proceedings of the National Academy of Sciences 109*, E2579-E2586.

[170] Rees, H. A., and Liu, D. R. (2018) Base editing: precision chemistry on the genome and transcriptome of living cells, *Nature reviews genetics 19*, 770-788.
[171] Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., and Raguram, A. (2019) Search-and-replace genome editing without double-strand breaks or donor DNA, *Nature 576*, 149-157.
[172] Lin, Y.-T., Seo, J., Gao, F., Feldman, H. M., Wen, H.-L., Penney, J., Cam, H. P., Gjoneska, E., Raja, W. K., and Cheng, J. (2018) APOE4 causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types, *Neuron 98*, 1141-1154. e1147.
[173] Meyer, K., Feldman, H. M., Lu, T., Drake, D., Lim, E. T., Ling, K.-H., Bishop, N. A., Pan, Y., Seo, J., and Lin, Y.-T. (2019) REST and neural gene network dysregulation in iPSC models of Alzheimer's disease, *Cell reports 26*, 1112-1127. e1119.
[174] Yang, H., Wang, H., Shivalila, C. S., Cheng, A. W., Shi, L., and Jaenisch, R. (2013) One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering, *Cell 154*, 1370-1379.
[175] Niu, Y., Shen, B., Cui, Y., Chen, Y., Wang, J., Wang, L., Kang, Y., Zhao, X., Si, W., and Li, W. (2014) Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos, *Cell 156*, 836-843.
[176] Chen, S., Lee, B., Lee, A. Y.-F., Modzelewski, A. J., and He, L. (2016) Highly efficient mouse genome editing by CRISPR ribonucleoprotein electroporation of zygotes, *Journal of Biological Chemistry 291*, 14457-14467.
[177] Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014) Genetic Screens in Human Cells Using the CRISPR-Cas9 System, *Science 343*, 80-84.
[178] 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.
[179] Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression, *Nature protocols 8*, 2180-2196.
[180] Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L. A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system, *Nucleic acids research 41*, 7429-7437.
[181] Noordermeer, J. N., Chen, C., and Qi, L. S. (2022) Genetic and Epigenetic Modulation of Gene Expression by CRISPR-dCas Systems, In *Crispr*, pp 195-212.
[182] Bikard, D., Hatoum-Aslan, A., Mucida, D., and Marraffini, L. A. (2012) CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection, *Cell host & microbe 12*, 177-186.
[183] Bikard, D., Euler, C. W., Jiang, W., Nussenzweig, P. M., Goldberg, G. W., Duportet, X., Fischetti, V. A., and Marraffini, L. A. (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials, *Nature biotechnology 32*, 1146-1150.
[184] Kim, J.-S., Cho, D.-H., Park, M., Chung, W.-J., Shin, D., Ko, K. S., and Kweon, D.-H. (2016) CRISPR/Cas9-mediated re-sensitization of antibiotic-resistant Escherichia coli harboring extended-spectrum β-lactamases, *Journal of microbiology and biotechnology 26*, 394-401.
[185] Sheth, R. U., Yim, S. S., Wu, F. L., and Wang, H. H. (2017) Multiplex recording of cellular events over time on CRISPR biological tape, *Science 358*, 1457-1461.
[186] Shipman, S. L., Nivala, J., Macklis, J. D., and Church, G. M. (2016) Molecular recordings by directed CRISPR spacer acquisition, *Science 353*, aaf1175.
[187] Gillmore, J. D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M. L., Seitzer, J., O'Connell, D., Walsh, K. R., Wood, K., Phillips, J., Xu, Y., Amaral, A., Boyd, A. P.,

Cehelsky, J. E., McKee, M. D., Schiermeier, A., Harari, O., Murphy, A., Kyratsous, C. A., Zambrowicz, B., Soltys, R., Gutstein, D. E., Leonard, J., Sepp-Lorenzino, L., and Lebwohl, D. (2021) CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis, *New England Journal of Medicine 385*, 493-502.

[188] Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., Ho, T. W., Kattamis, A., Kernytsky, A., Lekstrom-Himes, J., Li, A. M., Locatelli, F., Mapara, M. Y., de Montalembert, M., Rondelli, D., Sharma, A., Sheth, S., Soni, S., Steinberg, M. H., Wall, D., Yen, A., and Corbacioglu, S. (2020) CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia, *New England Journal of Medicine 384*, 252-260.

[189] Maikova, A., Boudry, P., Shiriaeva, A., Vasileva, A., Boutserin, A., Medvedeva, S., Semenova, E., Severinov, K., and Soutourina, O. (2021) Protospacer-adjacent motif specificity during clostridioides difficile type ib crispr-cas interference and adaptation, *MBio 12*, e02136-02121.

[190] Richter, H., Rompf, J., Wiegel, J., Rau, K., and Randau, L. (2017) Fragmentation of the CRISPR-Cas Type IB signature protein Cas8b, *Biochimica et Biophysica Acta (BBA)-General Subjects 1861*, 2993-3000.

[191] Rollie, C., Graham, S., Rouillon, C., and White, M. F. (2018) Prespacer processing and specific integration in a Type IA CRISPR system, *Nucleic acids research 46*, 1007-1020.

[192] Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014) In vitro assembly and activity of an archaeal CRISPR-Cas type IA Cascade interference complex, *Nucleic acids research 42*, 5125-5138.

[193] Hu, C., Ni, D., Nam, K. H., Majumdar, S., McLean, J., Stahlberg, H., Terns, M. P., and Ke, A. (2022) Allosteric control of type I-A CRISPR-Cas3 complexes and establishment as effective nucleic acid detection and human genome editing tools, *Molecular Cell 82*, 2754-2768.e2755.

[194] 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.

[195] O'Brien, R. E., Santos, I. C., Wrapp, D., Bravo, J. P. K., Schwartz, E. A., Brodbelt, J. S., and Taylor, D. W. (2020) Structural basis for assembly of non-canonical small subunits into type I-C Cascade, *Nature Communications 11*, 5931.

[196] Lin, J., Fuglsang, A., Kjeldsen, A. L., Sun, K., Bhoobalan-Chitty, Y., and Peng, X. (2020) DNA targeting by subtype I-D CRISPR–Cas shows type I and type III features, *Nucleic Acids Research 48*, 10470-10478.

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

[198] Chowdhury, S., Carter, J., Rollins, M. F., Golden, S. M., Jackson, R. N., Hoffmann, C., Bondy-Denomy, J., Maxwell, K. L., Davidson, A. R., and Fischer, E. R. (2017) Structure reveals mechanisms of viral suppressors that intercept a CRISPR RNA-guided surveillance complex, *Cell 169*, 47-57. e11.

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

[200] Pausch, P., Müller-Esparza, H., Gleditzsch, D., Altegoer, F., Randau, L., and Bange, G. (2017) Structural variation of type IF CRISPR RNA guided DNA surveillance, *Molecular cell 67*, 622-632. e624.

[201] 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.
[202] Singleton, M. R., Dillingham, M. S., and Wigley, D. B. (2007) Structure and Mechanism of Helicases and Nucleic Acid Translocases, *Annual Review of Biochemistry 76*, 23-50.
[203] He, L., St John James, M., Radovcic, M., Ivancic-Bace, I., and Bolt, E. L. (2020) Cas3 Protein-A Review of a Multi-Tasking Machine, *Genes (Basel) 11*.
[204] 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, *The EMBO Journal 32*, 385-394.
[205] 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, *The EMBO Journal 30*, 1335-1342.
[206] Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011) The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli, *Nucleic Acids Research 39*, 9275-9282.
[207] Zheng, Y., Han, J., Wang, B., Hu, X., Li, R., Shen, W., Ma, X., Ma, L., Yi, L., and Yang, S. (2019) Characterization and repurposing of the endogenous Type IF CRISPR–Cas system of Zymomonas mobilis for genome engineering, *Nucleic acids research 47*, 11461-11475.
[208] Zhang, J., Zong, W., Hong, W., Zhang, Z.-T., and Wang, Y. (2018) Exploiting endogenous CRISPR-Cas system for multiplex genome editing in Clostridium tyrobutyricum and engineer the strain for high-level butanol production, *Metabolic engineering 47*, 49-59.
[209] Csörgő, B., León, L. M., Chau-Ly, I. J., Vasquez-Rifo, A., Berry, J. D., Mahendra, C., Crawford, E. D., Lewis, J. D., and Bondy-Denomy, J. (2020) A compact Cascade-Cas3 system for targeted genome engineering, *Nat Methods 17*, 1183-1190.
[210] Maikova, A., Kreis, V., Boutserin, A., Severinov, K., and Soutourina, O. (2019) Using an endogenous CRISPR-Cas system for genome editing in the human pathogen Clostridium difficile, *Applied and environmental microbiology 85*, e01416-01419.
[211] Xu, Z., Li, Y., Cao, H., Si, M., Zhang, G., Woo, P. C. Y., and Yan, A. (2021) A transferrable and integrative type I-F Cascade for heterologous genome editing and transcription modulation, *Nucleic Acids Research 49*, e94-e94.
[212] Cameron, P., Coons, M. M., Klompe, S. E., Lied, A. M., Smith, S. C., Vidal, B., Donohoue, P. D., Rotstein, T., Kohrs, B. W., and Nyer, D. B. (2019) Harnessing type I CRISPR–Cas systems for genome engineering in human cells, *Nature biotechnology 37*, 1471-1477.
[213] Chen, Y., Liu, J., Zhi, S., Zheng, Q., Ma, W., Huang, J., Liu, Y., Liu, D., Liang, P., and Songyang, Z. (2020) Repurposing type I–F CRISPR–Cas system as a transcriptional activation tool in human cells, *Nature communications 11*, 3136.
[214] Dolan, A. E., Hou, Z., Xiao, Y., Gramelspacher, M. J., Heo, J., Howden, S. E., Freddolino, P. L., Ke, A., and Zhang, Y. (2019) Introducing a spectrum of long-range genomic deletions in human embryonic stem cells using type I CRISPR-Cas, *Molecular cell 74*, 936-950. e935.
[215] Morisaka, H., Yoshimi, K., Okuzaki, Y., Gee, P., Kunihiro, Y., Sonpho, E., Xu, H., Sasakawa, N., Naito, Y., Nakada, S., Yamamoto, T., Sano, S., Hotta, A., Takeda, J., and Mashimo, T. (2019) CRISPR-Cas3 induces broad and unidirectional genome editing in human cells, *Nature Communications 10*, 5302.

[216] Osakabe, K., Wada, N., Murakami, E., Miyashita, N., and Osakabe, Y. (2021) Genome editing in mammalian cells using the CRISPR type I-D nuclease, *Nucleic Acids Research 49*, 6347-6363.
[217] Pickar-Oliver, A., Black, J. B., Lewis, M. M., Mutchnick, K. J., Klann, T. S., Gilcrest, K. A., Sitton, M. J., Nelson, C. E., Barrera, A., and Bartelt, L. C. (2019) Targeted transcriptional modulation with type I CRISPR–Cas systems in human cells, *Nature biotechnology 37*, 1493-1501.
[218] Tan, R., Krueger, R. K., Gramelspacher, M. J., Zhou, X., Xiao, Y., Ke, A., Hou, Z., and Zhang, Y. (2022) Cas11 enables genome engineering in human cells with compact CRISPR-Cas3 systems, *Molecular Cell 82*, 852-867.e855.
[219] Luo, M. L., Mullis, A. S., Leenay, R. T., and Beisel, C. L. (2015) Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression, *Nucleic acids research 43*, 674-681.
[220] Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P. R., and Lundgren, M. (2015) Efficient programmable gene silencing by Cascade, *Nucleic acids research 43*, 237-246.
[221] Oost, J. v. d. (2013) New tool for genome surgery, *Science 339*, 768-770.
[222] Rouillon, C., Athukoralage, J. S., Graham, S., Grüschow, S., and White, M. F. (2019) Investigation of the cyclic oligoadenylate signaling pathway of type III CRISPR systems, In *Methods in Enzymology*, pp 191-218, Elsevier.
[223] Grüschow, S., Athukoralage, J. S., Graham, S., Hoogeboom, T., and White, M. F. (2019) Cyclic oligoadenylate signalling mediates Mycobacterium tuberculosis CRISPR defence, *Nucleic acids research 47*, 9259-9270.
[224] Singh, N., and Bose, K. (2022) Introduction to Recombinant Protein Purification, In *Textbook on Cloning, Expression and Purification of Recombinant Proteins* (Bose, K., Ed.), pp 115-140, Springer Nature Singapore, Singapore.
[225] Reid, S. L., Parry, D., Liu, H.-H., and Connolly, B. A. (2001) Binding and Recognition of GATATC Target Sequences by the EcoRV Restriction Endonuclease: A Study Using Fluorescent Oligonucleotides and Fluorescence Polarization, *Biochemistry 40*, 2484-2494.
[226] Zhu, W., McQuarrie, S., Grüschow, S., McMahon, S. A., Graham, S., Gloster, T. M., and White, M. F. (2021) The CRISPR ancillary effector Can2 is a dual-specificity nuclease potentiating type III CRISPR defence, *Nucleic Acids Research 49*, 2777-2789.
[227] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., and Potapenko, A. (2021) Highly accurate protein structure prediction with AlphaFold, *Nature 596*, 583-589.
[228] Holm, L. (2020) DALI and the persistence of protein shape, *Protein Science 29*, 128-140.
[229] Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., and Ferrin, T. E. (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers, *Protein Science 30*, 70-82.
[230] Shangguan, Q., Graham, S., Sundaramoorthy, R., and White, Malcolm F. (2022) Structure and mechanism of the type I-G CRISPR effector, *Nucleic Acids Research 50*, 11214-11228.
[231] 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 Research 47*, 5223-5230.
[232] Kieper, S. N., Almendros, C., Behler, J., McKenzie, R. E., Nobrega, F. L., Haagsma, A. C., Vink, J. N., Hess, W. R., and Brouns, S. J. (2018) Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation, *Cell reports 22*, 3377-3384.

[233] Lee, H., Dhingra, Y., and Sashital, D. G. (2019) The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation, *Elife 8*, e44248.
[234] Charpentier, E., Richter, H., van der Oost, J., and White, M. F. (2015) Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity, *FEMS microbiology reviews 39*, 428-441.
[235] Westra, Edze R., van Erp, Paul B. G., Künne, T., Wong, Shi P., Staals, Raymond H. J., Seegers, Christel L. C., Bollen, S., Jore, Matthijs M., Semenova, E., Severinov, K., de Vos, Willem M., Dame, Remus T., de Vries, R., Brouns, Stan J. J., and van der Oost, J. (2012) CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3, *Molecular Cell 46*, 595-605.
[236] 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.
[237] Pan, M., Morovic, W., Hidalgo-Cantabrana, C., Roberts, A., Walden, K. K. O., Goh, Y. J., and Barrangou, R. (2022) Genomic and epigenetic landscapes drive CRISPR-based genome editing in Bifidobacterium, *Proceedings of the National Academy of Sciences 119*, e2205068119.
[238] Guo, T. W., Bartesaghi, A., Yang, H., Falconieri, V., Rao, P., Merk, A., Eng, E. T., Raczkowski, A. M., Fox, T., Earl, L. A., Patel, D. J., and Subramaniam, S. (2017) Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex, *Cell 171*, 414-426.e412.
[239] Kempen, M. v., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2023) Fast and accurate protein structure search with Foldseek, *bioRxiv*, 2022.2002.2007.479398.
[240] Zhou, Y., Bravo, J. P. K., Taylor, H. N., Steens, J. A., Jackson, R. N., Staals, R. H. J., and Taylor, D. W. (2021) Structure of a type IV CRISPR-Cas ribonucleoprotein complex, *iScience 24*, 102201.
[241] McBride, T. M., Schwartz, E. A., Kumar, A., Taylor, D. W., Fineran, P. C., and Fagerlund, R. D. (2020) Diverse CRISPR-Cas Complexes Require Independent Translation of Small and Large Subunits from a Single Gene, *Molecular Cell 80*, 971-979.e977.
[242] Niu, Y., Yang, L., Gao, T., Dong, C., Zhang, B., Yin, P., Hopp, A.-K., Li, D., Gan, R., Wang, H., Liu, X., Cao, X., Xie, Y., Meng, X., Deng, H., Zhang, X., Ren, J., Hottiger, M. O., Chen, Z., Zhang, Y., Liu, X., and Feng, Y. (2020) A Type I-F Anti-CRISPR Protein Inhibits the CRISPR-Cas Surveillance Complex by ADP-Ribosylation, *Molecular Cell 80*, 512-524.e515.
[243] Pawluk, A., Shah, M., Mejdani, M., Calmettes, C., Moraes, T. F., Davidson, A. R., and Maxwell, K. L. (2017) Disabling a Type I-E CRISPR-Cas Nuclease with a Bacteriophage-Encoded Anti-CRISPR Protein, *mBio 8*, 10.1128/mbio.01751-01717.
[244] Schwartz, E. A., McBride, T. M., Bravo, J. P. K., Wrapp, D., Fineran, P. C., Fagerlund, R. D., and Taylor, D. W. (2022) Structural rearrangements allow nucleic acid discrimination by type I-D Cascade, *Nature Communications 13*, 2829.
[245] 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.
[246] Shangguan, Q., and White, M. F. (2023) Repurposing the atypical type I-G CRISPR system for bacterial genome engineering, *Microbiology 169*.

[247] Gomaa, A. A., Klumpe, H. E., Luo, M. L., Selle, K., Barrangou, R., and Beisel, C. L. (2014) Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems, *MBio 5*, e00928-00913.
[248] Vercoe, R. B., Chang, J. T., Dy, R. L., Taylor, C., Gristwood, T., Clulow, J. S., Richter, C., Przybilski, R., Pitman, A. R., and Fineran, P. C. (2013) Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands, *PLoS genetics 9*, e1003454.
[249] Cui, L., and Bikard, D. (2016) Consequences of Cas9 cleavage in the chromosome of Escherichia coli, *Nucleic Acids Research 44*, 4243-4251.
[250] 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, *Journal of Biological Chemistry 288*, 22184-22192.
[251] Chayot, R., Montagne, B., Mazel, D., and Ricchetti, M. (2010) An end-joining repair mechanism in Escherichia coli, *Proceedings of the National Academy of Sciences 107*, 2141-2146.
[252] Hao, Y., Wang, Q., Li, J., Yang, S., Zheng, Y., and Peng, W. (2022) Double nicking by RNA-directed Cascade-nCas3 for high-efficiency large-scale genome engineering, *Open Biology 12*, 210241.
[253] Grissa, I., Vergnaud, G., and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC bioinformatics 8*, 1-10.
[254] Yoshimi, K., and Mashimo, T. (2022) Genome editing technology and applications with the type I CRISPR system, *Gene and Genome Editing 3-4*, 100013.

Appendix

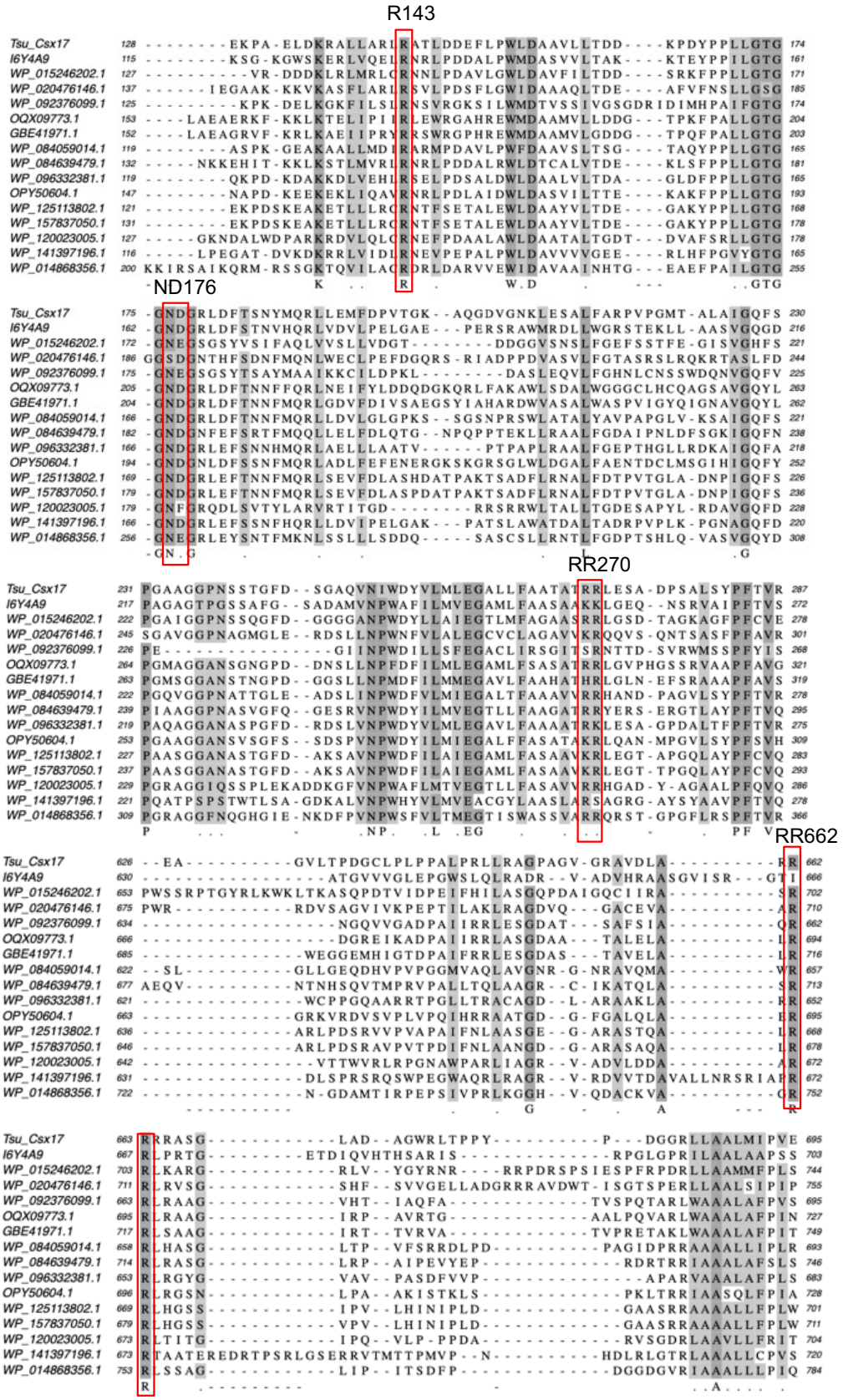


Figure legend next page

Figure 1. Cas8g (Csx17) sequence alignment

T. sulfidiphilus Cas8g protein sequence alignment with type I-G associated Cas8g from other organisms. Combining with *T. sulfidiphilus* Cas8g structure, following site were mutated : Arginine patch, 270-RR to EE for charge reversal; Arginine patch, 662-RR to EE for charge reversal; 172-GTGGNDG loop, 176-ND to AA; R1443 patch, mutate to E for charge reversal.

Table 1.1 CRISPR array

Name	Squence (5' to 3')	Note
TetTarget	GTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCAAGCTGTCCCTGATGGTCGTC ATCTACCTGCCTGGAGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCAAGCTG TCCCTGATGGTCGTCATCTACCTGCCTGGAGTCATCCGCGGCATTTAGCCGCGGCCT CATTGAAGCAAGCTGTCCCTGATGGTCGTCATCTACCTGCCTGGAGTCATCCGCGGC ATTTAGCCGCGGCCTCATTGAAGCAAGCTGTCCCTGATGGTCGTCATCTACCTGCCT GGAGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCAAGCTGTCCCTGATGGTC GTCATCTACCTGCCTGGAGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCAAG CTGTCCCTGATGGTCGTCATCTACCTGCCTGGAGTCATCCGCGGCATTTAGCCGCGG CCTCATTGAAGC	CRISPR array spacer targeting Tetracycline
lacZTarget	GTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCCAGCACATCCCCCTTTCGCCA GCTGGCGTAATAGCGGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCCAGCAC ATCCCCCTTTCGCCAGCTGGCGTAATAGCGGTCATCCGCGGCATTTAGCCGCGGCCT CATTGAAGCCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGGTCATCCGCGGC ATTTAGCCGCGGCCTCATTGAAGCCAGCACATCCCCCTTTCGCCAGCTGGCGTAATA GCGGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGC	CRISPR array spacer targeting *lacZ*
lpaTarget	GTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCATAGATGACAGTACACGCCCA CGTAGATTTCAGATCGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCATAGAT GACAGTACACGCCCACGTAGATTTCAGATCGTCATCCGCGGCATTTAGCCGCGGCCT CATTGAAGCATAGATGACAGTACACGCCCACGTAGATTTCAGATCGTCATCCGCGGC ATTTAGCCGCGGCCTCATTGAAGC	CRISPR array spacer targeting phage *lpa*

Table 1.2 Primers for separating Csb2

Name	Squence (5' to 3')	Note
Csb2 R260 to stop codon	5′- cgttcgctagcgccgtgtaccagagtgg -3′ 5′- cggcgctagcgaacgcccaacgcaccttg -3;	Primers for mutagenesis
C-terminus of Csb2 amplification	5′- **GCTGCCATGGCT**CATATGTTAGCAGTGGCTTGTCG -3′ 5′- **CAGCCTCGAGGGATCCTCA**GCGAACGCCCAACGCACC -3′	Primers with NcoI and BamHI site for Csb2 C-terminal domain amplification

Table 1.3 Oligonucleotides for in vitro assay

Name	Squence (5' to 3')	Note
crRNA repeat	5′-6-FAM- GUCAUCCGCGGCAUUUAGCCGCGGCCUCAUUGAAGC-3′	crRNA repeat with FAM label
3'- Haipin	5′-6-FAM- GUCAUCCGCGGCAUUUAGCCGCGGCCUC-3′	3'haipin with FAM label for anisotropy
5'-8nt-handle	5′- AUUGAAGC-6-FAM- -3′ 5′-6-FAM- AUUGAAGC-3′	5'-8nt handle for anisotropy, 5'FAM label or 3'FAM label
81nt target strand	5′GCCCGCGTTGCAGGCCATGCTGTCCAGGCAGGTAGATGACGACCA TCAGGGACAGCTTCAAGGATCGCTCGCGGCTCTTAC-3′	Tetracycline resistance target
80nt non-target strand	5′GTAAGAGCCGCGAGCGATCCTTGAAGCTGTCCCTGATGGTCGTCA TCTACCTGCCTGGACAGCATGGCCTGCAACGCGGG -3′	Complementary strand for dsDNA duplex

Table 2 Primers for type I-G genome targeting

Name	Squence (5′ to 3′)	Note
T7toaraBAD-F	gcaggaggaatacccatgggcagcgtcgacatggataaggatatgcacattaatgagatcgtcttgcg	Replace T7 promoter by araBAD promote on pACE-M1
T7toaraBAD-R	ctataacggtcctaaggtagcgacctaggtatcgttatgacaacttgacggctacatcattcac	
lacZ-507-F	TTGTGGAGCGACATCCAGAG	Verification on middle of lacZ
lacZ-507-R	GATGAAGACCAGCCCTTCCC	
lacZ-20kUP-F	GTCTTCCAGGGCGGAAATCA	Verification on 20k upstream lacZ
lacZ-20kUP-R	ATTACCCAGTCGAACCCACG	
lacZ-10kUP-F	CGTCCAGTAACCATGTCGCT	Verification on 10k upstream lacZ
lacZ-10kUP-R	AGAATACCCGGTGCAGAAGC	
lacZ-10kDown-F	ATGTTTTGCTGGTGGATCGC	Verification on 10k downstream lacZ
lacZ-10kDown-R	TATACGAATGCCCCACCACC	
lacZ-25kDown-F	CGCATGGCATCGAATACAGC	Verification on 25k downstream lacZ
lacZ-25kDown-R	GTTCGCACCAGCCAAGAATG	
lacZ-40kDown-F	AACATCATTAGCGGCCCCAG	Verification on 40k downstream lacZ
lacZ-40kDown-R	TTGCCTGGCTCTGGGATTTT	
lacZ-55kDown-F	AACTGGGCTTTCAGTCCTGG	Verification on 55k downstream lacZ
lacZ-55kDown-R	CTTGACGACGGGCAGGTTAT	
lacZ-1-veri-F	GCGAGTGGCAACATGGAAAT	Verification on lacZ-1 target site
lacZ-1-veri-R	TTAGGCACCCCAGGCTTTAC	
lacZ-4-veri-F	CCCCATATGGAAACCGTCGAT	Verification on lacZ-4 target site
lacZ-4-veri-R	TCTGACCACCAGCGAAATGG	
HR-5kdown-F	GGCTCATATGCCGCGCATTCCTCCAA	For homologous arm 5k downstream of lacZ
HR-5kdown-R	CAGCCTCGAGGAGCTGGAGGCAATTCCTTT	
HR-5kUP-F	GCTCCTCGAGAAACCGTTGTCCTGCTGCAT	For homologous arm 5k uptream of lacZ
HR-5kUP-R	gcagcctaggCCCCCAGACAATCAGGGTTT	
lacZ-HRveri-F	GACTGGGTTACAGCGAGCTT	Verification on HDR product
lacZ-HRveri-R	TTAAGGGCGTGCGAGGAAAT	

Table 3 Oligonucleotides for genome targeting

Name	Squence (5′ to 3′)	Note
Spacer-lacZ-1-T	AAGCACCGTAATGGGATAGGTCACGTTGGTGTAGATGGGC	Spacer targeting lacZ start site
Spacer-lacZ-1-C	TGACGCCCATCTACACCAACGTGACCTATCCCATTACGGT	
Spacer-lacZ-4-T	AAGCATGGTAGTGGTCAAATGGCGATTACCGTTGATGTTG	Spacer targeting lacZ end site
Spacer-lacZ-4-C	TGACCAACATCAACGGTAATCGCCATTTGACCACTACCAT	
Spacer-frmA-T	AAGCTCGTAGTCATTCGGGTTAATGCAGTCGGTAGCACCG	Spacer targeting frmA gene
Spacer-frmA-C	TGACCGGTGCTACCGACTGCATTAACCCGAATGACTACGA	
Spacer-yahK-T	AAGCGAAAACTACTGTGATCACATGACCGGCACCTATAAC	Spacer targeting yahK gene
Spacer-yahK-C	TGACGTTATAGGTGCCGGTCATGTGATCACAGTAGTTTTC	
Tsu-BpiI-rep-T-5′	catggATCGACTTTTCTGCGAGGGCCGTCATCCGCGGCATTTAGCCGCGGCCT CATTGAAGCgtgtctt	Introduce type I-G repeat sequence with two BpiI site
Tsu-BpiI-rep-T-3′	cgtaccttgaagaccaGTCATCCGCGGCATTTAGCCGCGGCCTCATTGAAGCG CCGAATCTCg	
Tsu-BpiI-rep-C-3′	tcgacGAGATTCGGCGCTTCAATGAGGCCGCGGCTAAATGCCGCGGATGACtg gtcttcaagg	
Tsu-BpiI-rep-C-5′	tacgaagacacGCTTCAATGAGGCCGCGGCTAAATGCCGCGGATGACGGCCCT CGCAGAAAAGTCGATc	

Table 4 *E. coli* strains

Name	Description
DH5α	For vector construction and gene cloning
C43(DE3)	For protein expression
BL21 star	For protein expression
MG1655	For genome targeting
LMG194	For phage propagation

Table 5 Plasmids

Name	Description
pACE-M1	Used for plasmid challenge assay and phage challenge assay. csb2, cas7 and cas8g genes included.
pCDF	Used for plasmid challenge assay and phage challenge assay. CRISPR array targeting tetR or phage lpa included.
pRAT-Duet	Used for plasmid challenge assay and phage challenge assay control. Lacking cas3.
pRAT-Cas3	Used for plasmid challenge assay and phage challenge assay. cas3 included.
pM2	Used for genome targeting. cas3, csb2, cas7 and cas8g genes included.
pSPACER	Used for genome targeting. CRISPR array included.
pHR	Used for genome targeting. CRISPR array and homologous arms included.