University of Nottingham
UK | CHINA | MALAYSIA

RNA-guided integration of diverse DNA molecules by casposase-Cas9 fusions

Chun Hang Lau

Student ID: 14290277

Supervised by Dr Edward L. Bolt

Submitted to the School of Life Science, University of Nottingham in partial fulfilment of the requirements for the degree of Doctor of Philosophy

October 2020

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins build adaptive immunity against mobile genetic elements (MGEs) in prokaryotes. The memory of previously encountered MGEs is established by DNA capture and integration into CRISPR loci catalysed by Cas1-Cas2 adaptation complexes. Cas1 is thought to have evolved from enzymes called casposases, which belong to a novel class of transposases. A previous study showed casposases integrate into target molecules single and double stranded DNA containing terminal inverted repeats (TIRs). In this project, *in vitro* biochemical assays showed that the substrate flexibility of *Acidoprofundum boonei* casposase extends to random integration of DNA without TIRs, including integration of a functional gene.

The DNA substrate tolerance of casposases may make them useful as a tool for biotechnology applications that require targeted DNA integration activity. Casposase-Cas9 fusions were engineered to investigate the targeting of DNA integration to specific DNA sites. The fusion proteins were able to form an R-loop with target DNA and demonstrated RNA-guided DNA integration *in vitro*. Full-site RNA-guided DNA integration products were not detected and DNA integration into some non-specific DNA sites was observed. Expression of fusion proteins in *Escherichia coli* cells could target both chromosomal and plasmid *lacZ* gene but casposase catalysed DNA integration was not detected. Casposase-Cas9 fusions might be a useful prototype of a genome editing tool for targeted DNA insertion that is independent of homology-directed DNA repair.

In addition, the project investigated alternatives to casposase-Cas9 fusions. A casposase-dCasX fusion construct yielded protein that could not be successfully purified. A tyrosine recombinase, Int^{pTN3} integrase, was fused to dCas9 and successfully purified. However, RNA-guided recombination between homologous sequences was not detected. Future work will require a deeper understanding of the low sequence specificity recombination activity of Int^{pTN3}.

Layperson summary

DNA, the genetic material in living organisms, is like a digital memory storing information for cell activity. In some bacteria, the DNA also stores information about invaders such as hostile viruses and this information is used in fighting against viral infection of the matching virus. This defence system is called CRISPR-Cas system and CRISPR is the stretch of DNA that would generate guide sequences, which direct Cas9 protein to cut matching DNA. The Cas9 protein is a molecular scissors and it can be mutated to become a DNA binding protein. CRISPR-Cas systems, in particular CRISPR-Cas9, are simple yet powerful tools for introducing desirable changes into the genetic information of an organism. In this project, the Cas9 or mutated Cas9 proteins were tethered to a casposase protein to investigate programmable DNA insertion that does not require the host cell's processing. This was because we showed that casposase could insert different DNA into random DNA sites. By fusing casposase to Cas9, targeted DNA integration into a user-defined target site was observed but DNA integration into random sites remained a big problem. Thus, this project showed a potential prototype of a programmable DNA insertion tool that requires further improvements.

Acknowledgements

I would like to thank Dr Edward L. Bolt for giving me the opportunity to work in his lab. Dr Bolt was always available for motivating discussion leading to new ideas and experiments. I would like to thank Dr Tom Killelea, a postdoctoral researcher in the lab, for reading part of my thesis and providing valuable comments. I would like to express my deep gratitude to all the current and former lab members that made working so much enjoyable, especially Tabitha Jenkins for organising lab outing events, Andrew Cubbon for being a teammate of team Cas9 and other members for bringing cakes and baked foods.

I am sending my deepest gratitude to my family for supporting me in all aspects. I am very grateful to my wife Yi Wang and my cat Duchess for their emotional support.

Abbreviations

1D- one dimensional

3D – three dimensional

Amp – ampicillin

Amp^R – ampicillin resistance

ATP – adenosine triphosphate

BFP – blue fluorescent protein

bp – base pairs

CARF – CRISPR-associated Rossman fold

Cas – CRISPR associated

Casp-Cas9 – casposase-$(GGS)_8$-Cas9

Casp-dCas9 – casposase-$(GGS)_8$-dCas9

CaspoR – casposase-Cas9 mediated recombination

CFE – cell-free extracts

CIP – alkaline phosphatase from calf intestinal

Cm – chloramphenicol

Cm^R – chloramphenicol resistance

cOA – cyclic oligoadenylate

CRISPR – clustered regularly interspaced short palindromic repeats

CRISPRa – CRISPR activation

crRNA – CRISPR RNA

Cyro-EM – cryo-electron microscopy

dCas9 – catalytically dead Cas9

DE – double end

DMEM – Dulbecco's Modified Eaglr Medium

DSB – double stranded break

dsDNA – double-stranded DNA

DTT – dithiothreitol

EDTA – Ethylenediaminetetraacetic acid

EGFP – enhanced green fluorescent protein

EMSA – Electrophoretic mobility shift assay

ES – embryonic stem

EtBr – ethidium bromide

EV – empty vector

FIAU – 1-(2-deoxy-2-fluoro-1-D-arabinofuranosyl)-5-iodouracil

HDR – homology-directed repair

HEPN – Higher Eukaryotes and Prokaryotes Nucleotide-binding

Hf – high-fidelity

HRP – horse radish peroxidase

HTH – helix-turn-helix

IHF – integration host factor

Indels – insertions/deletions

IPTG – isopropyl-β-D-thiogalactopyranoside

K_d – disassociation constant

LB – Luria-Bertani broth

LE – left end

MBP – maltose-binding protein

MGEs – mobile genetic elements

MW – molecular weight

MWCO – molecular weight cut off

nCas9 – nickase Cas9

NHEJ – non-homologous end joining

NLS – nuclear localisation signal

nt – nucleotides

OD – optical density

ONPG – ortho-Nitrophenyl-β-galactoside

PAM – protospacer adjacent motif

PBS – phosphate buffered saline

PCR – polymerase chain reaction

PI – PAM-interacting

PolBs – family B DNA polymerases

pre-crRNA – precursor CRISPR RNA

RNP – ribonucleoprotein

rpm – rotations per minute

RRM – RNA recognition motif

RT – room temperature

RVD – repeat variable diresidue

SDS-PAGE - SDS-polyacrylamide gel electrophoresis

SE – single end

sgRNA – single guide RNA

smFRET – small-molecule fluorescence resonance energy transfer

SOB – Super optimal broth

SOC – Super optimal broth with catabolite repression

SpCas9 – *Streptococcus pyogenes* Cas9

KDa – kilodalton

SSB – single stranded binding protein

ssDNA – single-stranded DNA

ssODN – single-stranded oligodeoxynucleotide

SSRs – site-specific recombinases

TAE – tris-acetate-EDTA

TALENs - transcription activator-like effector nucleases

TBE – tris-borate-EDTA

TIRs – terminal inverted repeats

tracrRNA – transactivating CRISPR RNA

TSD – target site duplication

TSL – target strand loading

Wt – wild-type

ZFNs – zinc finger nucleases

Table of contents

List of Figures

List of Tables

Chapter 1

1. Introduction

1.1. CRISPR-Cas system

1.1.1. Overview of CRISPR systems

Viruses can infect and replicate within prokaryotes. They are estimated to be ten times more abundant in the biosphere than their prokaryotic hosts and they act as a selective pressure to drive their hosts' evolution (Bikard and Marraffini, 2012). Prokaryotes have evolved countermeasures to prevent the invasion of mobile genetic elements (MGEs) such as viral DNA/RNA, transposons and plasmids. These can be grouped into innate immunity and adaptive immunity. The prokaryotic innate immunity mechanisms include modification of surface receptors, restriction modification and abortive infection systems. While these innate immunities are non-specific and recognises only generic features of invaders, the prokaryotic Argonaute based innate immunity and CRISPR-Cas adaptive immunity are nucleic acid-guided defence systems that target specific sequence (Barrangou et al., 2007; Koonin and Krupovic, 2015). In particular, the prokaryotic adaptive immunity can generate memory of previously encountered invaders so it confers a rapid, robust response upon reinvasion. CRISPR-Cas systems are the only one known prokaryotic adaptive immune system thus far (Koonin, 2017).

CRISPR stands for clustered regularly interspaced short palindromic repeats and the direct repeats are separated by DNA sequences called spacers that are derived from previously invading MGEs (Jansen et al., 2002; Mojica et al., 2005) (Figure 1A). Unlike human adaptive immune system, the immunological memory stored in the CRISPR locus is heritable and is passed on to progeny after division (Barrangou et al., 2007).

Upstream of the CRISPR is a region called leader, which contains the promoter for transcription of the CRISPR array. Adjacent to the CRISPR array, a cluster of CRISPR associated (CAS) genes is often found and they encode proteins catalysing and regulating adaptive immunity. The CRISPR-Cas mediated defence can be divided into three stages: adaptation, crRNA expression and interference (Marraffini and Sontheimer, 2010) (Figure 1B). Upon the invasion of foreign DNA, the DNA is degraded by the host RecBCD enzyme or CRISPR effector complex. In adaptation, the degraded DNA forms prespacer substrates that are recognised by the Cas1-Cas2 complex and are integrated into the CRISPR locus as spacers. The integrated spacers provide a memory of prior encounters with MGEs that through crRNA expression and interference, CRISPR Cas systems destroy the same or similar MGE in future encounters. During the expression stage, the CRISPR array is transcribed into a long precursor CRISPR RNA (pre-crRNA) which is then processed by Cas protein and/or other host factors to generate matured crRNA. Finally, a single crRNA is loaded onto Cas protein(s) to form a crRNA-effector complex. During interference, the first step in target search is finding and recognising a 2–6 bp protospacer adjacent motif (PAM). After the recognition of the correct PAM, the crRNA begins base pairing with the target. If the target DNA or RNA molecules are complementary to the spacer in the crRNA, the effector complex will degrade the target. Because the PAM is absent in spacers in the CRISPR array, it plays a role in discriminating between self and non-self in CRISPR-Cas immune systems (Jackson et al., 2017).

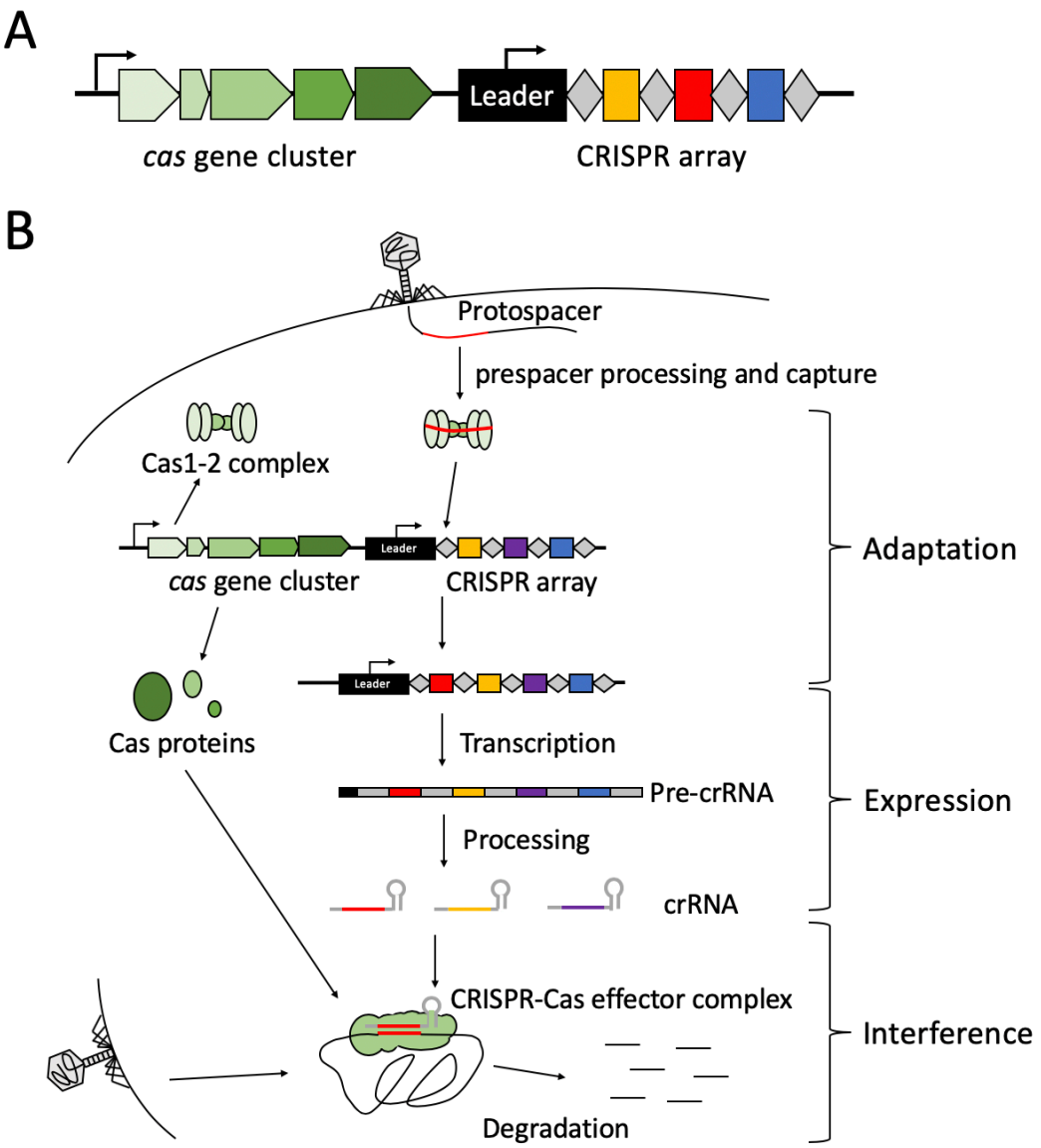


Figure 1. Overview of CRISPR-Cas mediated adaptive immunity.
(A) Genomic structure of CRISPR locus. The grey diamond shapes represent the CRISPR repeats and the coloured boxes represent spacers. The leader sequence is upstream of the CRISPR array and a cluster of *cas* genes is associated with the array. (B) CRISPR-Cas adaptive immunity involves three stages: adaptation, expression and interference. During adaptation, a new spacer from a foreign source is integrated into the CRISPR array. In expression, the long CRISPR array is transcribed and processed to give rise to individually matured crRNA. In the interference stage, the crRNA assembles with Cas proteins to form effector complex which recognises specific foreign nucleic acids by sequence complementarity to the crRNA and degrades it. Adapted with permission from AAAS and from Jackson, S. A., McKenzie, R. E., Fagerlund, R. D., Kieper, S. N., Fineran, P. C., & Brouns, S. J. (2017). CRISPR-Cas: Adapting to change. *Science*, 356(6333).

1.1.2. Classification

CRISPR-Cas systems have been found in 85% sequenced archaeal genomes and in 40% bacterial genomes (Makarova et al., 2020). They are divided into six types according to their signature genes present in the *cas* gene cluster and are further divided into 33 subtypes by gene compositions and architectures of the CRISPR-Cas loci. These six types are grouped into two classes based on the architecture of the effector complex in the interference stage (Figure 2). In class 1 systems (types I, III, IV), the effector complexes, comprising multiple subunits of different Cas proteins, bind crRNA and process the target whereas in class 2 (types II, V, VI), the effector complexes employ a single multi-domain Cas protein (Makarova et al., 2015). The signature protein in type I systems is Cas3, a protein containing helicase and nuclease domains that is recruited to the Cascade complex and carries out DNA degradation (Sinkunas et al., 2011). In type III systems, the signature protein is Cas10 that produces cyclic oligoadenylate for signal transduction and cleaves single-stranded DNA (ssDNA) in a Cascade-like complex called Csm or Cmr complex depending on the sub-type (Jung et al., 2015; Kazlauskiene et al., 2017). In type IV systems, the signature protein is an uncharacterised protein called Csf1 and interestingly there are often no detectable associated CRISPR arrays found in these systems (Wright et al., 2016). In class 2 systems, Cas9 is the signature protein in type II systems and it possesses multiple domains for target DNA recognition and cleavage (Nishimasu et al., 2014). Similar to type II, type V systems also utilise a large single endonuclease, Cas12, that targets and cleaves dsDNA with the aid of crRNA (Dong et al., 2016). In type VI systems, the signature protein Cas13 exclusively targets RNA. It is a single effector protein that has two HEPN domains; one for target RNA cleavage and one for non-specific *trans*-RNA

cleavage that is activated by target binding (Abudayyeh et al., 2016; Meeske et al., 2019).

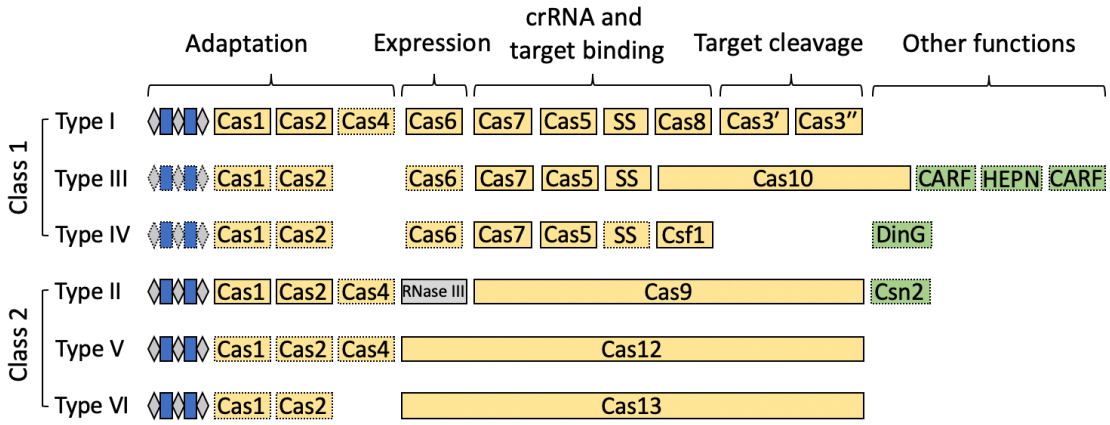


Figure 2. Modular organisation of different types of CRISPR-Cas systems.
The core Cas proteins in each type were coloured light yellow and the ancillary proteins were coloured in green. Dispensable or missing components in some variants are indicated by dashed outlines. In type II systems, crRNA processing and maturation is catalysed by RNase III (coloured in grey), a non-Cas protein from the host. In some type III variants, Cas10 and ancillary proteins containing CRISPR-associated Rossman fold (CARF) domain and Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domain function in signal transduction and non-specific RNA cleavage. Figure adapted from (Koonin and Makarova, 2019).

Apart from the CRISPR-Cas core proteins that contribute to adaptive immunity, many ancillary proteins have been found associated with various subsets of CRISPR-Cas systems (Athukoralage et al., 2018; Makarova et al., 2015, 2020). Computational analysis of these CRISPR-linked ancillary proteins predicts their functions in signal transduction pathways and in membrane association (Shmakov et al., 2018). This suggests some CRISPR-Cas variants might play other roles beyond adaptive immunity (Faure et al., 2019). Indeed, experimental results demonstrated Cas10 in Csm effector complexes from CRISPR type III-A systems catalyses the synthesis of cyclic oligoadenylate (cOA) molecules upon binding of the crRNA-Csm complex to the target RNA (Kazlauskiene et al., 2017; Niewoehner et al., 2017). These cOA second

messengers bind to a CRISPR-associated Rossman fold (CARF) domain in the Csm6 protein, a CRISPR-linked protein, and allosterically activate the non-specific RNase activity of a Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domain in the Csm6. This indiscriminate RNA cleavage causes growth arrest of the host cell and allows more time for the Cas10 DNase activity to degrade invading MGEs (Rostøl and Marraffini, 2019). Then a CRISPR ancillary ring nuclease containing a CARF domain degrades the cOA molecules acting as an off-switch for the signal transduction pathway and the cell resumes normal once the invading MGE has been cleared from the cell (Athukoralage et al., 2018).

1.1.3. The detailed mechanism of CRISPR-Cas systems

1.1.3.1. Adaptation

The success of CRISPR-Cas systems in defending invasion is not only due to the existing spacer records obtained from previous invaders, but also owing to the acquisition of new spacers to expand the arsenal. The spacer acquisition is mainly catalysed by the universally conserved Cas1-Cas2 complex despite the divergence of CRISPR-Cas systems. The X-ray crystal structure of *Escherichia coli* Cas1-Cas2 complex illustrates that two Cas1 dimers are connected by a central Cas2 dimer forming a $Cas1_4$-$Cas2_2$ complex (Nuñez et al., 2015a). Each Cas1 dimer carries the active site for spacer integration while the Cas2 dimer only confers structural stability to the complex (Nuñez et al., 2014). The fact that *cas1* and *cas2* genes present in nearly all CRISPR-Cas systems suggests that the ancestor of CRISPR-Cas systems encoded an integrase and the prokaryotic adaptive immune system began with the insertion of foreign sequences (Jackson et al., 2017). Indeed, once spacers from MGEs are stored in the

host's genome, distinct effector modules which can utilise this information are recruited to the locus and give rise to different types of CRISPR-Cas systems.

Prior to spacer integration, protospacers in MGEs need to be processed into prespacers for binding to Cas1-Cas2 complex. The acquisition of spacers from MGEs that are encountered the first time is called naïve adaptation. During the naïve adaptation, DNA of MGEs is degraded into small pieces and the ones containing the PAM sequence are recognised and bound to Cas1-Cas2 complex. Stalled replication forks occurred during DNA replication can lead to double stranded break (DSB), of which the repair is initiated by RecBCD complex. It was suggested that during the repair, RecBCD complex degrades DNA to an upstream Chi site to generate prespacers for Cas1-Cas2 complex. The acquisition of spacers from the host genome is deleterious so the adaptation bias towards MGEs is achieved by the higher frequency of stalled replication forks for MGEs during DNA replication and smaller numbers of Chi sites in MGEs (Levy et al., 2015). Other than naïve adaptation, type I systems show an additional measure called priming to acquire spacers from previously encountered foreign MGEs using an existing spacer. Priming occurs for both perfectly matched spacers and mismatch containing spacers and this enables efficient defence against fast replicating MGEs and highly divergent MGEs. During priming, the crRNA-Cascade complex binds to a DNA target and recruits Cas3 to the non-target strand unwinding the dsDNA in a 3' to 5' direction. Then the Cas1-Cas2 complex is recruited to the unwound DNA and leads to acquisition of new spacers distal to the initial priming site (Amitai and Sorek, 2016). Although RecBCD complex and Cas3 both produce ssDNA degradation products, it was shown that Cas1-Cas2 complex captures ssDNA in a PAM-dependent manner and facilitates the pairing to a complementary DNA strand leading

to precursor duplex DNA with long 3′ overhangs (Kim et al., 2020). In systems containing Cas4 protein, the Cas4 nuclease tightly interacts with Cas1 to form a Cas4-Cas1-Cas2 complex (Lee et al., 2019). The Cas4 trims the Cas1-Cas2 bound prespacer to the correct size up to a PAM sequence (Figure 3). In some systems, Cas4 remains bound to Cas1 during spacer integration and ensures the correct orientation of DNA integrated into the CRISPR array (Lee et al., 2018; Shiimori et al., 2018). In CRISPR-Cas systems devoid of Cas4 such as type I-E, a host 3′ to 5′ exonuclease DnaQ or ExoT are used to trim the exposed part of prespacer that is not protected by Cas1-Cas2 binding. The trimming of prespacer 3′ overhangs is asymmetric that the PAM containing 3′ overhang is slightly longer than the non-PAM 3′ overhang, because a C-terminal tail of a non-catalytic Cas1 subunit interacts with the PAM and protect it from degradation (Kim et al., 2020; Ramachandran et al., 2020; Wang et al., 2015).

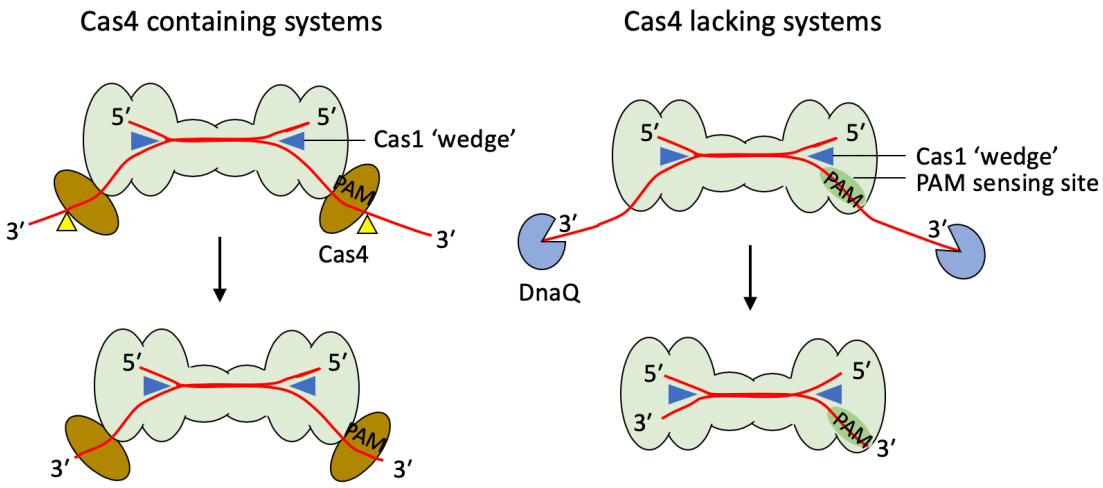


Figure 3. Prespacer processing in CRISPR systems containing or lacking Cas4.
Schematic diagram showing Cas4 dependent (left) and Cas4 independent (right) prespacer processing in CRISPR-Cas systems. In CRISPR-Cas systems containing Cas4, Cas4 interacts with Cas1 and cleaves the 3′ overhangs at the PAM site. In Cas4 independent prespacer processing, a host exonuclease such as DnaQ trims the 3′ overhangs of a Cas1-Cas2 (green) bound prespacer to the correct size. Prespacer (red) is splayed at both ends in the Cas1-Cas2 complex and the double fork structure is stabilised by a Cas1 tyrosine ‘wedge’ (blue triangle) at each end.

The adaptation process can be divided into three steps: substrate capture by Cas1-Cas2 complex, recognition of CRISPR array and integration into CRISPR array. Depending on the type of systems, the process may require help of other Cas proteins or host proteins. Because adaptation in different types of CRISPR systems is similar and most of the knowledge of spacer acquisition has been gained from studying the *E. coli* type I-E system, here we focus on the adaptation mechanism of this system. The crystal structure of *E. coli* Cas1-Cas2-prespacer complex reveals that the protospacer is splayed into a double fork structure with the double stranded region lying on the central Cas2 dimer and the 3' single stranded overhangs being inside the Cas1 subunits (Nuñez et al., 2015a) (Figure 3). The crystal structure supports previous work (Nuñez et al., 2014) that Cas2 does not contribute to catalytic activity of spacer integration. The Cas2 subunit forms a large part of the 'arginine clamp' which interacts and stabilises the phosphate backbone of the duplex region of the protospacer. At both ends of the duplex region, there is a tyrosine residue from Cas1 acting as a wedge to stabilise the unwinding and bend the 3' overhang into the Cas1 catalytic active site. In addition to the stacking interaction, the fixed distance between two Cas1 wedges also acts as a molecular ruler to dictate the length of the new spacer. The 3' single stranded overhang is stabilised by an arginine channel in each catalytic Cas1 subunit.

The recognition of CRISPR repeat is accomplished by the Cas1-Cas2-prespacer complex having a strong affinity to the palidromic motif in the repeat. It has been observed that new spacers are strictly inserted into the leader end of CRISPR array *in vivo*, however, *in vitro* study using Cas1-Cas2 complex alone resulted in integration at every CRISPR repeat in type I systems (Nuñez et al., 2015b). Indeed, diverse mechanisms are used to ensure the recognition of the leader-proximal CRISPR repeat

by the Cas1-Cas2 complex in different CRISPR types and subtypes. For example, an integration host factor (IHF) heterodimer binds to the leader sequence in type I-E systems and it sharply bends DNA at the leader region (Nuñez et al., 2016). This recruits the Cas1-Cas2-prespacer complex to the leader-repeat junction by allowing interaction between a non-catalytic Cas1 subunit and upstream leader sequence (Figure 4). It was also shown that the IHF-α subunit might directly interact with a Cas1 subunit and this interaction play a part in efficeint spacer integration (Wright et al., 2017). In type II systems, the recognition of the leader-repeat junction solely relies on a region of intrinsic sequence called leader anchoring site which is immediately upstream of the first CRISPR repeat (Kim et al., 2019b; McGinn and Marraffini, 2016). The orientation of the Cas1-Cas2-prespacer complex is determined by the PAM at the 3' end in one of the prespacer strands and the PAM is detected by the PAM sensing site in a Cas1 subunit (Wang et al., 2015). After the binding to the CRISPR repeat, the hydroxyl group (-OH) of the non-PAM 3' overhang makes the first nucleophilic attack at the leader-repeat boundary to form a half-site intermediate (Figure 4). Then the PAM region of the PAM containing 3' overhang is released by a Cas1 subunit and trimmed by an exonuclease such as DnaQ or ExoT to remove the PAM. This fully processed 3' overhang makes the second nucleophilic attack at the repeat-spacer boundary to form a full-site integration product (Kim et al., 2020). During this process, the CRISPR repeat is duplicated and the single stranded region is repaired by host proteins such as DNA polymerase and ligase (Jackson et al., 2017).

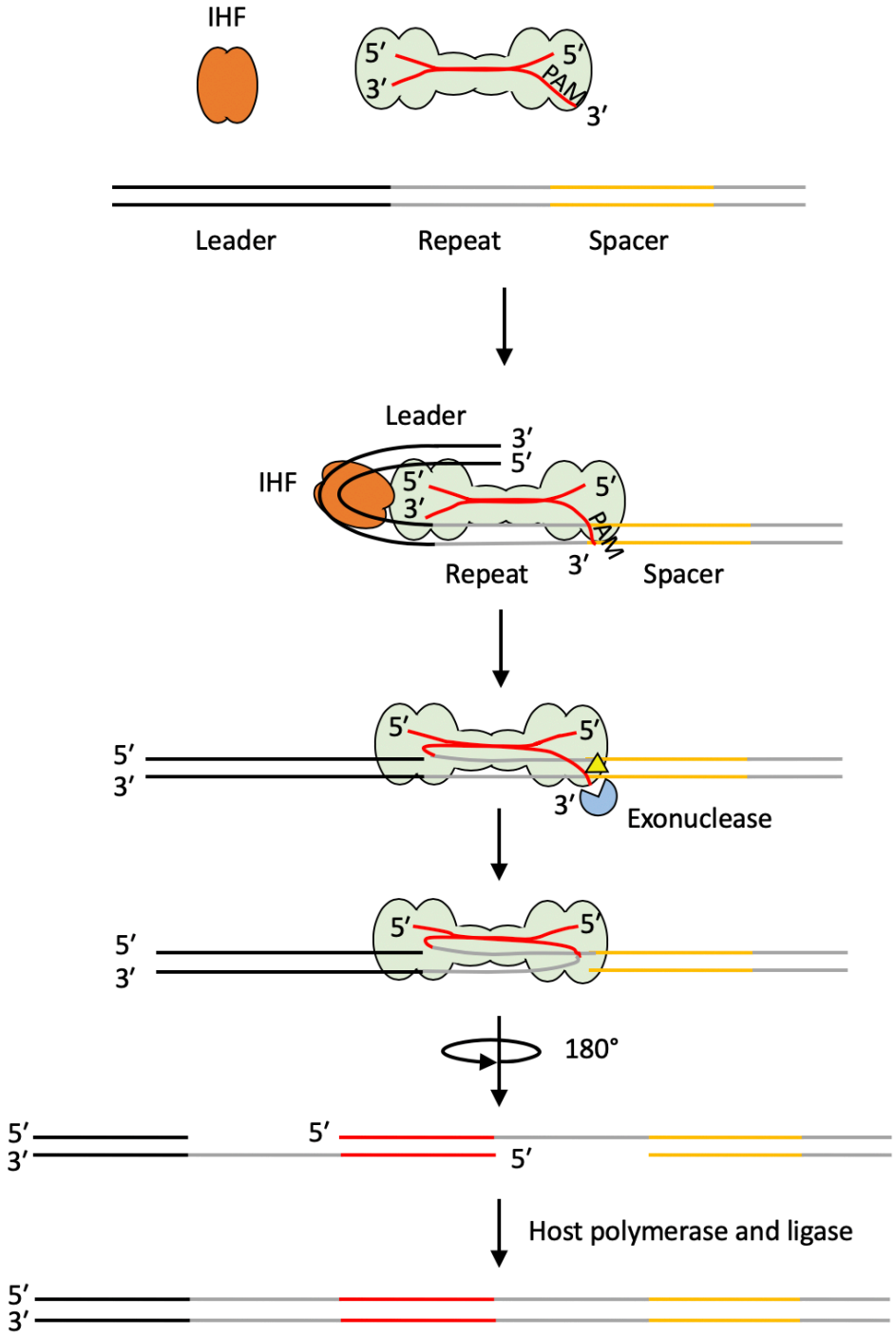


Figure 4. E. coli CRISPR type I-E Cas1-Cas2 catalysed spacer integration.
Schematic diagram showing steps of spacer integration. IHF sharply bends the leader sequence and recruits the Cas1-Cas2 to the leader-proximal repeat. The Cas1-Cas2 complex catalyses the first nucleophilic attack at the leader-repeat junction by the non-PAM 3′ overhang. Then the PAM containing 3′ overhang is further trimmed by an exonuclease and the processed 3′ overhang makes the second nucleophilic attack at the repeat-spacer boundary. During the process, the repeat is duplicated and the single-stranded regions are repaired by host DNA polymerase and ligase. Figure adapted from (Lau et al., 2019).

1.1.3.2. Biogenesis of crRNA

The CRISPR-Cas system is a RNA-guided system that recognises and cleaves specific nucleic acids. The action of nuclease activity of the effector complex requires the matured crRNA which contains a spacer sequence for target recognition and part of repeat seuqnece for effector complex binding. During expression, the CRISPR array is transcribed into a long pre-crRNA from the promoter in the leader sequence. Then the pre-crRNA is processed by endoribonucleases to generate individual functional crRNAs. In most type I and type III systems, the long pre-crRNA is processed by Cas6 proteins (Behler and Hess, 2020). In types I-D, I-E and I-F, palidromic sequences within a CRISPR repeat form a hairpin structure that recruits Cas6 binding (Charpentier et al., 2015; Gesner et al., 2011; Kunin et al., 2007) (Figure 5). Cas6 specifically cleaves at the base of the hairpin located at the 3' end resulting in a crRNA containing 8 nucleotides of the repeat sequence at the 5' end, a complete spacer in the middle and a hairpin structure of repeat sequence at the 3' end (Wakefield et al., 2015). In other type I and type III systems that lack stable hairpin structures in pre-crRNA, the Cas6 protein recognises and cleaves within the repeat sequence to yield individual crRNA. The crRNA is then loaded into the interference effector complex and undergoes further maturation through trimming of the 3' end by an unknown ribonuclease (Charpentier et al., 2015).

In type II systems, the processing of pre-crRNA requires trans-activating crRNA (tracrRNA) which has sequence complementarity to the repeat region of pre-crRNA (Deltcheva et al., 2011). Cas9 promotes the tracrRNA binding to pre-crRNA and this recruits the host RNase III to cleave at the duplex region resulting in a spacer sequence flanked by part of repeat sequence at both ends (Karvelis et al., 2013) (Figure 5). The

5’ end of crRNA is cleaved within the spacer sequence by an unknown nuclease to generate matured crRNA containing 20 nucleotides spacer-derived guide seqeunce and around 20 nucleotides of repeat seuqence at the 3’ end forming a partial duplex structure with the antirepeat sequence on the tracrRNA. The tracrRNA has three stem loop structures that play an important role in stabilising the crRNA-Cas9 complex (Nishimasu et al., 2014). In type V and VI systems, the single interference effector proteins Cas12 and Cas13 also possess catalytic activity for crRNA processing in addition to crRNA-guided target cleavage (East-Seletsky et al., 2016; Swarts et al., 2017).

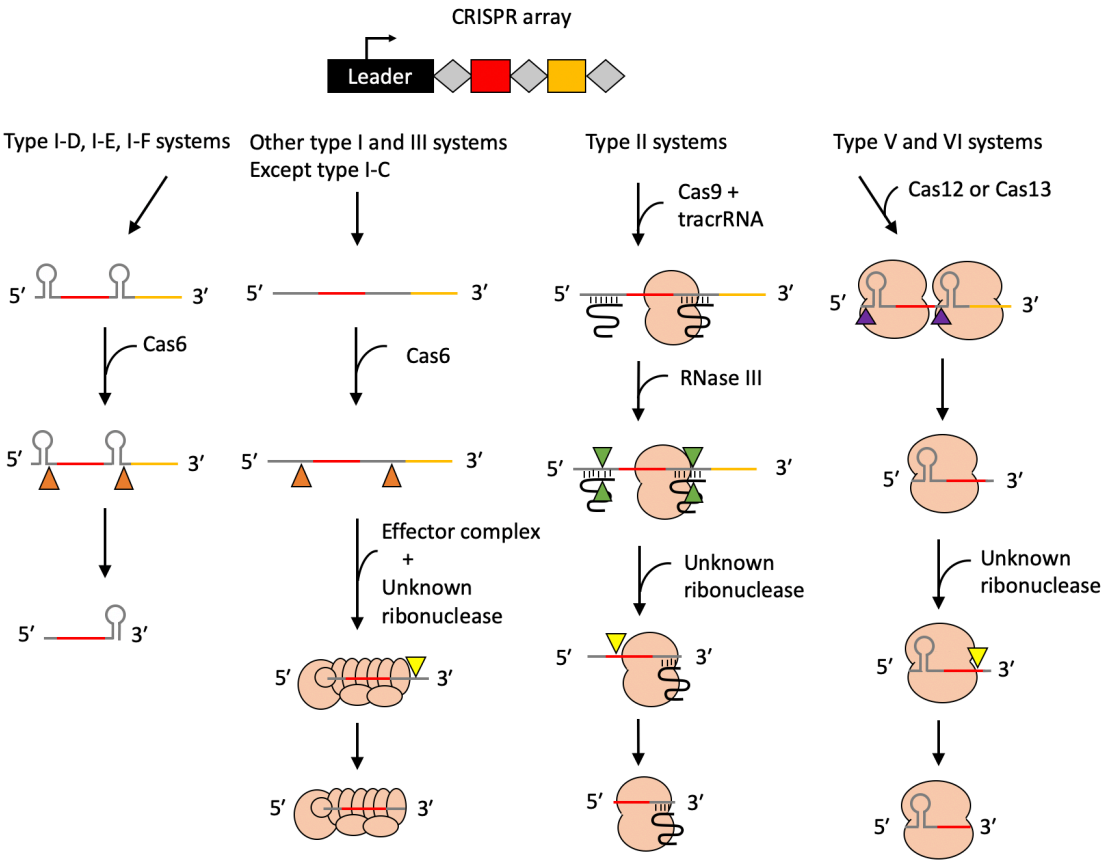


Figure 5. Biogenesis of crRNA in different types of CRISPR-Cas systems.
CRISPR array is transcribed from the leader sequence into a long pre-crRNA. In type I-D, I-E, I-F systems, Cas6 protein cleaves at repeat sequence (orange triangle), which forms a hair pin structure, to generate individual crRNA. In other type I and III systems except I-C, Cas6 cleaves within the repeat sequence (orange triangle). The crRNA is further cleaved at the 3′ end by an unknown nuclease (yellow triangle). In type II systems, tracrRNA binds to the repeat sequence in pre-crRNA facilitated by Cas9 and this triggers cleavage of the RNA duplex region (green triangle) by RNase III. The resulting crRNA is further cleaved by an unknown nuclease at the 5′ end (yellow triangle) to give matured crRNA for interference. In type V systems that lack tracrRNA and in type VI systems, the repeat sequence in pre-crRNA forms a secondary structure that recruits Cas12 or Cas13 effector protein for crRNA processing. The 3′ end of crRNA was processed by unknown nucleases (yellow triangle).

1.1.3.3. Interference

The formation of crRNA-effector complex allows target searching, which is started with recognition of a protospacer adjacent motif (PAM). Single molecule studies have shown that crRNA-effector complexes from types I, II and V search for target DNA via facilitated diffusion (Dillard et al., 2018; Globyte et al., 2019; Jeon et al., 2018).

Facilitated diffusion is a combination of three-dimensional (3D) diffusion and one-dimensional (1D) sliding or hopping along the DNA and it is a common strategy for many site-specific DNA binding proteins such as transcription factors and Argonaute proteins (Cui and Joo, 2019; Halford, 2004). However, a study using single-molecule fluorescence resonance energy transfer (smFRET) revealed that the *E. coli* Cascade complex sampled DNA merely through 3D diffusion (Xue et al., 2017). Another study using *Thermobifida fusca* Cascade, which is also from type I-E systems, showed 1D diffusion along the DNA and identified a Cas8 protein in the Cascade complex is important in facilitated diffusion (Dillard et al., 2018). A structure-based multiple sequence alignment showed that the *T. fusca* Cas8 protein contains a positive channel formed by conserved lysine/arginine residues on the outer surface and this positive channel is disrupted in *E. coli* Cas8, probably explaining the discrepancy in DNA searching of these two Cascade complexes. After the optimal PAM binding, crRNA forms duplex with the target strand leading to the destruction of MGEs. The PAM recognition precedes the crRNA-target duplex formation, thus preventing from targeting the CRISPR locus in the host's genome (Xiao et al., 2017).

The CRISPR-Cas systems are divided into six types due to the diversity of effector complexes during the interference stage. The effector complex of types I, III and IV systems consist of multiple Cas proteins while types II, V and VI systems employ a single Cas protein with multiple domains. In addition to the architecture of the effector, the diversity of CRISPR types comes from the target recognition and mechanism of action. For instance, types I, II and V recognise DNA but types III and VI recognise RNA targets (Jackson et al., 2017). Here, we focus on CRISPR-Cas systems that target and degrade DNA, i.e. types I, II and V systems.

In *E. coli* type I-E system, Cas6 protein remains bound to the 3' hairpin structure of the crRNA after crRNA maturation and this may help the recruitment of other Cas proteins to form the Cascade complex. Six Cas7 proteins polymerise along the crRNA and a Cas5 protein caps the 5' end of the crRNA (Jackson et al., 2014). The surveillance complex also includes two small subunits (Cse2 in *E. coli*) lying in the middle of the complex and one large subunit Cas8, also known as Cse1, next to the Cas5 protein. A crystal structure of crRNA-Cascade complex bound to dsDNA revealed that the PAM in the DNA is recognised in double stranded form and Cas8 is responsible for PAM recognition (Hayes et al., 2016) (Figure 6). Upon optimal PAM recognition, a glutamine wedge of Cas8 is inserted underneath the PAM sequence causing the bending and unwinding of dsDNA. The same research group later presented two cryo-electron microscopy (cryo-EM) snapshots of crRNA-Cascade complex bound to dsDNA showing an R-loop intermediate and a full R-loop formation (Xiao et al., 2017). In the R-loop intermediate snapshot, only 11 base pairs (bp) next to a PAM sequence in the target DNA are unwound forming a partial R-loop with crRNA compared with 32 bp in full R-loop formation. This PAM-proximal region contains the seed sequence that determines the specificity of CRISPR systems. Single point mutations inside the seed region can severely decrease the affinity of the crRNA-complex to the target DNA while mutation outside the seed region can be tolerated (Semenova et al., 2011). When the seed region matches to the crRNA, the R-loop propagates and extends to form a full R-loop. Upon the full R-loop formation, the Cascade complex undergoes conformational changes by rotating the Cas8 large subunit and sliding the Cse2 dimer. This recruits a Cas3 protein which nicks the non-target strand of the R-loop using a HD nuclease domain (Figure 6). Then a helicase domain of the Cas3 utilises ATP to unwind

the dsDNA by a DNA reeling mechanism while the Cas3 remains bound to the Cascade complex. The helicase domain feeds the single-stranded non-target strand to the HD nuclease domain, which degrades the ssDNA in a 3′ to 5′ direction (He et al., 2020; Huo et al., 2014; Loeff et al., 2018).

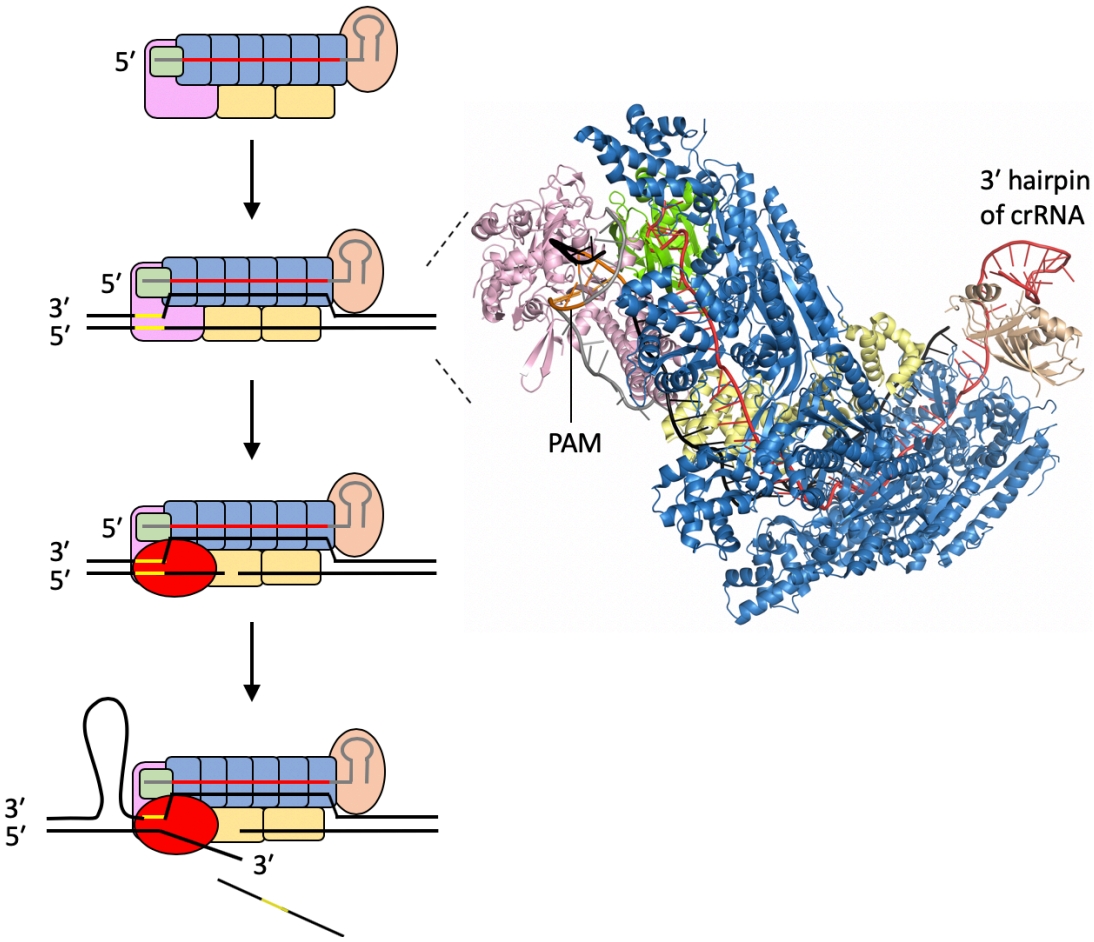


Figure 6. Target DNA cleavage during interference stage of CRISPR type I-E system in E. coli.
In *E. coli* type I-E system, the DNA surveillance Cascade complex comprises of Cas5 (green), Cas6 (orange), six Cas7 (blue), Cas8 (pink), two Cse2 (yellow) proteins and crRNA. After identifying the correct PAM (yellow line) and checking the target sequence complementary to the spacer in crRNA, a full R-loop is formed. This recruits the helicase/nuclease Cas3 (red), which nicks the non-target strand for binding. The helicase domain of Cas3 unwinds dsDNA and degrades the non-target strand intermittently. A crystal structure of *E. coli* Cascade forming a full R-loop with target DNA (PDB: 5H9E) is shown on the right with the same colour code of Cascade described in the schematic diagram except the non-target strand being grey and crRNA being red. The target DNA was unwound after the PAM (orange) and the target strand (black) base paired with the crRNA (red).

In type II systems, the RNA-guided target recognition and degradation are accomplished by a single protein, Cas9. The *Streptococcus pyogenes* Cas9 (SpCas9) is the best characterised Cas9 protein in type II systems and its structural studies provides extensive insights into the molecular mechanisms of RNA-guided DNA binding and cleavage. Cas9 has a bilobed structure comprising of a recognition (REC) lobe and a nuclease (NUC) lobe connected by an arginine-rich bridge helix (Nishimasu et al., 2014) (Figure 7A). The REC lobe consists of three alpha helical domains (REC1-3) and it is responsible for nucleic acid recognition. The NUC lobe is formed by an HNH nuclease domain, a RuvC nuclease domain and a PAM-interacting (PI) domain that recognises 5'-NGG-3' PAM in SpCas9.

Different Cas9 orthologs recognise different PAM sequences and their PI domains share a similar protein fold despite of low sequence similarity. Structural studies of these Cas9 orthologs revealed that different PAM specificities are caused by different PAM-interacting residues in the PI domain (Nishimasu et al., 2015). By comparing the structure of apo-Cas9 with single guide RNA (sgRNA, crRNA connected with tracrRNA by a tetranucleotide loop) bound Cas9, it depicts that Cas9 undergoes drastic conformational changes upon binding to the sgRNA. A cleft appears between the REC lobe and NUC lobe for accommodating the sgRNA and the position of REC1-3 domains is rearranged. Cas9 makes extensive interactions with the repeat-antirepeat partial duplex and stem loop 1 in the sgRNA. The stem loops 2 and 3, on the other hand, make fewer contacts with Cas9 (Jiang and Doudna, 2017). Experimental evidence shows that mutations changing the structure of the repeat-antirepeat duplex or stem loop 1 significantly decrease the Cas9 function whereas mutations or deletion of stem loops 2 and 3 only marginally affect the Cas9 function (Jinek et al., 2012; Nishimasu et al.,

2014). This suggests stem loops 2 and 3 are not essential to the Cas9 function but they stabilise the sgRNA-Cas9 complex.

The binding of sgRNA turns Cas9 from an inactive state to an active DNA surveillance complex. A crystal structure of sgRNA-Cas9 complex illustrates that the seed region of sgRNA is preordered into an A-form conformation facilitating the formation of RNA-DNA heteroduplex (Jiang et al., 2015a) (Figure 7B). In addition, the two arginine residues in the PI domain that interacts with the 5'-NGG-3' PAM are pre-positioned upon sgRNA binding, thus making Cas9 ready for target DNA searching and binding. Similar to type I systems, Cas9 first scans for the correct PAM to rapidly identify potential targets from a long DNA molecule. Then the dsDNA is bent and unwound from the seed region and a perfect complementarity between the seed region and target DNA leads to a full R-loop formation (Jiang et al., 2016). Upon the full R-loop formation, Cas9 undergoes conformational changes which position the HNH active site towards the target strand. As the HNH domain connects to the RuvC domain via a loop linker, the linker also undergoes conformational changes bringing RuvC active site close to the non-target strand (Figure 7B). Both nuclease domains cleave their corresponding strand of the dsDNA 3 bp from the PAM site giving rise to a blunt ended DSB (Jiang and Doudna, 2017; Nishimasu and Nureki, 2017).

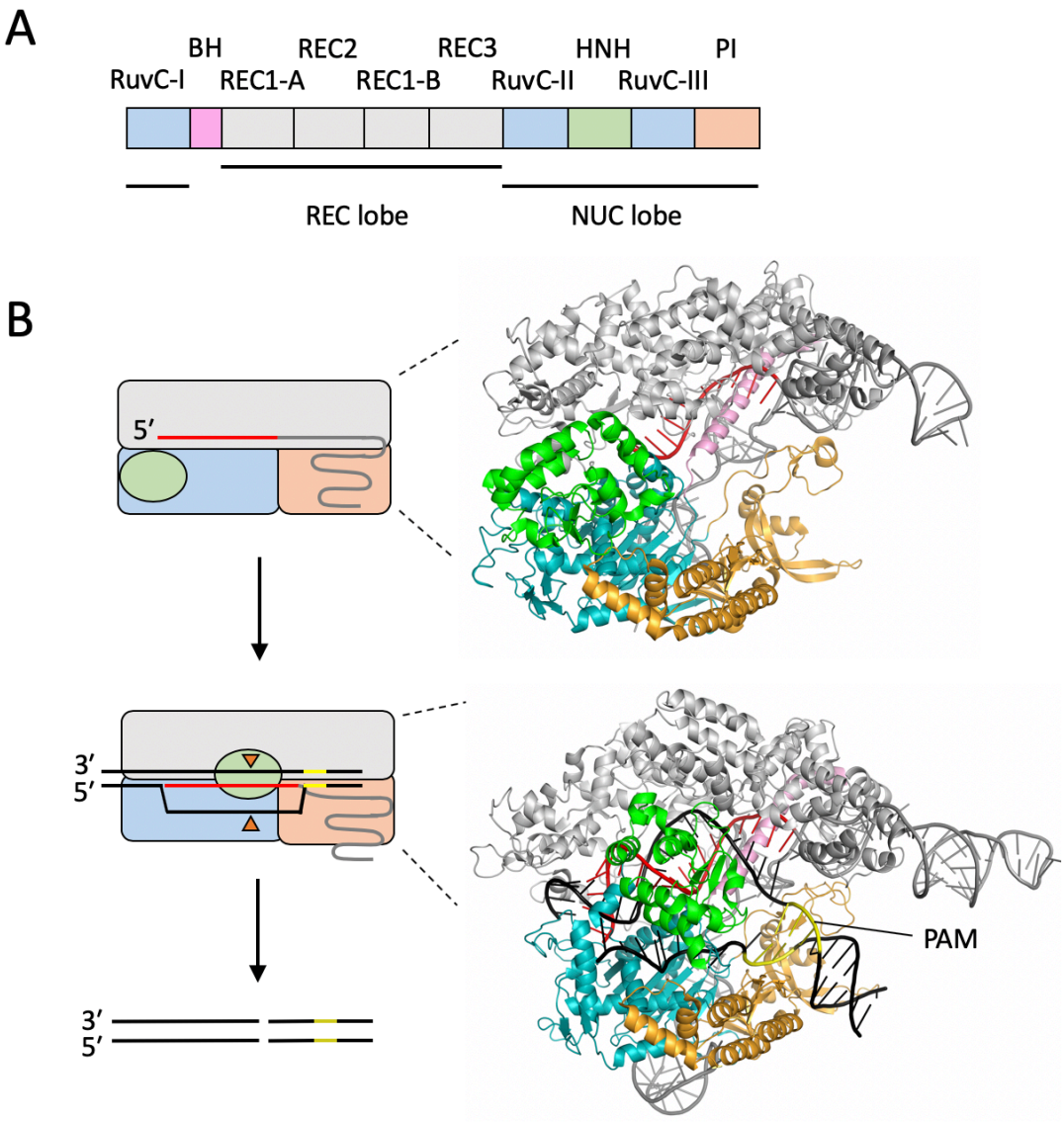


Figure 7. SpCas9 catalysed target DNA cleavage in CRISPR type II-A system.
(A) Domain organisation of SpCas9 containing a REC lobe and a NUC lobe connected by a bridge helix (BH, pink). In type II systems, the interference effector only contains one Cas9 protein. SpCas9 contains multiple domains: REC1-3 (grey), HNH (green), RuvC (blue) and PI (orange). (B) Schematic diagram showing sgRNA-Cas9 forms an R-loop with PAM containing target DNA and cuts the target DNA. The colour code of each domain is the same as in (A) and crystal structures at different steps are shown on the right. In the sgRNA-Cas9 binary structure (top right, PDB: 4ZT0), only the seed region of guide RNA is shown because the non-seed nucleotides were disordered. In the sgRNA-Cas9-dsDNA ternary structure (bottom right, PDB: 5F9R), it illustrates that the PAM (yellow) in target DNA (black) is recognised in duplex form and the HNH nuclease domain is placed close to the target DNA strand. This conformational change activates the HNH and RuvC nuclease domains that cleave the target strand and non-target strand respectively creating a blunt-ended DSB 3 bp upstream of the PAM.

In CRISPR-Cas type V systems, subtype V-A is the best characterised and studied most. The effector protein Cas12a (formerly known as Cpf1) contains a bilobed structure formed by a REC lobe and a NUC lobe similar to Cas9 in type II systems. The N-terminal REC lobe consists of the REC1 and REC2 domains. The C-terminal NUC lobe consists of the RuvC, wedge (WED), PI and target strand loading (TSL) domains (Dong et al., 2016) (Figure 8A). The TSL domain is also known as nuclease (NUC) domain but this domain is actually responsible for presenting the target strand to the catalytic site rather than target strand cleavage (Liu et al., 2019).

After transcription of pre-crRNA, the repeat region in each crRNA forms a pseudoknot structure recruiting Cas12a and leading to subsequent processing and maturation of crRNA in the WED domain in Cas12a (Swarts et al., 2017). In the Cas12a-crRNA complex, a seed sequence containing five PAM-proximal nucleotides is preordered into A-form facilitating target DNA binding. The target DNA binding is initiated by recognition of a T-rich PAM in Cas12a (Yamano et al., 2017). A crystal structure of *Francisella novicida* Cas12a showed that a loop-lysine helix-loop region in the PI domain contains three conserved residues, K667, K671 and K677, which involve in PAM recognition and target DNA unwinding immediately after the PAM (Stella et al., 2017). The unwound target DNA strand forms a partial R-loop with the seed sequence in crRNA acting as a checkpoint as mismatches in the seed sequence are intolerant (Fonfara et al., 2016). The full complementarity between the seed sequence and target strand enables the formation of a full R-loop, which induces conformational change to allosterically activates Cas12a (Stella et al., 2018). Three protein segments called ‘REC linker’, the ‘lid’ and ‘REC finger’ start to interact with crRNA when R-loop is formed. Notably, in Cas12a-crRNA complex, the ‘lid’ in the RuvC domain contains

residues interacting with the catalytic residues, thus restricting access of DNA substrates to the catalytic site. Once a stable R-loop is formed, the 'lid' switches to interact with crRNA to open the catalytic site and other residues in the RuvC domain interact with the catalytic residues to stabilise the open conformation of the catalytic site.

Cas12 relies on one nuclease domain, RuvC, to generate a DSB at target DNA so the cleavage of target strand and non-target strand takes place sequentially. When the target DNA forms a R-loop with crRNA, the displaced non-target strand is stabilised by residues in the PI domain and by a positively charged groove in the RuvC domain that guides the non-target strand to the catalytic site (Stella et al., 2017). Therefore, the non-target strand is cleaved first (Stella et al., 2018; Swarts and Jinek, 2018) (Figure 8B). Cas12 cleaves the target strand downstream of the R-loop generating a staggered DSB with 5-nt 5' overhangs. To achieve this, the target DNA downstream of the R-loop needs to unwind probably by fraying of the nicked DNA (Ferreira et al., 2015) and structures of Cas12b and Cas12e (CasX) showed that the target strand is bent sharply to enter the catalytic site mediated by the TSL domain (Liu et al., 2019; Yang et al., 2016). After generating a DSB, Cas12a releases the cleaved PAM-distal target DNA while it remains bound to the PAM-proximal cleaved product (Stella et al., 2018). Because the 'REC linker', 'lid' and 'Rec finger' are still interacting with the partial R-loop, the catalytic site remains open and is solvent exposed to allow access of ssDNA. This accounts for a non-specific single-stranded DNase activity of Cas12a upon target binding (Chen et al., 2018). The ssDNA *trans*-cleavage activity of Cas12a may help the cell to degrade MGEs during transcription and clear ssDNA phages more rapidly. It was shown that new crRNA displaces the partial R-loop in Cas12a to terminate the non-

specific ssDNA *trans*-cleavage activity and reset the Cas12a to target new DNA (Stella et al., 2018).

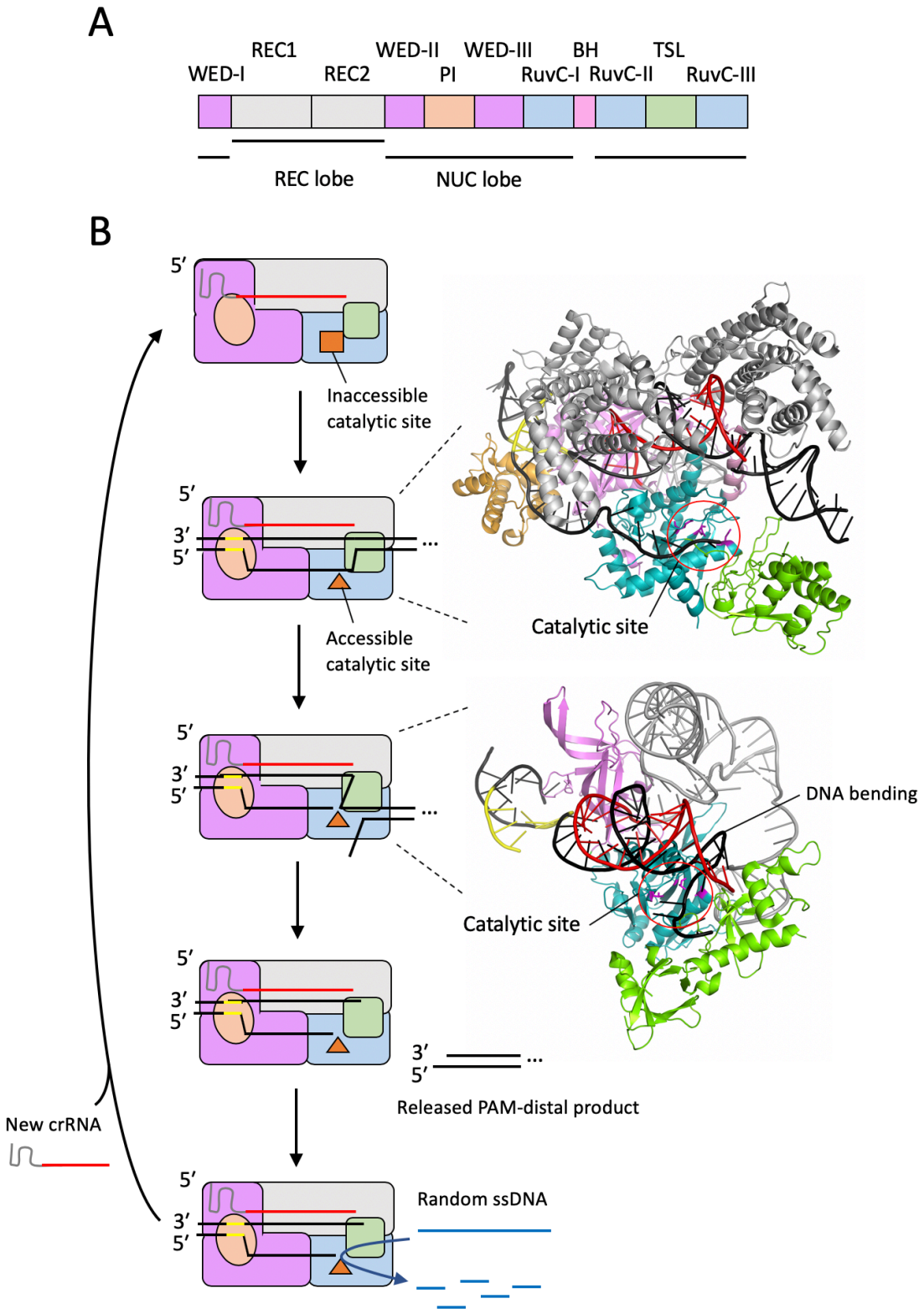


Figure 8. Cas12 catalysed cis- and trans-cleavage of ssDNA in CRISPR type V systems. (A) Domain organisation of Cas12a containing a REC lobe and a NUC lobe connected by a bridge helix (BH, pink). In type V systems, the interference effector only contains one Cas12 protein. Cas12a contains multiple domains: REC1-2 (grey), WED (purple), RuvC (blue), PI (orange) and TSL (green). (B) Schematic diagram showing target binding activates Cas12a *cis*- and *trans*-cleavage of ssDNA and recycling of Cas12a by new crRNA. The colour code of each domain is the same as in (A). The Cas12-crRNA forms an R-loop with target DNA. A crystal structure of *F. novicida* Cas12a forming R-loop with dsDNA is shown and the RuvC catalytic residues are shown as sticks and highlighted with a red circle (top right, PDB: 6I1K). The formation of the full R-loop induces conformational changes to activate the RuvC nuclease domain, which cleave the non-target strand first. Then the target strand bends sharply to enter the RuvC catalytic site for cleavage. A crystal structure of *Alicyclobacillus acidoterrestris* Cas12b-crRNA-tracrRNA binding to a target strand illustrates that the target strand bends sharply to enter the RuvC catalytic site highlighted with a red circle (bottom right, PDB: 5U30, the REC domain is omitted for clarity). Cas12 generates a staggered cut with 5' overhangs and the PAM-distal cleaved product is released. Cas12 remains associated with the PAM-proximal R-loop and stays activated. The accessible catalytic site catalyses ssDNA cleavage non-specifically and this is stopped by new crRNA displacing the R-loop.

1.1.4. Evolutionary origin of CRISPR systems

Given that class 1 and class 2 CRISPR-Cas systems contain fundamentally different effector modules, it is believed that they have evolved independently and through repeatedly recruiting genes from different MGEs to contribute to adaptive immunity (Koonin and Makarova, 2019). For instance, casposons from a novel DNA transposon family have given rise to the highly conserved adaptation module and CRISPR repeats (Krupovic et al., 2017). Casposon transposition is catalysed by a Cas1 homologue called casposase, which is suggested to be the ancestor protein of Cas1. Upon insertion of an ancestral casposon, it generates a 15-bp target site duplication (TSD) that serves as the ancestral CRISPR repeats (Hickman and Dyda, 2015). Casposon is described more in section 1.2.2.

In class 1 systems, it is evident that type I and type III systems share common ancestry because of the architecture of the effector complex. The Cascade complex in type I

and Csm or Cmr complexes in type III all consist of multiple Cas7 subunits as backbone, a Cas5 protein, small subunits and a large subunit (Nishimasu and Nureki, 2017). It is suggested that type III systems are more related to the ancestral state of CRISPR systems and have given rise to types I and IV because they shows complex gene composition containing diverse ancillary genes (Makarova et al., 2020). In addition, this assumption fits a model that the ancestor of the CRISPR-Cas effector module originated from a stress response system analogous to abortive infection systems (Short et al., 2018). This putative stress response system contained a Cas10 homologue containing a cyclic oligoA polymerase palm domain (a variant of RNA recognition motif, RRM) and proteins containing one or both domains of CRISPR-associated Rossman fold (CARF) and a Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domain (Koonin and Makarova, 2019). This system was proposed to function analogously to the signal transduction pathway identified in most of the type III systems. The details of the signal transduction pathway are described in section 1.1.2. In brief, cyclic oligoadenylate (cOA) molecules produced by the palm domain of the Cas10 homologue in response to phage infection was bound by the CARF domain containing protein leading to allosteric activation of the indiscriminate RNases activity from the HEPN domain (Kazlauskiene et al., 2017; Niewoehner et al., 2017). This would then result in cell growth arrest or programmed cell death (Rostøl and Marraffini, 2019). The ancestral class 1 CRISPR-Cas system was proposed to result from the merging of two modules: the adaptation module along with CRISPR repeats from immobilised casposon and the type III-like effector module from the stress response system. Then the system evolved through domain duplication and divergence of the RRM domain in the Cas10 homologue and recruitment of more ancestral *cas* genes

from MGEs to give rise to ancestral type III systems. Then the ancestral type III systems lost the signal transduction pathway and gave rise to type I systems by recruiting Cas3 helicase and nuclease and replaced Cas10 with Cas8 as the large subunit. Type IV systems probably originates from multiple gene loss of the ancestral type III systems (Koonin and Makarova, 2019; Makarova et al., 2020).

In class 2 CRISPR-Cas systems, although Cas9 from type II and Cas12 from type V systems share a RuvC domain, this domain shows very low sequence similarity between the two proteins. Outside of this RuvC domain, the two proteins shares no sequence similarity so they are not homologous and have evolved independently. Even within the same type, it was proposed that Cas12 from different subtypes have evolved independently from different RuvC domain containing TnpB nucleases encoded by IS605-like transposons (Shmakov et al., 2015). This is because Cas12 from different subtypes contains unrelated N-terminal regions and distinct target strand loading (TSL) domains. For type II systems, Cas9 was found homologous to IscB proteins encoded by a distinct family of IS605 transposons called insertion sequences Cas9-like (ISC) (Chylinski et al., 2014; Kapitonov et al., 2016). The IscB proteins are TnpB-like proteins with an HNH nuclease domain inserted into a RuvC domain and they are the likely ancestors of Cas9.

1.2. Casposon

1.2.1. Transposable elements

Transposable elements (TEs) are DNA sequences that are capable of moving from one location to another location within a genome. They are found in virtually all organisms and can make up a large fraction of the genome. For example, 45% of the human

genome is comprised of TEs (Munoz-Lopez and Garcia-Perez, 2010). The movement of TEs within a genome creates genetic diversity of the species and TE insertions can influence the host's fitness by inserting into a gene or affecting expression of surrounding genes. TEs have helped shaping genome architecture and evolution by causing chromosomal rearrangements, insertions, deletions and horizontal gene transfer (Chandler, 2017). TEs can be divided into two classes according to their mechanisms of transposition. Class 1 TEs, also known as retrotransposons, mobilise via a 'copy and paste' mechanism that is catalysed by reverse transcriptases. Retrotransposons are first transcribed into an RNA intermediate, which is then reverse transcribed to cDNA for insertion elsewhere in the genome. Class 2 TEs, also known as DNA transposons, mostly mobilise via a 'cut and paste' mechanism that does not involve an RNA intermediate and is catalysed by transposases (Bourque et al., 2018). During transposition, transposases are recruited to terminal inverted repeats at both ends of the transposon and excise the transposon from the genome. Then the transposon bound transposase makes a stagger cut at a target site elsewhere in the genome and ligates the transposon to the sticky ends of the target site. Finally DNA polymerase fills the gap generating target site duplication.

1.2.2. Casposon gave rise to the CRISPR adaptation module and CRISPR repeats

Spacer acquisition is the foundation of CRISPR-Cas adaptive immunity and Cas1 is required in this step. As Cas1 is so important for the system, it was thought that Cas1 was exclusive to the CRISPR-Cas system and Cas1 was used as a marker to identify new CRISPR-Cas systems. However, a comparative genomic study discovered that two groups of Cas1 homologues are not associated with any CRISPR systems (Makarova et

al., 2013). It was subsequently found out that one group of stand-alone Cas1 homologues are associated with family B DNA polymerases (PolBs). By analysing the genomic loci containing *cas1* homologue and *polB*, it revealed that these loci are flanked by terminal inverted repeats (TIRs) and target site duplication (TSD) (Figure 9A), which are the hallmark of integrated DNA transposons. Thus, these Cas1 homologue containing DNA transposons are termed 'casposons' and the authors put forward that casposon played a crucial role in the origin of CRISPR-Cas system (Krupovic et al., 2014).

It was then experimentally shown that the Cas1 homologue encoded by an *Acidoprofundum boonei* casposon integrated TIR-derived substrates and the results suggest that casposons utilise these Cas1 homologues for transposition. These Cas1 homologues are therefore called 'casposases' (Hickman and Dyda, 2015). Indeed, casposase catalyses the insertion of casposon in a similar way to spacer insertion catalysed by the Cas1-Cas2 adaptation complex in CRISPR-Cas systems (Figure 9B). It was shown that casposase has a target site preference which requires a 24-30 bp sequence upstream of the target site and this resembles the requirement of the leader sequence upstream of the CRISPR array in CRISPR-Cas systems (Béguin et al., 2019; Hickman et al., 2020). This 'leader' sequence varies according to different casposon families, which will be discussed in more details in section 1.2.4. For family 2 casposons, the 'leader' sequence lies in the 3' distal region of different tRNA genes (Krupovic et al., 2017). This 'leader' sequence recruits casposase to the target site and defines the 'leader'-target site boundary in the first nucleophilic attack by the 3'-OH group of the casposon catalysed by casposase (Béguin et al., 2016). The second nucleophilic attack occurs at the 'leader'-distal region a fixed distance from the first

attack and this distance is dictated by the spanning sequence of casposase (Hickman and Dyda, 2015). Casposase generates a TSD upon casposon integration which resembles the CRISPR repeat duplication during adaptation. The resemblance of the two systems emphasises their evolutionary relationships that CRISPR-Cas systems evolved from an immobilised ancestral casposon.

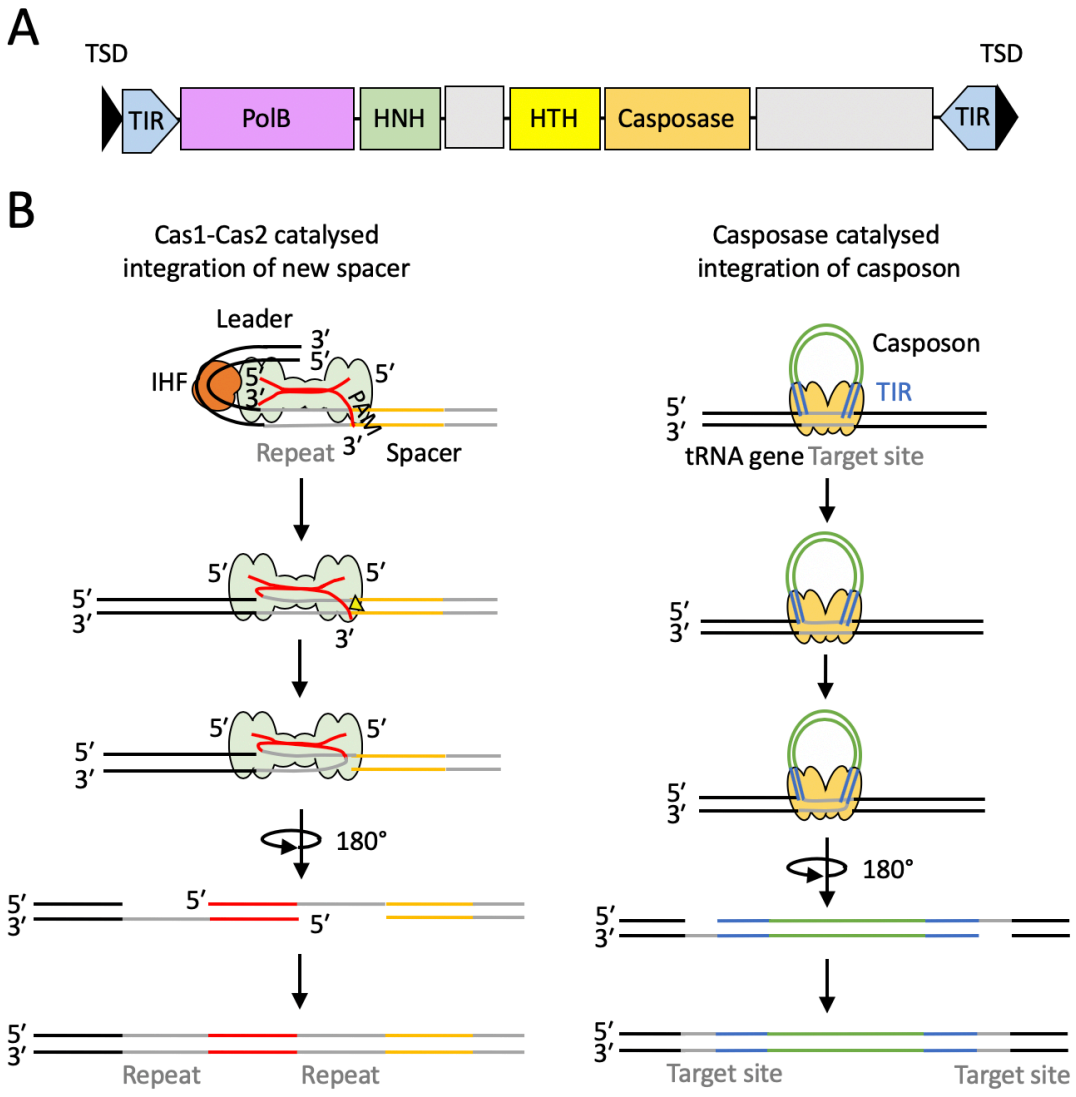


Figure 9. Casposase catalyses casposon integration similar to spacer acquisition in CRISPR-Cas system.

(A) Schematic representation of genome organisation of a general family 2 casposon. TSD: target site duplication. TIR: terminal inverted repeat. PolB: family B DNA polymerase. HNH: HNH endonuclease. HTH: helix-turn-helix protein. (B) Comparison of Cas1-Cas2 complex catalysed spacer integration in *E. coli* CRISPR-Cas type I-E system

(left) with casposon integration catalysed by casposase from family 2 casposons (right). The Cas1-Cas2 complex is recruited to the leader-proximal repeat by IHF bending the leader sequence. Casposase is recruited at a target site downstream of the 3′ end of a tRNA gene by sequence recognition. The first nucleophilic attack occurs at the leader-repeat boundary in the CRISPR-Cas system and at the tRNA-target site boundary in casposon integration. The second nucleophilic attack occurs at the opposite strand of the target DNA distal to the leader or tRNA gene. The ssDNA gaps are repaired by host polymerase resulting in duplication of the CRISPR repeat or target site. Adapted from Lau, C. H., Reeves, R., & Bolt, E. L. (2019). Adaptation processes that build CRISPR immunity: creative destruction, updated. *Essays in biochemistry*, 63(2), 227-235.

1.2.3. Comparison of casposase structure to Cas1 structure

A recent study revealed a X-ray structure of *Methanosarcina mazei* casposase bound to DNA mimicking post-integration product (Hickman et al., 2020) (Figure 10A). The DNA used in crystallography contained a 14-bp double-stranded target region and two 8-nt 5′ overhangs at each side representing post-nucleophilic attacks by 3′-OH group of casposon ends. Casposase forms a tetramer around the DNA and two single-stranded casposon ends are in a channel leading to the active site in two catalytic casposases (Figure 10A). When the post-integration structure of casposase is compared with that of CRISPR-Cas type I-E Cas1-Cas2 complex in *E. coli* (Figure 10B), the mode of DNA binding by casposase resembles that of CRISPR Cas1 but they differ that Cas2 is required by Cas1 to stabilise the binding of double-stranded region of integrating spacer (Hickman et al., 2020). When the catalytic monomers of the two structures are compared, the C-terminal α-helical domains of two proteins align very well and the single-stranded casposon end after the nucleophilic attack resembles single-stranded region of the integrating spacer (Figure 10C). The conserved catalytic residues of the two proteins are in the same position relative to the covalently joined single-stranded DNA (Figure 10C and Appendix 2). The non-catalytic subunits of casposase interacts with the target DNA and the interacting residues mostly locates in

the C-terminal region (Figure 10D). This interaction between target DNA and non-catalytic subunits is also observed in CRISPR Cas1 and most of the interacting residues are in the C-terminal region too (Figure 10E). The Cas2 dimer that bridges two Cas1 dimers interacts with both the duplex region of the integrating spacer and the target DNA. It seems that in the evolution of CRISPR adaptation module, the recruitment of Cas2 into the ancestral Cas1 tetramer does not only fix the spacer length by the molecular ruler mechanism, but it also increases the length of target site. This is supported by the fact that CRISPR repeats are generally longer than target sites of casposases (Hickman and Dyda, 2014).

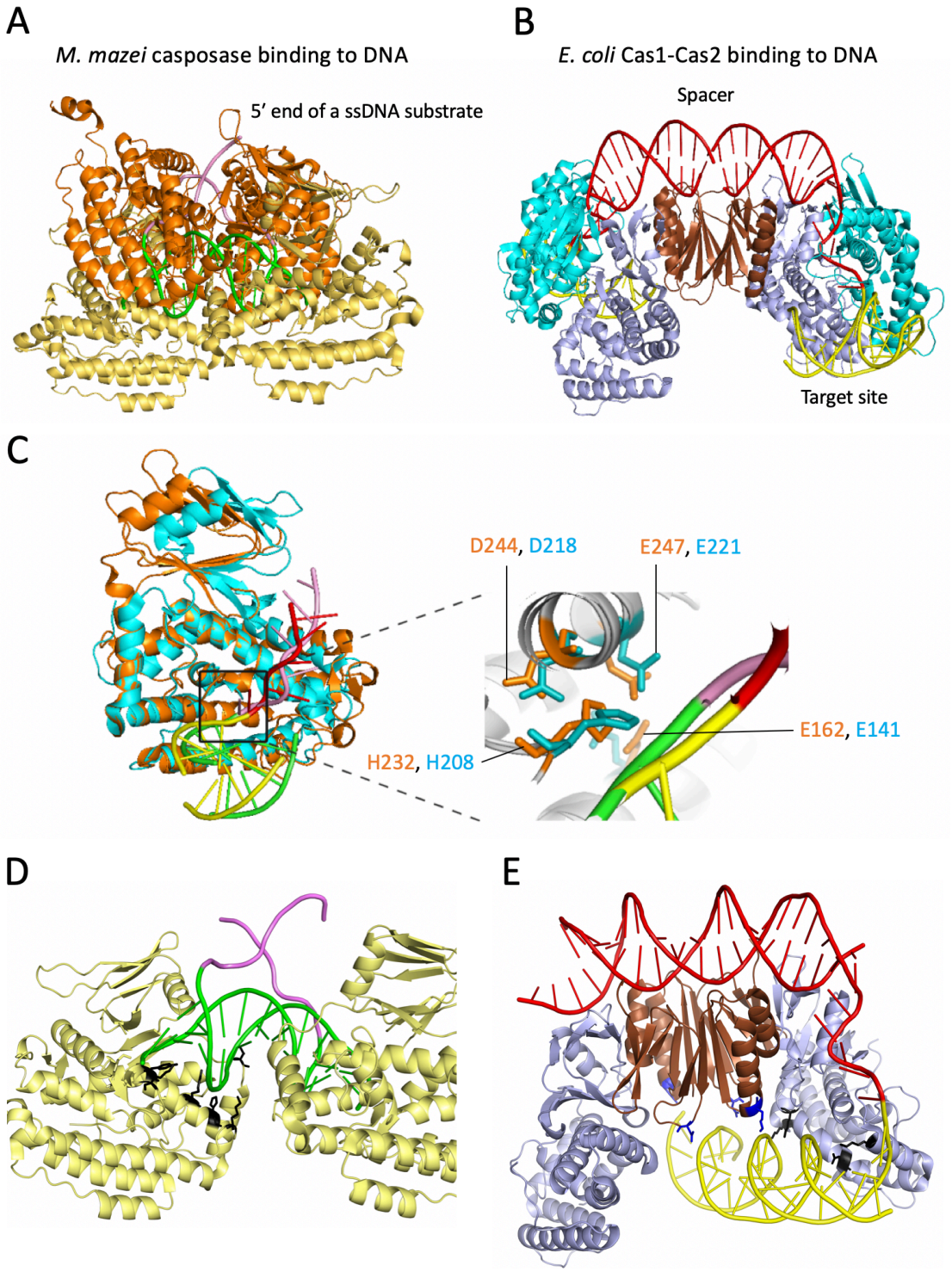

A
M. mazei casposase binding to DNA
5′ end of a ssDNA substrate
B
E. coli Cas1-Cas2 binding to DNA
Spacer
Target site
C
D244, D218
E247, E221
E162, E141
H232, H208
D
E

Figure 10. Structure of M. mazei casposase and comparison to E. coli CRISPR Cas1
(A) *M. mazei* casposase forms tetramer around DNA mimicking post-integration complex (PDB: 6OPM). Orange subunits are catalytic casposases and yellow subunits are non-catalytic. The DNA contains double-stranded target site region (green) and 5' overhangs representing casposon ends after nucleophilic attacks (pink). (B) *E. coli* CRISPR type I-E Cas1-Cas2 hexamer bound to DNA mimicking full-site integration (PDB: 5VVK). The brown object is a Cas2 dimer. Cyan subunits are catalytic Cas1 and light blue subunits are non-catalytic. The DNA contains integrating spacer (red) with ssDNA joining part of the target site (yellow) representing post-integration state. (C) Alignment of DNA (pink-green) bound catalytic casposase subunit (orange) from (A) to DNA (red-yellow) bound catalytic Cas1 subunit (cyan) from (B). Right, enlarged view superimposing of catalytic casposase active site residues (orange) with the catalytic Cas1 residues (blue). Pink and red parts of DNA represent nucleophilic attack strand and green and yellow parts of DNA represent target DNA strand. (D) Non-catalytic casposase subunits in the casposase tetramer interact with target DNA. The interacting residues in a non-catalytic casposase subunit in (A) are shown as sticks and coloured black. The colour code is the same as (A). (E) Non-catalytic Cas1 subunits and a Cas2 dimer in the Cas1-Cas2 complex interact with target DNA (PDB: 5VVJ) using the same colour code as (B). The interacting residues are shown as sticks and coloured black in a non-catalytic Cas1 and blue in the Cas2 dimer. Images were created using PyMOL.

1.2.4. Classification of casposons

All casposons encode casposase and DNA polymerase PolB. Because of this, they are suggested to be self-synthesising DNA transposons but it lacks experimental evidence showing the function of PolB in casposons. Casposons are divided into four families based on gene contents, taxonomic distribution and casposase phylogeny (Krupovic et al., 2017). Family 1 casposons are distinct from the other three families that they are exclusively found in the archaeal phylum *Thaumarchaeota* and the PolBs they encode are protein-primed PolBs that are closely related to those in archaeal viruses (Krupovic et al., 2014). One defining feature of family 2 casposons is that casposases they encode contain additional C-terminal helix-turn-helix (HTH) domain compared with those in other families (Hickman and Dyda, 2015). Family 2 casposons are widespread in Euryarchaeota and family 3 is exclusive to bacteria. The PolBs encoded

by both families are RNA-primed PolB (Makarova et al., 2014). Family 4 casposons are newly identified in some species of euryarchaeon *Methanosarcina mazei* that also contains family 2 casposons (Krupovic et al., 2016). Apart from casposase and PolB, families 2, 3 and 4 also share genes encoding an HNH endonuclease and a protein containing an HTH domain. The function of the HNH endonuclease and HTH protein remain elusive.

Casposons from different families are integrated into different genomic loci representing different target site preferences of the encoding casposases. Family 1 casposons are found in the 3' region of the gene encoding aEF-2. Both families 3 and 4 casposons target intergenic regions. Family 2 casposons are found in intergenic regions or 3' end of tRNA genes (Krupovic et al., 2017). Biochemical characterisation of casposases from family 2 casposons illustrated that *A. boonei* casposase specifically targets the 3' end of tRNA-Pro gene and *M. mazei* casposase targets the tRNA-Leu gene (Béguin et al., 2016, 2019; Hickman et al., 2020).

1.3. Genome engineering

1.3.1. Gene targeting mediated by homologous recombination

Genetics research employs two opposing strategies of forward and reverse genetics, which has significantly improved our knowledge of the roles of genes in different processes such as cellular metabolisms, development and diseases. Forward genetics involves linking of the observed phenotype to specific gene mutation(s), thus determining the function of gene(s). Reverse genetics requires certain prior knowledge of a gene such as DNA sequence and involves introducing user-defined mutations into specific sites of the gene to cause a phenotypic change. From that, the

function of the gene or of a domain in the gene is deduced (Gurumurthy et al., 2016). While forward genetics works in situations of a phenotype controlled by a gene, reverse genetics is advantageous over forward genetics in more complexed phenotype that is governed by more than one gene.

The understanding of DNA repair by homologous recombination (HR) has built the foundation of reverse genetics. It was shown that the *his3* gene in yeast, *Saccharomyces cerevisiae*, was knocked out by transforming a plasmid containing homologous sequence to the chromosomal *his3* gene (Scherer and Davis, 1979). This gene replacement mediated by homologous recombination is also known as gene targeting and extensively used in yeast leading to the generation of the yeast deletion collection (Giaever and Nislow, 2014). The gene targeting strategy developed in yeast inspired the development of recombineering in bacteria and gene targeting in mouse embryonic stem cells (Urnov, 2018; Zhang et al., 1998). An important study done by Mario R. Capecchi's group showed that the *hprt* gene in mouse embryonic stem (ES) cells was knocked out by gene targeting by replacing one of the exons with a selection marker (*neo*^{r}) (Thomas and Capecchi, 1987). However, the targeting efficiency was very low that only 1 in 950 neomycin resistant colonies showed desired targeted gene replacement. Subsequently, the same group developed a strategy of positive-negative selection to enrich homologous targeting events (Mansour et al., 1988). A targeting vector contains a positive selectable marker, *neo*^{r}, within the homology arms for selecting successful integration of the targeting vector (Figure 11). The targeting vector also contains a negative selectable marker, *HSV-tk*, outside the homology arm for selecting against random insertion of the targeting vector. Cells resulting from random insertion of the targeting vector express herpes simplex virus thymidine

kinase from the *HSV-tk* gene that makes the cells sensitive to ganciclovir and FIAU. This improved gene targeting strategy became one of the main methods in generating a collection of mouse ES cell lines that has every gene in the genome knocked out (Collins et al., 2007). The significance of this kind of research was recognised by the award of the 2007 Nobel Prize in Physiology or Medicine being awarded to Mario R. Capecchi, Martin J. Evans and Oliver Smithies, for gene targeting in mouse ES cells.

Gene targeting is a very powerful tool but it also has limitations. The gene targeting efficiency is limited by homologous recombination efficiency, which is extremely low and confined to specific cell types and cell cycle (Horii and Hatada, 2016). By using gene targeting only, conditional genome modification cannot be achieved and this can lead to embryo lethality in transgenic mice containing knockout of an essential gene.

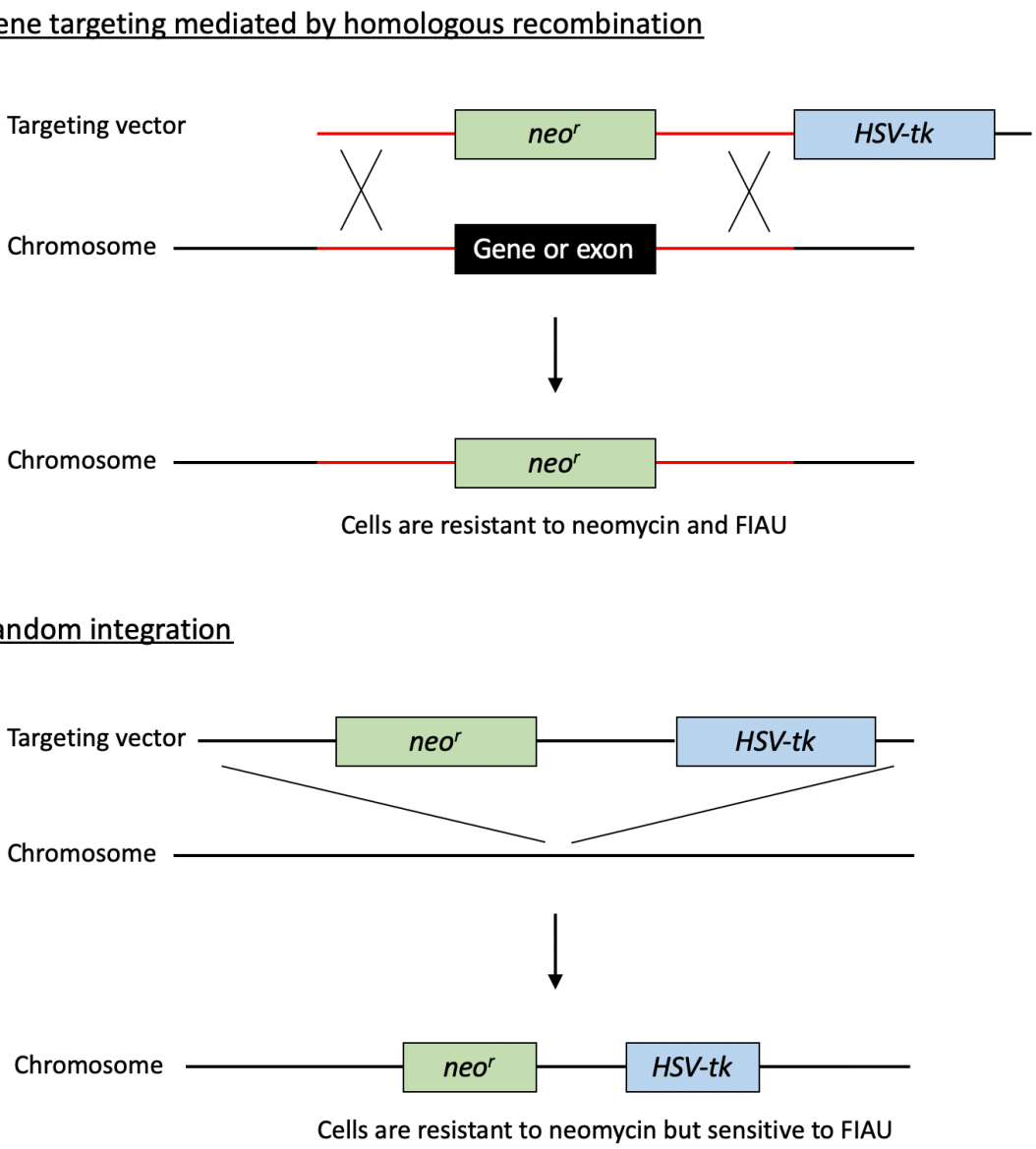


Figure 11. Gene targeting strategy using positive-negative selection.
A targeting vector contains a positive selectable marker (*neo*^{r}) flanked by homology arms (red) targeting to the genome and a negative selectable marker (*HSV-tk*) outside the homology arm. In gene targeting, a *neo*^{r} gene flanked by homology arms replaces a gene of interest or an exon by homologous recombination. Because *HSV-tk* is not integrated into the genome, cells with the gene targeted genome are resistant to neomycin and FIAU. In random integration, the whole linear vector including *HSV-tk* is integrated into the genome making the cells sensitive to FIAU.

1.3.2. Site-specific recombination

Conservative site-specific recombination is a type of genetic recombination between a pair of short DNA sequences catalysed by site-specific recombinases (SSRs). Each DNA sequence contains a pair of inverted repeats for recruiting a dimer or two monomers of recombinases (Figure 12A). The inverted repeat pair is separated by the cross-over site where the DNA breakage occurs and this asymmetric sequence could determine the polarity of the recombination site. Site-specific recombination differs from homologous recombination that the process does not require long stretches of DNA homology in the two SSR binding sites and a single protein SSR does all the jobs, i.e. catalyses DNA cleavage, strand exchange and DNA religation. The outcomes of site-specific recombination are predictable and include integration, deletion and inversion depending on the location and orientation of the two recombination sites (Figure 12B). SSRs catalysed site-specific recombination reactions are generally more specific and more efficient than other genome engineering tools such as gene targeting and targeted nucleases (Olorunniji et al., 2016). However, the use of SSRs in genome manipulation requires pre-integration of SSR binding site(s) into the target genome that is normally done by gene targeting mentioned in section 1.3.1. By combining gene targeting and conservative site-specific recombination, cell type specific and temporal control of genome modification can be achieved *in vivo* (Ahn and Joyner, 2004; Lagace et al., 2007).

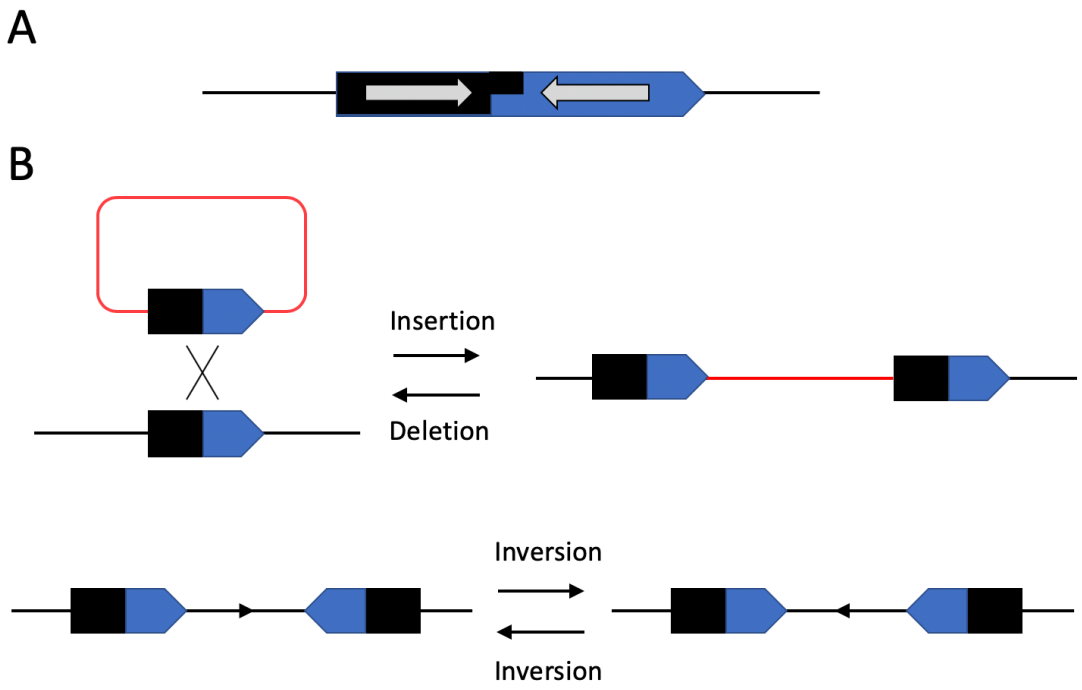


Figure 12. Outcomes of SSR catalysed site-specific recombination.
(A) Schematic representation of a recombination site in site-specific recombination. Grey arrowheads represent a pair of inverted repeats that recruit recombinases. The inverted repeats are separated by the cross-over site where the DNA breakage occurs. (B) Site specific recombination can result in insertion, deletion and inversion depending on the location and orientation of the recombination sites (black and blue boxes).

SSRs perform a variety of biological functions including integration and excision of mobile genetic elements and controlling gene expression in some bacteria (Grindley et al., 2006). They are classified into two families, tyrosine recombinases and serine recombinases, according to the conserved amino acid residue that makes the nucleophilic attack to DNA and forms a covalent bond with the DNA during the recombination reaction. Although they both catalyse DNA recombination, the two families showed no structural or sequence similarity and displayed different mechanisms of action during strand exchange. When SSRs bind to inverted repeats in a recombination site to form a dimer, the two recombination sites are paired to form a synaptic complex through protein-protein interaction of SSRs. For tyrosine

recombinases, they generate a single-stranded break at each recombination site and join the ssDNA between the two recombination sites to form a Holliday junction intermediate (Figure 13). The Holliday junction is resolved by a second round of ssDNA breakage and rejoining. Examples of widely used tyrosine recombinase systems in genome engineering are Cre-*loxP* from the *E. coli* bacteriophage P1 and Flp-*FRT* from *S. cerevisiae* 2µ plasmid (Gaj et al., 2014). For serine recombinases, the enzymes are modular containing a N-terminal catalytic domain, which is responsible for core site recognition and catalysing recombination, and a C-terminal DNA binding domain for binding to the inverted repeat. After the formation of the synaptic complex, serine recombinases generate a DSB at each recombination site and strand exchange is completed by a 180° rotation of two subunits followed by rejoining of the DNA strands (Li et al., 2005) (Figure 13). Because serine recombinases contain separate catalytic and DNA binding domains, their catalytic domains were fused to other DNA binding proteins to achieve DNA recombination at user-defined sites (Akopian et al., 2003; Standage-Beier et al., 2019). This is discussed in more details in section 1.3.4.

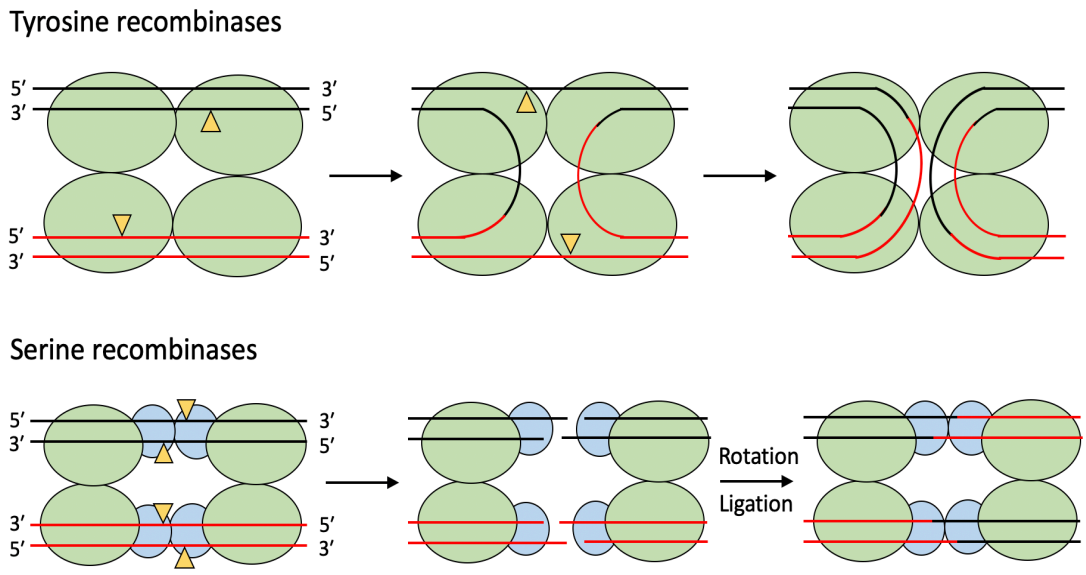


Figure 13. Strand exchange mechanisms of two different families of SSRs.
Tyrosine recombinases cut and exchange one DNA strand at each recombination site to form a Holliday junction intermediate. The Holliday junction is resolved by cutting and exchanging the other pair of DNA strands. Serine recombinases generate a DSB in each recombination site and the broken DNAs are swapped between the two recombination sites by a 180° rotation of two subunits.

1.3.3. Genome editing by targeted endonucleases

A key discovery leading to the era of genome editing using endonucleases was that a DSB generated by a rare endonuclease, I-SceI could be repaired precisely from a donor DNA by homology-directed repair (HDR) in mitotic dividing mammalian cells (Rouet et al., 1994). The study also showed the DSB could be repaired by the error-prone non-homologous end joining (NHEJ) pathway leading to small insertions/deletions (indels). Subsequent studies showed that the efficiency of gene targeting was enhanced by I-SceI induced DSBs and hereafter, these genome engineering techniques relying on DSB repair were collectively termed genome editing (Choulika et al., 1995; Donoho et al., 1998; Rocha-Martins et al., 2015). Early genome editing techniques utilised meganucleases, a family of naturally occurring rare cutting restriction enzymes, laid the foundation of genome editing strategies used today including gene correction,

insertion and disruption (Figure 14). However, meganucleases such as I-SceI recognises an 18-bp sequence that is hard to find in the genome. Thus, early genome editing required pre-integration of the enzyme recognition site into the genome limiting the potential of the tool (Silva et al., 2011).

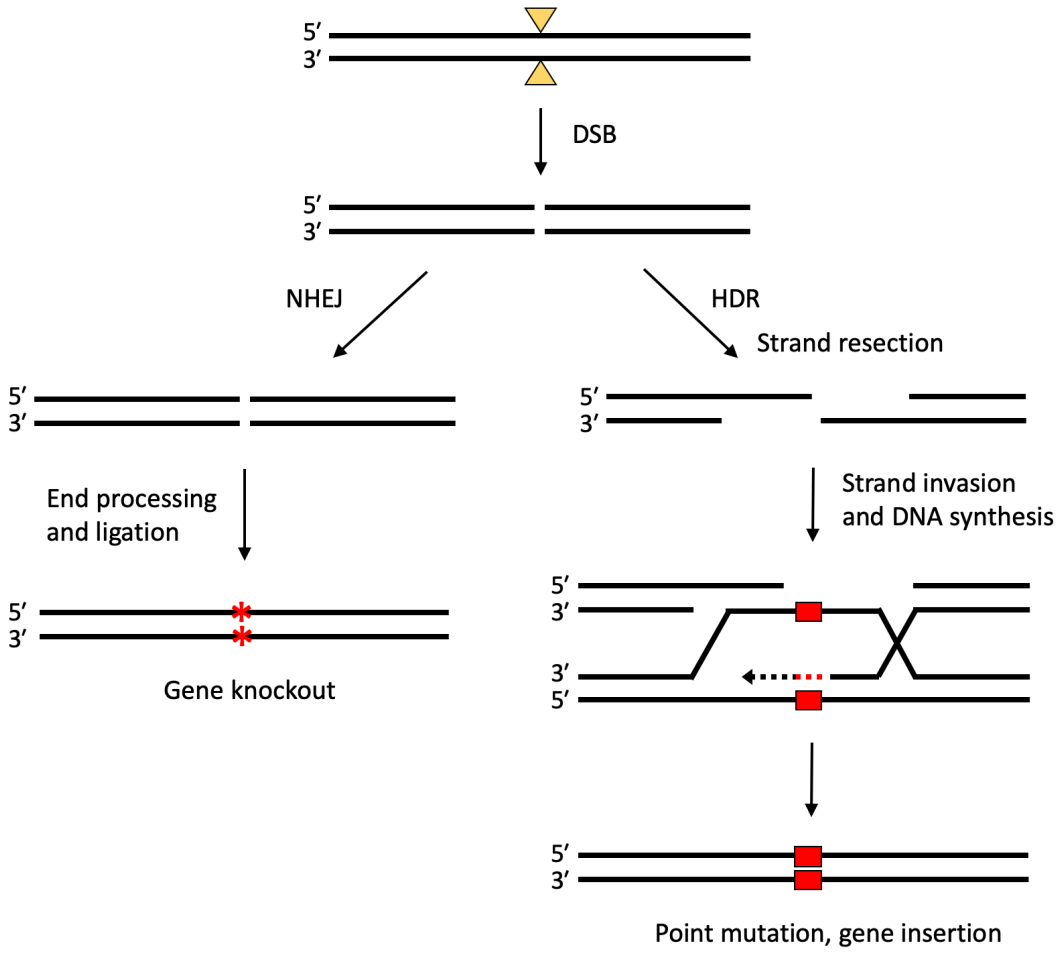


Figure 14. Endonucleases generated DSBs are repaired by two competing pathways leading to different outcomes.
DSBs generated by endonucleases in genome editing are repaired by either error-prone non-homologous end joining (NHEJ) resulting in small indels (red asterisk) and gene knockout or repaired by homology-directed repair (HDR) using donor DNA to introduce precise point mutation and gene insertion (red box).

Discoveries of the zinc-finger DNA binding domain and the catalytic domain of FokI revolutionised genome editing (Li et al., 1992; Miller et al., 1985). The crystal structure

of a DNA bound Zif268 protein containing three zinc fingers showed that each zinc finger recognises a 3-bp sequence and the amino acid residues that interact with the bases were identified (Pavletich and Pabo, 1991). Since then, researchers mutated zinc fingers for binding to different DNA triplet sequences (Choulika et al., 1995). By linking several zinc fingers into an array, it generates a DNA binding domain that recognises user-defined sequences of n×3-bp long where n is the number of zinc fingers in the array. The first zinc finger nuclease (ZFN) was constructed by fusing a FokI catalytic domain to a DNA binding domain containing three zinc fingers (Kim et al., 1996). The efficiency of this ZFN was lower than the wild-type FokI because FokI dimerisation is required for efficient double-strand break (Bitinaite et al., 1998). Thus, subsequent ZFNs were designed to work in pair that the zinc finger domain of each ZFN bind to the opposite DNA strand to allow dimerisation of FokI in the middle (Figure 15A). The drawback of ZFNs is the unpredicted specificity or unexpected failure in target DNA binding arising from neighbouring zinc fingers influencing DNA binding specificity (Garton et al., 2015).

Transcription activator-like effector nucleases (TALENs) provides an alternative to ZFNs and they were developed by fusing a FokI catalytic domain to a transcription activator-like effector (TALE) DNA binding domain. TALE proteins are naturally occurring proteins secreted by *Xanthomonas* bacteria to help bacterial infection of plants. They bind to promoter sequences of the host plant cell and activate gene expression leading to programmed cell death (Römer et al., 2007). It was shown that TALE proteins bind DNA through a tandem repeat domain of each TALE repeat containing 34 residues. Each repeat contains two adjacent hypervariable amino acid residues called repeat variable diresidue (RVD) (Moscou and Bogdanove, 2009). The

crystal structure of a DNA bound TALE protein revealed that each TALE repeat forms two alpha helices linked by a RVD containing loop where the second RVD residue interacts with a DNA base in a base-specific manner and the first RVD residue stabilises the RVD-containing loop (Deng et al., 2012). Therefore, altering the RVD can affect the specificity for each of the four DNA bases and the RVD codes for each base were deciphered (Boch et al., 2009; Cong et al., 2012). A TALEN can target any sequences by assembly of TALE repeats containing corresponding RVDs (Figure 15B). TALENs are advantageous over ZFNs that TALE repeats are easier to design than zinc finger motifs and the DNA binding specificity of TALE is more predictable and reliable (Gupta and Musunuru, 2014).

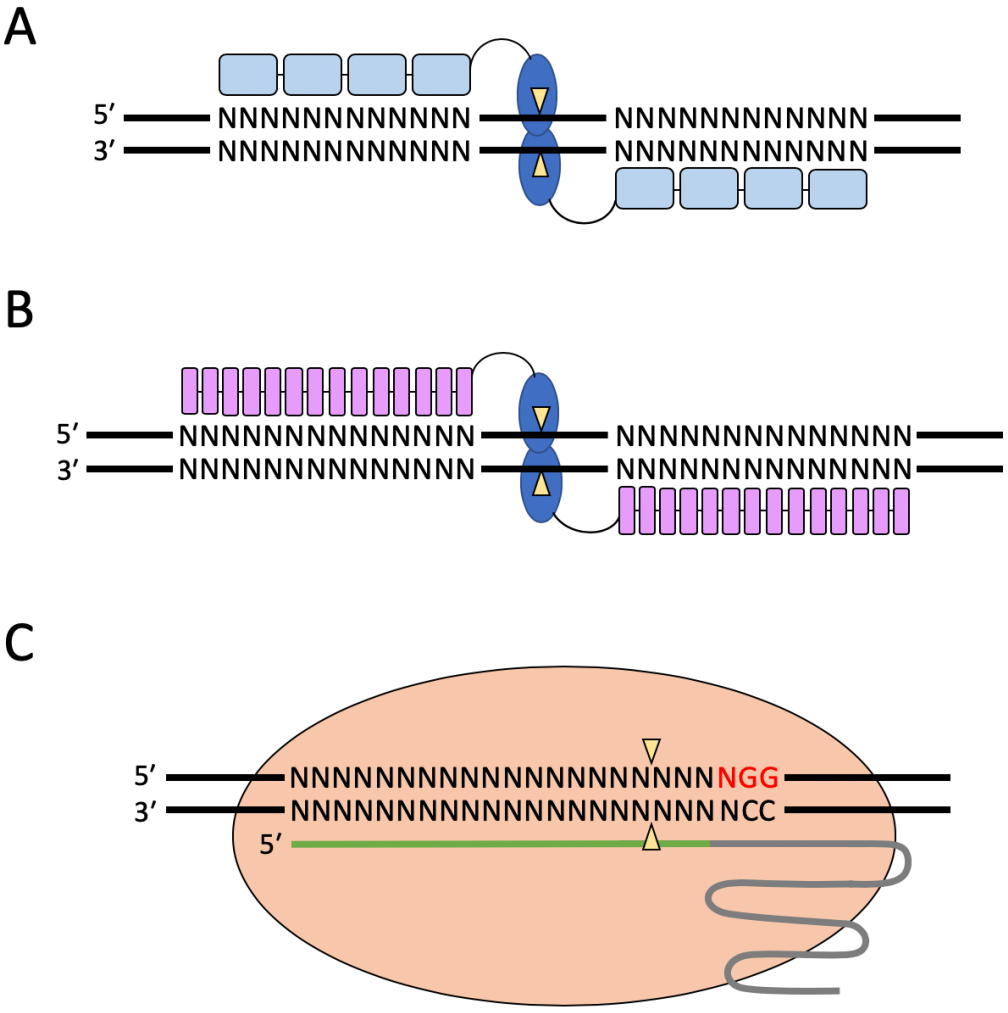


Figure 15. Mechanisms of target DNA recognition and cleavage of ZFNs, TALENs and Cas9.

(A) Schematic representation of a DSB generated by a pair of ZFNs. The DNA binding domain in this figure binds a 12-bp sequence with each zinc finger (light blue) recognises 3 bp. A pair of ZFNs bind to the opposite DNA strands allowing dimerisation of FokI (dark blue) and efficient DNA cleavage in the middle (yellow triangle). (B) Schematic representation of a DSB generated by a pair of TALENs. The TALEN shown here contains 14 TALE repeats (pink) in tandem and each repeat recognise a single base specifically. A pair of TALENs bind to the opposite DNA strands allowing dimerisation of FokI (dark blue) and efficient DNA cleavage in the middle (yellow triangle). (C) Schematic representation of a DSB generated by Cas9-sgRNA. Cas9-sgRNA binds to a PAM (red) containing target DNA that is complementary to the guide sequence in sgRNA (green). Then Cas9 generates a DSB 3 bp upstream of the PAM (yellow triangle).

The discovery of CRISPR loci and subsequent characterisation of the interference effector protein SpCas9 from type II systems revolutionised the genome editing field again (Lander, 2016). The DNA recognition and binding of Cas9 is dependent on a crRNA-tracRNA dual RNA molecule. Cas9 generates a blunt ended DSB at a PAM containing DNA sequence that is complementary to the crRNA (Figure 15C). Details of target DNA binding and cleavage by Cas effector proteins are described in section 1.1.3.3. The CRISPR-Cas9 system was further simplified by linking crRNA and tracrRNA to form a single guide RNA (sgRNA) (Jinek et al., 2012). Cas9 is advantageous over ZFNs and TALENs that by simply changing the guide sequence in sgRNA, Cas9 can be retargeted to a new DNA sequence allowing for multiplexing. ZFNs and TALENs, on the other hand, require reconstruction of the DNA binding domain, which is time consuming (Gupta and Musunuru, 2014). Although Cas9 requires a PAM sequence in the target DNA for binding, researchers have identified different Cas9 variants that recognise different PAMs (Hu et al., 2018). In addition, Cas12 interference effector proteins from type V systems show completely different PAM specificity making them a new potential genome editing tool (Dong et al., 2016). However, it was later shown that Cas12 is not ideal as a targeted endonuclease because it cleaves ssDNA non-specifically when it binds to target DNA. However, this indiscriminate ssDNA cleavage activity has been exploited as a reporter in Cas12-based high sensitivity nucleic acid detection tool (Chen et al., 2018; Gootenberg et al., 2018).

The biggest concern regarding using targeted nucleases in genome editing is the off-target effect. Cas9 can tolerate a few mismatches outside the seed region and still cuts DNA (Zhang et al., 2015). Increased specificity can be achieved by fusing catalytically

inactive dCas9 to FokI and the fusion protein works similarly to ZFNs and TALENs (Guilinger et al., 2014). For in vivo genome editing, it was shown that directly deliver Cas9-sgRNA RNP complex into cells along with donor DNA increases editing efficiency and specificity by limiting the working time-window (Shapiro et al., 2020).

1.3.4. Cas proteins expand research tools

The RNA-guided DNA targeting activity of CRISPR-Cas9 has inspired the development of many Cas9 chimeric proteins, expanding molecular research tools. Because DSBs are repaired predominantly by the error-prone NHEJ and precise editing requires HDR, one of the goals of Cas9 chimeras is to efficiently edit the genome without causing a DSB. A recent study on the fusion of dCas9 to a catalytic domain of Tn3 resolvase, a serine recombinase, showed the fusion protein is able to perform RNA-guided site-specific deletion and insertion despite of low efficiency of integration (0.08%) (Standage-Beier et al., 2019). In order to integrate a large DNA fragment without inducing DSBs, RNA-guided DNA transposition was performed by fusing dCas9 to transposase (Bhatt and Chalmers, 2019) or by using CRISPR associated transposases found in naturally occurring Tn7-like transposons (Klompe et al., 2019; Strecker et al., 2019). Precise point mutations can be introduced into the genome using base editors without causing DSBs. Current base editors are able to convert C to T, A to G and C to G in a narrow editing window and they were developed from fusions of nickase Cas9 (nCas9) to adenine deaminase or cytosine deaminase enzymes (Gaudelli et al., 2017; Kim et al., 2019a; Komor et al., 2016). These base editors showed editing efficiency ranging from 11% to 75% in different targets and cell lines.

In addition to directly edit the genome, Cas9 chimeras can also edit the epigenome and regulate gene expression. At the beginning, CRISPR activation (CRISPRa) tools were dCas9 fusions to a single transcription activator such as Vp64. Subsequent development of CRISPRa tools involved fusing transcription activators p65 and Rta to the Vp64 moiety of dCas9-Vp64 to further increase the expression level of target genes (Chavez et al., 2015). Similarly, targeted gene repression was achieved by fusing dCas9 to a transcription repressor KRAB domain (Qi et al., 2013).

1.4. Aims of the project

Casposases are Cas1 homologues that are found in a novel family of transposons named casposons. It is suggested that ancestral casposon gave rise to CRISPR adaptation module and CRISPR repeats, which means casposases might be the ancestral state of Cas1. I was interested in how casposases differ from Cas1 in terms of substrate specificity.

The characterisation of *A. boonei* casposase showed it can integrate a wide variety of DNA substrates. I attempted to exploit this promiscuous substrate specificity of casposase in targeted integration by fusing it to SpCas9 mutants from a type II-A system or dCasX from a type V-E system. We wondered if a Casposase Cas9 fusion protein can achieve targeted integration in a one-step process to bypass HDR (Figure 16). Fusion proteins were studied biochemically to investigate RNA-guided DNA integration property.

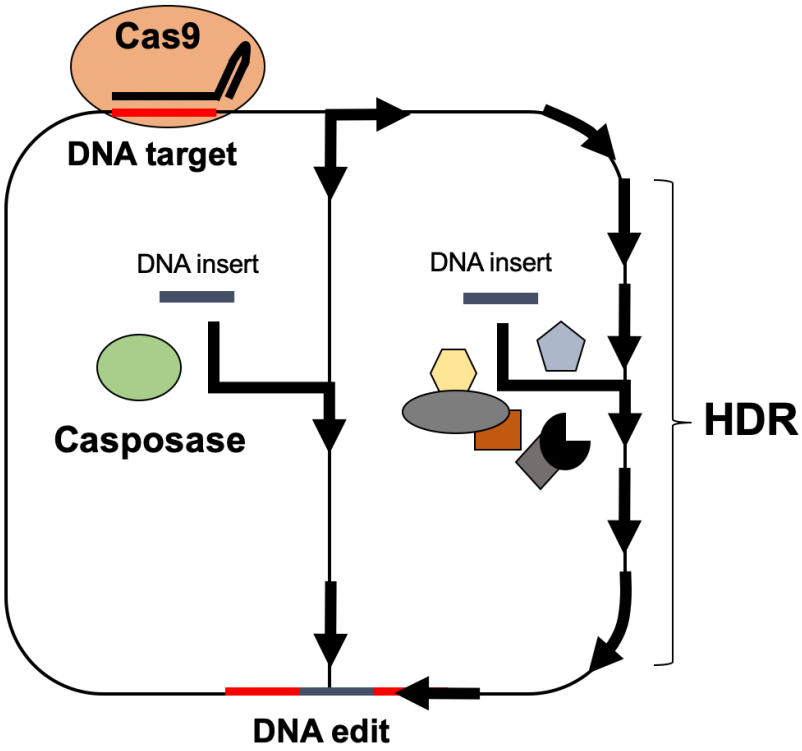


Figure 16. A hypothesis of casposase-Cas9 fusion protein can bypass HDR to achieve targeted DNA insertion.

A recent study biochemically characterised a tyrosine recombinase, Int^{pTN3} integrase and illustrated that the enzyme catalysed low sequence specificity recombination. Therefore, in a parallel study, the Int^{pTN3} integrase was fused to dCas9 to investigate RNA-guided recombination between homologous sequences.

Chapter 2

2. Materials and Methods

2.1. Molecular cloning

2.1.1. Polymerase chain reaction (PCR)

All primers used for polymerase chain reactions (PCRs) listed in appendix 1 were purchased from Sigma-Aldrich (St. Louis, Missouri, US) at 0.05 µmole scale and were delivered purified by desalting. All DNA polymerases were purchased from New England Biolabs (NEB, Ipswich, Massachusetts, US). Vent® DNA Polymerase was used to amplify DNA fragments for cloning and PCR reactions were performed with 1 unit of Vent® DNA Polymerase, 1x ThermoPol Reaction Buffer (20 mM Tris-HCl pH 8.8, 10 mM KCl, 2 mM $MgSO_4$, 10 mM $(NH_4)_2SO_4$, 0.1% Triton® X-100), 200 µM dNTP mix (NEB) and 0.5 µM of forward and reverse primers. One*Taq*® Hot Start DNA Polymerase was used to amplify double-stranded DNA fragments for an integration assay and amplify DNA fragments after an integration assay. PCR reactions were made the same as Vent PCR reactions except with 1 unit of One*Taq*® Hot Start DNA Polymerase and 1x One*Taq* Standard Reaction Buffer (20 mM Tris-HCl pH 8.9, 22 mM KCl, 1.8 mM $MgCl_2$, 22 mM NH_4Cl, 0.05% Tween® 20, 0.06% IGEPAL® CA-630). PCR amplification programmes (Table 2.1 and Table 2.2) were carried out using an Applied Biosystems® Veriti™ 96-Well Thermal Cycler. The annealing temperature of each primer pair was calculated using the NEB Tm calculator (https://tmcalculator.neb.com/#!/main). PCR products were resolved on a 0.7% agarose gel containing 0.5 µg/ml ethidium bromide (EtBr, Sigma-Alrich) in 1x Tris-borate-EDTA (TBE) buffer (89 mM Tris, 89 mM boric acid, 2 mM EDTA) and visualised on a UV-transilluminator (Syngene, Bangalore, India). The

correct sized band was excised and DNA was gel purified by QIAquick® Gel Extraction Kit (Qiagen, Hilden, Germany) following manufacturer's instructions.

Table 2.1. Thermocycling conditions for a Vent PCR

Step	Temperature (˚C)	Time
Initial denaturation	95	2 min
30 cycles	95 55-65 72	15s 20s 60s per kb
Final extension	72	5 min

Table 2.2. Thermocycling conditions for a OneTaq PCR

Step	Temperature (˚C)	Time
Initial denaturation	94	30s
30 cycles	94 45-68 68	15s 20s 60s per kb
Final extension	68	5 min

2.1.1.1. Overlap extension PCR

The first round of PCR included amplification of *casposase* gene and *cas9* gene respectively. The genomic DNA of *Aciduliprofundum boonei T469* was purchased from Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ, Braunschweig, Germany). The *A. boonei* CRISPR-associated endonuclease Cas1, *casposase* gene (NCBI gene ID: 8827333) was PCR amplified by Vent® DNA Polymerase using Assem Casp F and Assem Casp R primers (see Appendix 1). The *cas9* gene was

amplified from pMJ806 (Addgene, Watertown, Massachusetts, U.S.) using Assem Cas9 F and Cas9 pDuet R primers. The PCR products were run on agarose gel and the correct sized DNA was gel purified as described in 2.1.1. Equal molar amounts of purified products from the first round of PCR reactions were mixed and used as template for the second round of PCR reaction. Primers Assem Casp F and Cas9 pDuet R were used and the extension time was lengthened corresponding to the total length of the fusion product.

2.1.2. Restriction digestion and DNA end modification

All restriction enzymes used in this work were purchased from NEB. DNA was digested with 1 unit of restriction enzymes in 1x CutSmart® Buffer (NEB, 20 mM Tris-acetate pH 7.9, 50 mM Potassium acetate, 10 mM Magnesium acetate, 100 µg/ml BSA) at 37°C for one hour. The digested PCR products were purified by QIAquick® PCR purification Kit (Qiagen). For the digestion of backbone vectors, 1 unit of alkaline phosphatase from calf intestinal (CIP, NEB) was added to the reaction at the end of digestion and incubated at 37°C for 30 min. The digested plasmid vectors were then run on 0.7% agarose gel and purified by QIAquick® Gel Extraction Kit (Qiagen).

2.1.3. Ligation and transformation

A total of 100 ng of DNA containing 3:1 molar ratio of insert: vector was incubated with 1 µL of T4 DNA Ligase (NEB) in 1x T4 DNA Ligase Reaction Buffer [50 mM Tris-HCl pH 7.5, 10 mM MgCl2, 10 mM dithiothreitol (DTT), 1 mM adenosine triphosphate (ATP)] at 16°C overnight. Then the ligation mix was chemo-transformed into DH5α (see section 2.1.6.1).

2.1.4. Site-directed mutagenesis

Site-directed mutagenesis was based on PCR to generate insertions, deletions and substitutions in plasmid DNA. PCR reactions were performed with 1 unit of Q5® High-Fidelity DNA Polymerase (NEB), 1x Q5 Reaction Buffer (25 mM TAPS-HCl pH 9.3, 50 mM KCl, 2 mM $MgCl_2$, 1 mM β-mercaptoethanol), 200 µM dNTP mix and 0.5 µM of forward and reverse primers and 2 ng of template plasmid DNA. The PCR amplification programme was shown in table 2.3. After Q5 site-directed mutagenesis PCR, the template plasmid was degraded by DpnI (NEB) at 37°C for one hour. The linear PCR product was added with 5' phosphate by T4 Polynucleotide Kinase (NEB) in 1x T4 DNA Ligase Reaction Buffer at 37°C for 30 min and re-circularised by T4 DNA Ligase at 16°C overnight. The ligated plasmid was transformed into chemically competent DH5α (see section 2.1.6.1).

Table 2.3. Thermocycling conditions for a Q5 PCR

Step	Temperature (˚C)	Time
Initial denaturation	98	30s
25 cycles	98 50-72 72	10s 20s 30s per kb
Final extension	72	2 min

2.1.5. Vectors

The DNA sequences of constructed plasmids were verified by Sanger sequencing performed by Source BioScience, Nottingham, U.K., and plasmids were purified from a single colony by QIAprep® Spin Miniprep Kit (Qiagen) following manufacturer's instructions. The *A. boonei casposase* gene was PCR amplified and inserted into

pACYC-Duet plasmid between BamHI and PstI restriction sites. The resulting plasmid pCHL2 encodes casposase protein with a N-terminal His tag. Q5 site-directed mutagenesis PCR was used to generate pCHL13 and pCHL14 using pCHL2 as a template. pCHL13 and pCHL14 encoded a casposase active site mutant H242A and D254A reactively. The 38 bp terminal inverted repeats (TIRs) of *A. boonei* casposon were inserted at both ends of amp^R gene in pET-14b to generate pCaspamp plasmid containing a mini-casposon. The vectors constructed for chapter 3 are listed in table 2.4.

For fusion proteins, the genes encoding SpCas9 and SpdCas9 were amplified from pMJ806 and pMJ841 respectively that were purchased from Addgene. The appropriate gene was inserted into pCHL2 downstream of the *casposase* gene between PstI and NotI restriction sites. The sequence between the *casposase* stop codon and *cas9/dcas9* was replaced by different linkers using Q5 site directed mutagenesis PCR. The flexible linkers comprised either the 16-residue 'SGSETPGTSESATPES' 'XTEN' sequence (Guilinger et al., 2014), 18-residue $(GGS)_6$ or 24-residue $(GGS)_8$ sequence. The vectors constructed for chapter 4 were listed in table 2.4.

A DNA fragment of 250 bp from pACYC-Duet containing integration sites A and B (see section 4.2.5) was inserted into pMA-T using GeneArt® Gene Synthesis service (Thermo Fisher) and the resulting plasmid was called pABwt. Another plasmid pABmut was also synthesised and it contained the same sequence as pABwt except 15 bp flanking integration sites A and B were mutated to random sequences.

In order to detect full-site integration of oligonucleotide duplex, a pCHL2 plasmid with the insertion of mini-casposon was used and called pCHL41. This pCHL41 was derived

from a mini-casposon integration assay and was resistant to both ampicillin (amp) and chloramphenicol (cm). The chloramphenicol resistance gene (cm^R) in pCHL41 was inserted with 67 bp at the start of the gene and this caused frameshift inactivating the gene. The 67 bp sequence included 32 bp DNA sequence flanking the fusion protein integration site (site B, see section 4.2.5). The resulting plasmid was called pCHL42.

The sgRNA expressing plasmid pgRNA-bacteria conferring ampicillin resistance was purchased from Addgene. Q5 mutagenesis PCR was used to insert guide sequence targeting to *Escherichia coli lacZ* gene into pgRNA-bacteria to generate pglacZ. The sglacZ RNA expressing locus was PCR amplified and subcloned into pCDF-1b between PstI and XbaI sites to generate a sgRNA expressing plasmid, sglacZ smR, conferring spectinomycin resistance.

For high-fidelity (hf) fusion proteins, three rounds of Q5 mutagenesis PCR were used to introduce substitutions, K848A, K1003A, R1060A in a His-MBP-Cas9 plasmid pMJ806 (Addgene). The *hfcas9* gene was subcloned into pCHL2. A $(GGS)_8$ linker was inserted to replace the sequence between *casposase* stop codon and *hfcas9* by Q5 PCR to generate a plasmid pCHL29 encoding a Casp-hfCas9 fusion protein. H840A was introduced into *cas9* in pCHL29 to generate pCHL30 turning hfCas9 into hfnCas9. D10A was introduced into *ncas9* in pCHL30 to generate pCHL31 turning hfnCas9 into hfdCas9. A PstI restriction site was introduced into the hf-fusion protein plasmids and sglacZ RNA expressing locus was subcloned into the PstI site to generate a plasmid expressing both hf-fusion protein and sgRNA.

For experiments in human cells, proteins need to contain a nuclear localisation signal (NLS) for transport of the protein into the nucleus. A pNLS-His-StrepII plasmid containing 6xHis tag-NLS-restriction sites-NLS-Strep-tag® II was constructed by a

BBSRC DTP student (Mr Andrew Cubbon). pNLS-Cas9 containing *cas9* gene inserted between PstI and NotI was also made by Andrew Cubbon. *Casp-hfnCas9* and *Casp-hfdCas9* genes from pCHL30 and pCHL31 were subcloned into pNLS-His-StrepII between PstI and EagI sites respectively.

The gene encoding casposase from pCHL2 was subcloned into pNLS-His-StrepII between EcoRI and SalI sites to generate pNLS-Casp. pET 2C-T10-dCasX was purchased from Addgene and *dcasX* gene was subcloned into pNLS-Casp between SalI and EagI sites. A $(GGS)_8$ linker was inserted to replace the sequence between *casposase* stop codon and *dCasX* by Q5 PCR to generate a casposase-$(GGS)_8$-dCasX fusion protein, Casp-dCasX. The vectors constructed for chapter 5 were listed in table 2.5.

An integrase from *Thermococcus nautili* plasmid pTN3 (Int^{pTN3}, NCBI gene ID: 17125032) was codon-optimised for expression in *E. coli* and the gene fragment was synthesised using GeneArt® Gene Synthesis service. The Int^{pTN3} fragment was cloned into pACYC-Duet between BamHI and PstI sites to generate pCHL11. Then *dcas9* gene from pMJ841 was subcloned into pCHL11 downstream of the Int^{pTN3} gene between PstI and NotI restriction sites. The sequence between the Int^{pTN3} stop codon and *dcas9* was replaced by a $(GGS)_8$ linker using Q5 site directed mutagenesis PCR and then another round of Q5 mutagenesis PCR was used to insert a Strep-tag® II sequence at the C terminal end of the protein. The resulting plasmid pCHL49 encoded a His-tagged and StrepII-tagged fusion protein Int^{pTN3}-dCas9. To test the inversion and deletion activity of Int^{pTN3}, a XhoI restriction site was first introduced into a pUC19 plasmid downstream of the amp^R gene. The *lacZ*α gene in pUC19 was PCR amplified and a XhoI site was added at both ends of the gene fragment. After XhoI digestion and T4 ligation (NEB) of the gene fragment and the XhoI site containing pUC19 plasmid, the resulting

plasmids contained two *lacZ*α fragments facing either in the same direction or in the opposite direction.

Table 2.4. Plasmids constructed and used in chapter 3 and 4

Plasmid	Feature	Antibiotic resistance
pCHL2	casposase in pACYC-Duet	chloramphenicol
pCHL5	casposase-$(GGS)_6$-Cas9 in pACYC-Duet	chloramphenicol
pCHL6	casposase-$(GGS)_8$-Cas9, Casp-Cas9 in pACYC-Duet	chloramphenicol
pCHL7	casposase-XTEN-Cas9 in pACYC-Duet	chloramphenicol
pCHL8	casposase-$(GGS)_6$-dCas9 in pACYC-Duet	chloramphenicol
pCHL9	casposase-$(GGS)_8$-dCas9, Casp-dCas9 in pACYC-Duet	chloramphenicol
pCHL10	casposase-XTEN-dCas9 in pACYC-Duet	chloramphenicol
pCHL13	casposase^{H242A} in pACYC-Duet	chloramphenicol
pCHL14	casposase^{D254A} in pACYC-Duet	chloramphenicol
pCHL20	casposase^{D254A}-$(GGS)_8$-dCas9 in pACYC-Duet	chloramphenicol
pCaspamp	*amp^R* gene flanked by 38 bp *A. boonei* casposon TIR	ampicillin
pABwt	250 bp from pACYC-Duet including integration sites A and B from fusion protein integration assay was inserted in pMA-T.	ampicillin
pABmut	Same as pABwt except integration sites A and B were mutated to random sequences	ampicillin

psglacZ1	sgRNA targeting to the middle position of *E. coli lacZ* gene in pgRNA-bacteria	ampicillin
psglacZ1 smR	sgRNA targeting to the middle position of *E. coli lacZ* gene in pCDF-1b	spectinomycin
psglacZ2	sgRNA targeting to the start position of *E. coli lacZ* gene in pgRNA-bacteria	ampicillin
psglacZ2 smR	sgRNA targeting to the start position of *E. coli lacZ* gene in pCDF-1b	spectinomycin
pCHL27	His-MBP-hfCas9 (K848A, K1003A, R1060A) in pMJ806	kanamycin
pCHL29	casposase-$(GGS)_8$-hfCas9, Casp-hfCas9 in pACYC-Duet	chloramphenicol
pCHL30	casposase-$(GGS)_8$-hfnCas9 H840A, Casp-hfnCas9 in pACYC-Duet	chloramphenicol
pCHL31	casposase-$(GGS)_8$-hfdCas9, Casp-hfdCas9 in pACYC-Duet	chloramphenicol
pCHL35	Casp-hfCas9, sgRNA targeting *lacZ* (sglacZ1) in pACYC-Duet	chloramphenicol
pCHL36	Casp-hfnCas9, sgRNA targeting *lacZ* (sglacZ1) in pACYC-Duet	chloramphenicol
pCHL37	Casp-hfdCas9, sgRNA targeting *lacZ* (sglacZ1) in pACYC-Duet	chloramphenicol

pCHL41	pCHL2 with the insertion of mini-casposon	ampicillin and chloramphenicol
pCHL42	pCHL41 with 67 bp insertion immediately downstream of cm^R start codon	ampicillin
pCHL50	Casp-hfdCas9, sgRNA targeting *lacZ* (sglacZ2) in pACYC-Duet	chloramphenicol
pNLS-Cas9	His-NLS-Cas9-NLS-Strep in pACYC-Duet	chloramphenicol
pNLS-hfnCaspoR	His-NLS-Casp-hfnCas9-NLS-Strep in pACYC-Duet	chloramphenicol
pNLS-hfdCaspoR	His-NLS-Casp-hfdCas9-NLS-Strep in pACYC-Duet	chloramphenicol

Table 2.5. Plasmids constructed and used in chapter 5

Plasmid	Feature	Antibiotic resistance
pNLS-Casp	His-NLS-casposase-NLS-Strep in pACYC-Duet	chloramphenicol
pNLS-dCaspoX	His-NLS-Casp-dCasX-NLS-Strep in pACYC-Duet	chloramphenicol
pCHL11	Int^{pTN3} in pACYC-Duet	chloramphenicol
pCHL43	*lacZ*α gene inserted downstream of amp^R gene in pUC19, same orientation as the original *lacZ*α	ampicillin

pCHL44	*lacZα* gene inserted downstream of *amp*^{R} gene in pUC19, opposite orientation as the original *lacZα*	ampicillin
pCHL49	His-Int^{pTN3}-dCas9-Strep in pACYC-Duet	chloramphenicol

2.1.6. Bacterial cell strains

The details of cell strains are listed in table 2.6. DH5α was used in cloning and used in transforming plasmids after an integration assay. L-arabinose inducible T7 RNA polymerase strain BL21-AI was used in protein expression for protein purification and *in vivo* assays. EB286 was used to generate CL003 by P1 transduction. CL003 and EB377 were used in *in vivo* assays.

Table 2.6. E. coli cell strains used in this project

Strain	Features
DH5α	F–, ϕ80*lacZ*ΔM15, Δ(*lacZYA-argF*)U169, *rec*A1, *end*A1, *hsd*R17(rK–, mK+), *pho*A, *sup*E44, λ–, *thi*-1, *gyr*A96, *rel*A1
BL21-AI	F–, *omp*T, *hsd*SB (rB–, mB–), *gal, dcm, ara*B::T7RNAP-*tet*A
EB286	Wild type K-12 MG1655, Δ*lacIYZA::FRT*
CL003	Wild type K-12 MG1655, Δ*lacIYZA::FRT*, *ara*B::T7RNAP-*tet*A
EB377	Wild type K-12 MG1655 background, *ara*B::T7RNAP-*tet*A

2.1.6.1. Making chemocompetent cells and chemical transformation

Luria-Bertani (LB) medium (1% tryptone, 0.5% yeast extract, 1% NaCl, 2mM NaOH) was inoculated with one cell colony and it was incubated at 37°C with shaking overnight. Fresh LB with inoculation of a hundredth of overnight culture was grown at 37°C with shaking until optical density at 600 nm wavelength (OD_{600}) reached 0.4-0.6. The cells were pelleted and resuspended in ice-cold 0.1 M $CaCl_2$ to a fifth of LB's volume. After incubation on ice for one hour, cells were pelleted again. The cells were resuspended in ice-cold 0.1 M $CaCl_2$ and 30% glycerol to a fifth of LB's volume. The cells can be used directly for transformation or flash frozen for storage at -80°C.

For transformation of chemically competent *E. coli*, 100 µL competent cells were mixed with 20-100 ng of plasmid DNA and incubated on ice for 5 min. Then the cells were heat shocked at 42°C for 1 min following incubation on ice for 5 min. LB was added to the cells to a final volume of 1 mL and the culture was grown at 37°C for one hour. The cells were resuspended in 100 µL of LB and spread on a LB agar plate (1.5% w/v agar in LB) containing appropriate antibiotics. The details of antibiotics used are listed in table 2.7.

Table 2.7. Antibiotics used in this project

Antibiotic	Working concentration	Supplier
Ampicillin	50µg/mL	Sigma
chloramphenicol	25µg/mL	Sigma
Kanamycin	50µg/mL	AppliChem
Spectinomycin	50µg/mL	Sigma
Tetracycline	10µg/mL	AppliChem

2.1.6.2. Making electrocompetent cells and electroporation

An overnight culture was set up in LB. Super optimal broth (SOB, 2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$) was inoculated with a hundredth of overnight culture and grown at 37°C with shaking until OD_{600} of 0.5-0.7 was reached. Cells were pelleted and resuspended in ice-cold sterile 10% glycerol in the same volume as SOB and this procedure was repeated twice. After cells were spun down and the supernatant were poured off, residual 10% glycerol was used to resuspend the cell pellet. The cells can be used directly for electroporation or flash frozen for storage at −80°C.

Electrocompetent cells were mixed with DNA and transferred to a cold 1mm electroporation cuvette. The cells were electroporated using an electroporator that was set at 1.8 kV, 200 ohms and 25 µF. Super optimal broth with catabolite repression (SOC, SOB + 20mM glucose) was added to the cells to a final volume of 1 mL and incubated at 37°C for one hour. The cells were pelleted and spread on LB plate containing appropriate antibiotics.

2.1.6.3. P1 transduction of genes between *E. coli* strains

The procedure of P1 transduction was followed using a protocol from Thomason et al., (2007). To generate CL003, BL21-AI was used as a donor strain. Overnight culture of BL21-AI was diluted 100-fold in LB containing 0.2% of glucose and 5 mM $CaCl_2$ and was grown at 37°C for 45 min. A volume of 100 µL P1*vir* was added to the culture and incubated for three hours to lyse cells. A few drops of chloroform were added, and the culture was spun down. The supernatant containing P1*vir*·BL21-AI was collected and stored at 4°C.

Overnight culture of a recipient strain EB286 was pelleted and resuspended in one-half original culture volume of sterile P1 salts solution (10 mM $CaCl_2$, 5 mM $MgSO_4$). Resuspended cells in 100 µL P1 salts solution were mixed with 10 µL of P1*vir*·BL21-AI and incubated at room temperature (RT) for 30 min. The cells were added with 1mL LB and 200 µL of 1 M sodium citrate and incubated at 37°C for one hour. The cells were spun down and resuspended in 100 µL of LB before spreading on a LB agar plate containing 5 mM sodium citrate and 5 µg/ml tetracycline. The plate was incubated at 37°C overnight and a single colony was streaked on a new plate the next day. This was repeated a second time to remove residual P1 phage.

2.2. Protein expression and purification

2.2.1. Recombinant protein expression by classical induction and cell harvesting

A single colony of BL21-AI harbouring plasmid for expression of a target protein was used to inoculate 50 mL LB containing antibiotics to establish an overnight culture. The overnight culture was diluted 100-fold and grown in two litres of LB containing antibiotics at 37°C until OD_{600} reached 0.6. Protein expression was induced by addition of L-arabinose and isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.1% (w/v) and 0.5 mM respectively. The culture was incubated at 18°C overnight. Cells were pelleted at 4,000 rpm for 10 min and resuspended in buffer containing 25 mM Tris pH 7.5, 500 mM NaCl, 20 mM imidazole and 1x protease inhibitor cocktail (Roche, Basel, Switzerland). The resuspended cells were stored at -80°C. Proteins used in all assays in this study were purified from expression by classical induction because auto-induction did not work.

2.2.2. Recombinant protein expression by auto-induction

The procedure of auto-induction was followed using a protocol from Studier (2005). An overnight culture of BL21-AI harbouring plasmid for expression of a target protein was prepared as described in 2.2.1. The overnight culture was diluted 100-fold in one litre of ZYM-5052 medium (1% tryptone, 0.5% yeast extract, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, 2 mM $MgSO_4$, 0.5% glycerol, 0.05% glucose, 0.2% α-lactose, 0.05% L-arabinose, 0.2x trace elements) containing antibiotics. The culture was grown at 37°C for eight hours and then at 25°C for 24 hours. Cells were harvested as in section 2.2.1.

2.2.3. Recombinant protein purification

Resuspended cells were thawed and sonicated on ice for 3 min for every 5 mL with 10s pulses. The lysed cells were incubated with DNase I (Sigma-Aldrich) on ice for 20 min and centrifuged at 35,000 g for 30 min at 4°C. A 5 mL HiTrap® Chelating High Performance column (GE Healthcare, Chicago, Illinois, U.S.) was connected with a ÄKTA start protein purification system (GE Healthcare) and charged with $NiCl_2$ followed by equilibration with cell resuspension buffer. The soluble protein fraction in supernatant was loaded onto the column and washed with the cell resuspension buffer. Proteins were eluted by a linear gradient from 0% to 100% Ni buffer B (25 mM Tris pH 7.5, 500 mM NaCl, 400 mM imidazole, 10% glycerol) and a graph of absorbance at 280 nm (A_{280}) against eluted fractions was given by the ÄKTA start system. Eluted fractions was run on 8% SDS polyacrylamide gel for analysis (For SDS-polyacrylamide gel electrophoresis, SDS-PAGE, see section 2.3). Fractions containing the desired protein were dialysed at 6°C against heparin buffer A (25 mM Tris pH 7.5, 150 mM

NaCl, 10% glycerol) overnight. Then the protein sample was loaded onto a 1 mL HiTrap™ Heparin HP column (GE Healthcare) and washed by heparin buffer A. The proteins were eluted by a linear gradient from 0% to 100% heparin buffer B (25 mM Tris pH 7.5, 1 M NaCl, 10% glycerol). Fractions containing desired protein were identified using Coomassie staining of a 8% SDS-PAGE gel, and these fractions were concentrated by a Microsep™ Advance centrifugal device (VWR, Radnor, Pennsylvania, U.S.) with 100 kDa molecular weight cut off (MWCO) for fusion proteins and 10 kDa MWCO for casposase. The buffer was also exchanged at this step to storage buffer (25 mM Tris pH 7.5, 500 mM NaCl, 15% glycerol) and the concentrated proteins were stored at -80°C. All purified proteins were listed in table 2.8.

The purification step of Cas9 and dCas9 was the same as above. After the heparin column, the protein sample was added with 1 mg of TEV protease (Sigma-Aldrich) and was dialysed at 6°C against TEV buffer (25 mM Tris pH 8.0, 500 mM NaCl, 10% glycerol) overnight to remove the MBP tag. The cleaved protein sample was concentrated by 100 kDa MWCO Microsep™ Advance centrifugal device and then subjected to size exclusion chromatography on a HiPrep™ 16/60 Sephacryl® S-300 HR column (GE Healthcare) using the protein storage buffer. SDS-PAGE was run for analysis and fractions containing the protein with correct size were concentrated by a centrifugal device and stored at −80°C.

For purification of NLS-Cas9 and Int^{pTN3}-dCas9, which contained an addition C-terminal strep-tag® II, the pH of all buffers was pH 8.0. Proteins eluted from a 5 mL $NiCl_2$ charged HiTrap® Chelating High Performance column was run through 2 mL of Strep-Tactin® Macroprep® resin (IBA Lifesciences, Göttingen, Germany). The column was washed with buffer W (100 mM Tris pH 8.0, 300 mM NaCl) and protein was eluted

with buffer E (100 mM Tris pH 8.0, 300 mM NaCl, 5mM desthiobiotin). Fractions containing the protein with correct size were concentrated and stored at −80°C.

Table 2.8. Proteins purified in this study

Protein	**Expressing plasmid**	**Purification steps**
Casposase	pCHL2	$NiCl_2$ column then heparin column
Casposase^{H242A}	pCHL13	$NiCl_2$ column then heparin column
Casposase^{D254A}	pCHL14	$NiCl_2$ column then heparin column
Cas9	pMJ806	$NiCl_2$ column, heparin column, TEV protease cleavage, S-300 gel filtration column
dCas9	pMJ841	$NiCl_2$ column, heparin column, TEV protease cleavage, S-300 gel filtration column
Casposase-(GGS)$_6$-Cas9	pCHL5	$NiCl_2$ column then heparin column
Casposase-(GGS)$_8$-Cas9 (Casp-Cas9)	pCHL6	$NiCl_2$ column then heparin column
Casposase-XTEN-Cas9	pCHl7	$NiCl_2$ column then heparin column
Casposase-(GGS)$_6$-dCas9	pCHL8	$NiCl_2$ column then heparin column

Casposase-$(GGS)_8$-dCas9 (Casp-dCas9)	pCHL9	$NiCl_2$ column then heparin column
Casposase-XTEN-dCas9	pCHL10	$NiCl_2$ column then heparin column
Casposase^{D254A}-$(GGS)_8$-dCas9	pCHL20	$NiCl_2$ column then heparin column
Casp-hfCas9	pCHL29	$NiCl_2$ column then heparin column
Casp-hfnCas9	pCHL30	$NiCl_2$ column then heparin column
Casp-hfdCas9	pCHL31	$NiCl_2$ column then heparin column
NLS-Cas9	pNLS-Cas9	$NiCl_2$ column then Strep-Tactin® resin
NLS-Casp-hfnCas9	pNLS-hfnCaspoR	$NiCl_2$ column then heparin column
NLS-Casp-hfdCas9	pNLS-hfdCaspoR	$NiCl_2$ column then heparin column
Casp-dCasX	pNLS-dCaspoX	$NiCl_2$ column then Strep-Tactin® resin
Int^{pTN3}	pCHL11	$NiCl_2$ column then heparin column
Int^{pTN3}-dCas9	pCHL49	$NiCl_2$ column then Strep-Tactin® resin

2.3. SDS-PAGE and western blot

Protein containing samples were mixed with 4x SDS loading buffer (120 mM Tris pH 6.8, 8% SDS, 20% glycerol, 0.5% bromophenol blue) and 30 mM DTT, boiled at 95°C for 10 minutes and were run through an 8% SDS polyacrylamide gel in buffer (25 mM

Tris, 190 mM glycine, 0.1% SDS) at 140V for 70 min. To visualise proteins present, gels were stained using staining buffer (10% acetic acid, 40% methanol, 1 g/L brilliant blue R-250) and de-stained using de-staining buffer (10% acetic acid, 20% methanol).

For western blotting, proteins in the gel were transferred onto an Amersham Hybond Polyvinylidene fluoride (PVDF) membrane (GE Healthcare). The wet transfer process was carried out in transfer buffer (25 mM Tris, 190 mM glycine, 5% methanol) at 30 V at 6°C overnight. The membrane was blocked with 5% milk in TBST buffer (20 mM Tris pH 7.5, 150 mM NaCl, 0.1% Tween 20) at RT for one hour following three washes with TBST. Then the membrane was incubated with 1:1000 biotin-conjugated anti-6xHis antibody (Invitrogen, Carlsbad, California, U.S.) in TBST containing 5% BSA and 0.05% sodium azide at RT for one and a half hours. The membrane was washed with TBST three times and subsequently incubated with 1:3000 HRP-conjugated anti-biotin antibody (Cell Signaling Technology, Danvers, Massachusetts, U.S.) in TBST containing 5% skimmed milk at RT for one hour. After three washes of TBST, the proteins on the membrane were visualised using Immobilon® Forte Western HRP Substrate (Merck Millipore, Burlington, Massachusetts, U.S.) to capture and the chemiluminescence in a Fujifilm LAS-3000 Imager (Fujifilm, Minato City, Tokyo, Japan).

2.4. Measurement of protein and oligonucleotide concentration

The concentration of purified proteins and oligonucleotides was determined using the Beer–Lambert law as shown below:

$$c = \frac{A}{\varepsilon \times l}$$

Where c is the concentration and A is the absorbance at 280 nm (A_{280}) for protein and at 260 nm (A_{260}) for DNA and RNA. The A_{280} and A_{260} were measured by a NanoDrop™

spectrophotometer (Thermo Scientific, Waltham, Massachusetts, U.S.). l is the path length in cm and the value is 1 in this case. ε is the extinction coefficient which of proteins was calculated using protein amino acid sequences and ExPASy ProtParam (https://web.expasy.org/protparam/) and of oligonucleotides was calculated using oligonucleotide sequences and OligoAnalyzer (Integrated DNA Technologies, https://eu.idtdna.com/calc/analyzer/).

2.5. Generation of nucleic acid substrates

2.5.1. Oligonucleotides

Synthetic oligonucleotides were purchased from Sigma at 0.2 µmole scale and delivered purified by High-performance liquid chromatography. Oligonucleotides used in this project were listed in table 2.9.

Table 2.9. Oligonucleotides used in this project.
Oligonucleotides used in human cells are highlighted with the bases showing introduced mutations.

Name	**Sequence from 5′ to 3′**
Cy5 MW14	/Cy5/CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA
MW14	CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA
MW12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
PM16	TGCCGAATTCTACCAGTGCCAGTGAT
EW3	TCGGATCCTCTAGACAGCTCCATGTAGCAATGTAATCGTCTATGACGTTG

CL1	CGGTCACGTCTATCCCCACTACGAGGAGAGTCTTGAGTGTAAAATTGT
CL2	/Cy5/AGGGGATGTATATATATCCCCCTCCTCGTAGTGGGGATAGACGT GACCG
CL3	ACAATTTTACACTCAAGACT
Cy5 PM32	/Cy5/TGTAATCGTCTATGACGTT
Cy5 LE30 top	/Cy5/GGGGATATATATACATCCCCTCTTAAGTTC
LE30 top	GGGGATATATATACATCCCCTCTTAAGTTC
Cy5 LE30 atk	/Cy5/GAACTTAAGAGGGGATGTATATATATCCCC
LE30 atk	GAACTTAAGAGGGGATGTATATATATCCCC
Cy5 TK24	/Cy5/GCAGTCCCCTCGCCTCAGCTACGCTCGT
TK24	GCAGTCCCCTCGCCTCAGCTACGCTCGT
TK25	CGTAGCTGAGGCGAGGGGACTGCTGGGC
FAM TK24	/6-FAM/GCAGTCCCCTCGCCTCAGCTACGCTCGT
CL4	GCCGAATTCTACCAGTGCCAGTGATCATGGAGCTGTCTAGAGGATCCG A
CL5	GCAGTCCCCTCGCCCAGTACGCTCG
CL6	CGTACTGGGCGAGGGGACTGCTGGG
CL7	GGGTAACGCCAAGGGTTTTCCCAGTC
ssODN for Cas9	AGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGT GACCACCCTGAGCCACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCAC ATGAAGCAGCACGACTTCTTCA
EGFP CaspoR oligo	TAGCGGCTGAAGCACTGCACGCCGTGGCT

site B hairpin	ATCAGCGCTAGCGGAGTGTATACTGGCTTACTATGTTGGCACTGATGAG GGTGTCAGTGAAGTGCTTCATGTGGCGAAAGCCACATGAAGCACTTCA CTGACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTA GCGCTGAT

2.5.2. Assembly and purification of short double-stranded substrates

To make double-stranded DNA substrates containing Cyanine5 (Cy5) labelling, 5μM of C5 oligo was mixed with 6μM unlabelled oligo(s) in annealing buffer (10 mM Tris pH 7.5, 50 mM NaCl, 1 mM EDTA). The mixture was heated at 95°C for 10 min and slowly cooled down to RT. The annealed substrate was separated from the unannealed oligos by running the mixture on an 8% native polyacrylamide gel in 1x TBE buffer. Substrate containing band was cut out and eluted in elution buffer (20 mM Tris pH 8.0, 50 mM NaCl).

Unlabelled double-stranded DNA substrates were made from mixing equal molar concentrations of complementary oligos in annealing buffer. The mixture was heated to 95°C for 10 min and cooled to RT for at least three hours. The substrates were stored at -20°C and the concentration of DNA substrates was measured at A_{260} using NanoDrop™ spectrophotometer (see section 2.4).

2.5.3. Generation of Cas9 sgRNA

The ssDNA in table 2.10 was amplified by PCR to generate dsDNA containing a T7 promoter, a guide sequence and a scaffold of the sgRNA. Then the sgRNA was transcribed from the dsDNA by a HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). The reaction was treated with RNase-free RQ1 DNase (Promega, Madison, Wisconsin, U.S.)

to remove the template DNA and the sgRNA was run on a pre-run 8% polyacrylamide, 7M urea denaturing gel in 1x TBE buffer at 10 W for two and a half hours. The sgRNA was visualised by UV shadowing and the band was excised for elution in RNase-free water at 4°C for 24 hours prior to ethanol precipitation (see section 2.6). The concentration of sgRNA was measured at A_{260} (see section 2.4).

Table 2.10. ssDNAs that were used to generate Cas9 sgRNA.
Bases in red were T7 promoter sequence and bases in light blue were spacer DNA sequence.

Name	Sequence from 5′ to 3′
sgMW14 ssDNA	TAATACGACTCACTATAGGTAGACGATTACATTGCTACAGTTT TAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTAT CAACTTGAAAAAGTGGCACCGAGTCGGTGCTT
sgpACYC ssDNA	TAATACGACTCACTATAGGAGCGCTAGCGGAGTGTATACGTT TTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTA TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTT
sgpUC19 ssDNA	TAATACGACTCACTATAGGGTGCTGCAAGGCGATTAAGTGTT TTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTA TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTT
sgpCHL42 ssDNA	TAATACGACTCACTATAGGATCTGCTGATGGGTAGGGAGGTT TTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTA TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTT
sgEGFP HDR ssDNA	TAATACGACTCACTATAGGCACTGCACGCCGTAGGTCAGTTTT AGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATC AACTTGAAAAAGTGGCACCGAGTCGGTGCTT

sgEGFP CaspoR ssDNA	TAATACGACTCACTATAGGCTTCATGTGGTCGGGGTAGGTTTT AGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATC AACTTGAAAAAGTGGCACCGAGTCGGTGCTT

2.6. Ethanol precipitation of nucleic acid

One tenth volume of 3 M sodium acetate, pH 5.2 was added to DNA and 3 times volumes of 100 % absolute ethanol was added to the mix. After overnight storage at -20°C, the mix was centrifuged at 13,000 rpm for 30 min at 4°C. The supernatant was discarded and the pellet was washed with 70 % ethanol followed by centrifugation at 13,000 rpm for 30 min at 4°C. The supernatant was discarded and the pellet was air-dried followed by resuspension in sterile distilled water.

2.7. Protein biochemical assays

2.7.1. Disintegration

Fluorescent DNA fork 3 substrate was formed by annealing a 5' Cy5 labelled MW14 to unlabelled MW12 and PM16. Fork casposon was generated by annealing Cy5 labelled CL2 to CL1 and CL3. The disintegration assay was carried out by incubating proteins (concentrations indicated in the relevant figures) with 25 nM fork substrates in an integration buffer (final concentration of 25 mM Tris pH 7.5, 106 mM KCl, 50 µg/ml BSA, 5 mM $MnCl_2$, 2 mM DTT, 9% glycerol) for 1 hr at 37°C. The reaction was then quenched by incubating with proteinase K stop buffer to a final concentration of 10 mM Tris pH8, 24 U proteinase K (Invitrogen), 20 mM EDTA at 37°C for 30 min. Formamide loading dye (80 % formamide, 0.25 % orange G, 20 % glycerol) and 100 nM unlabelled MW14 were added to reactions and the samples were heated at 95°C

for 10 min. Reaction products were run on a pre-run 20% 7 M urea polyacrylamide gel in 1x TBE at 7 W for two hours. The gel was analysed by a fluorescent image analyser FLA-3000 (Fujifilm).

2.7.2. DNA oligos integration

Integration assays which contained fluorescent DNA substrates, were carried out by incubating 150 nM proteins with 100 nM fluorescent DNA substrates and 150 ng plasmid DNA in an integration buffer (see section 2.7.1) for one hour at 37°C. In the case of Cas9 fusion proteins, three-fold molar excess of sgRNA was first incubated with the fusion proteins at RT for 10 min to form RNP complexes. Target plasmid was then added to the RNP complexes and incubated at 37°C for 20 min. DNA substrates were added and the reaction was incubated at 37°C for one hour before being quenched by proteinase K stop buffer at 37°C for 30 min. Reaction products were run on a 0.8 % agarose gel in 1X TBE at 120 V for 70 min. The gel was imaged using an Amersham Typhoon Biomolecular Imager (GE Healthcare) to detect Cy5 signal for observing assay products. After this step the gel was stained with ethidium bromide to observe all DNA by a UV-transilluminator (Syngene). Bands corresponding to DNA from integrated plasmids were excised and DNA extracted using Qiagen gel extraction kit. PCR reactions were carried out as described in section 2.1.1 to amplify integrated region for assessment on a 0.8 % agarose gel. Bands corresponding to DNA from targeted integration were excised and DNA extracted followed by sending to Sanger sequencing.

In integration reactions using site B hairpin as a target, 150 nM proteins were first incubated with 450 nM sgpACYC RNA followed by incubation with 450 nM site B

hairpin. Cy5 labelled DNA substrate TK2425 at 30 nM was added and incubated at 37°C for up to 60 min followed by proteinase K treatment at 37°C for 30 min. Reaction products were heated at 95°C for 10 min and run on 10% polyacrylamide, 7 M urea denaturing gel in 1x TBE at 10 W for two and a half hours. The gel was imaged by Amersham Typhoon Biomolecular Imager to detect a band shift corresponding to DNA integration.

For integration assays using unlabelled DNA substrates, the plasmid DNA was ethanol precipitated as described in section 2.6 after proteinase K treatment to remove proteins from reactions. Plasmid DNA was then electroporated into *E. coli* to select for integrated plasmid following the method described in section 2.1.6.2, or PCR reactions were carried out to detect integrated products as described in section 2.1.1. When a target plasmid pUC19 was used, a blue/white screening was carried out when the integrated plasmids were transformed into *E. coli*. The transformed cells were plated on LB agar plates containing antibiotics, 20 µg/ml X-gal, 1 mM IPTG. The plates were incubated at 37°C overnight and blue and white colonies appeared the next day.

2.7.3. Integration of longer DNA molecules

The amp^R gene was amplified from pCaspamp and by using different primers, the amp^R gene was flanked by *A. boonei* casposon TIRs (primers: Caspamp F and Caspamp R) or AatII restriction site at both ends (primers: AmpR F and AmpR R). After AatII (NEB) restriction digestion of the amp^R gene, the resulting amp^R gene had 4 nucleotides 3' overhangs at both ends. Integration reactions were carried out in a 1:1 molar ratio of plasmid: insert. Proteins at 150 nM were incubated with 150 ng pACYC-Duet and 45 ng long DNA molecules in 1x integration buffer at 37°C overnight. In the case of Cas9 fusion proteins, three-fold molar excess of sgRNA was first incubated with the fusion

proteins at RT for 10 min to form RNP complexes. Target plasmid was then added to the RNP complexes and incubated at 37°C for 20 min. Long DNA molecules were added to reactions and incubated at 37°C overnight. The plasmid DNA was ethanol precipitated as described in section 2.6 after proteinase K treatment of reactions. Then resuspended plasmid DNA was electroporated into *E. coli* DH5α as described in section 2.1.6.2 to select for integrated plasmid products or PCR reactions were carried out to detect integrated products using OneTaq® Hot Start DNA Polymerase (see section 2.1.1). If the plasmid was purified from reactions using Casp-Cas9, the plasmid was ligated using T4 DNA ligase before electroporation.

2.7.4. Casposon excision assay

Casposase at 150 nM was incubated with 150 ng pCaspamp, 1x integration buffer at 37°C for one hour. The reaction was then quenched by incubating with proteinase K stop buffer at 37°C for 30 min. Reaction products were run on a 0.8% agarose gel in 1x TBE at 120 V for 70 min. The gel was stained with ethidium bromide to detect excised mini-casposon by a UV-transilluminator.

2.7.5. Fluorescence anisotropy measurements

A two-fold serial dilution of casposase ranging from 0 to 900 nM was incubated with 3 nM of 6-FAM labelled single-stranded TK24 or 6-FAM labelled TK24 annealed to TK25 (TK2425) in an integration buffer. After 10 min, the fluorescent polarisation of the labelled substrates was measured by EnVision™ multilabel plate reader (PerkinElmer, Waltham, Massachusetts, U.S.). The EnVision™ plate reader was set at 480 nm for

excitation and at 535 nm for emission scanning. The millipolarisation unit (mP) was calculated using the formula shown below:

$$mP = 1000 \times [(I_v — GI_h) / (I_v + GI_h)]$$

Where Iv and Ih are parallel and perpendicular emission intensity measurements corrected for background (buffer only), G is G-factor.

Millipolarisation values at different casposase concentrations were normalised to the substrate only control (0 nM casposase) to obtain the values of change in mP. The values of change in mP were plotted against $\log_{10}$ scale of casposase concentrations and the equilibrium dissociation constant K_d was identified at the 50% of change in mP.

2.7.6. Electrophoretic Mobility shift assays (EMSAs)

A concentration titration of casposase (0, 25, 50, 75, 150, 250 nM) was incubated with 25 nM of Cy5 labelled fork substrates (fork 3 and fork casposon) in a binding buffer (final concentration of 25 mM Tris pH 7.5, 106 mM KCl, 50 µg/ml BSA, 5 mM EDTA, 2 mM DTT, 9% glycerol) at 37°C for 20 min. Reactions were loaded on a 5% native polyacrylamide gel and run at 120 V for two hours. The gel was imaged by Amersham Typhoon Biomolecular Imager to detect a band shift.

For binding to double-stranded MW14 (Cy5 MW14+EW3), proteins (casposase, Cas9 and fusion proteins) at concentrations (75, 150, 300, 500, 1000 nM) were incubated with three-fold molar excess of sgMW14 RNA at RT for 10 min to form RNP complexes. Then 25 nM of Cy5 labelled dsMW14 and 75 nM unlabelled dsMW12 (MW12+CL4) were added to the RNP complexes in a binding buffer at 37°C for 20 min. Reactions

were loaded on an 8% native polyacrylamide gel and run at 120 V for two hours. The gel was imaged by Amersham Typhoon Biomolecular Imager to detect a band shift.

For binding to pUC19, a concentration titration (40, 75, 150 and 300 nM) of casposase, dCas9, Casp-dCas9 were incubated with three-fold molar excess of sgpUC19 RNA at RT for 10 min. Then 150 ng of pUC19 and 100 nM of dsMW12 were added and incubated in a binding buffer at 37°C for 20 min. Reactions were run on a 0.8% agarose gel in 1x Tris-acetate-EDTA (TAE) buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) at 10 V overnight.

2.7.7. Assay for R-loop formation

A concentration titration (75, 150, 300, 500, 1000 nM) of Cas9, dCas9 and fusion proteins were incubated with three-fold molar excess of sgMW14 RNA at RT for 10 min to form RNP complexes. Then 25 nM of Cy5 labelled dsMW14 and 75 nM unlabelled dsMW12 were added to the RNP complexes in a binding buffer at 37°C for 20 min. Subsequently, a proteinase K stop buffer was added to remove proteins and reactions were loaded on an 8% native polyacrylamide gel and run at 120 V for two hours. The gel was imaged by Amersham Typhoon Biomolecular Imager to detect a band shift.

Cas9 and Casp-hfdCas9 fusion protein at 150 nM was first incubated with 450 nM sgpACYC RNA followed by incubation with 450 nM site B hairpin at 37°C for 20 min. Then, a proteinase K stop buffer was added to remove proteins and reactions were loaded on an 10% native polyacrylamide gel and run at 120 V for two hours. DNA was stained by Diamond™ Nucleic Acid Dye (Promega) and visualised by a UV-transilluminator.

2.7.8. DNA recombination assays catalysed by Int^{pTN3}

Int^{pTN3} at concentrations 0, 500, 2500 nM was incubated with pCHL43 in a recombination buffer containing 25 mM Tris pH 8.0, 300 mM KCl, 1 mM DTT, 5 mM $MgSO_4$ at 65°C for three and a half hours. Then the reaction was added with a proteinase K stop buffer and incubated at 37°C for 30 min. Reactions that used pCHL43 for detecting deletion were run on a 0.8% agarose gel in 1X TBE at 120 V for 70 min. The gel was stained with ethidium bromide to detect excised fragments by a UV-transilluminator.

For Int^{pTN3}-dCas9, 300 nM of fusion protein was incubated with 900 nM sgpUC19 RNA at RT for 10 min. Then the RNP complex was incubated with pCHL43 or pCHL44 in a recombination buffer at 37°C overnight. Reactions that used pCHL43 for detecting deletion were run on a 0.8% agarose gel in 1X TBE at 120 V for 70 min. The gel was stained with ethidium bromide to detect excised fragments by a UV-transilluminator. In reactions using pCHL44 for detecting inversion, plasmids were ethanol precipitated as described in section 2.6 and PCR was carried out to detect inversion using OneTaq® Hot Start DNA Polymerase as described in section 2.1.1. PCR products was run on a 0.8% agarose gel and stained with ethidium bromide.

2.7.9. Gel analysis and statistical analysis of experiments

The band intensity on gels was quantified using ImageJ (National Institutes of Health, Bethesda, Maryland, U.S.) and normalised to a control on each gel. Graphs were plotted using GraphPad Prism version 8.4.0 (GraphPad Software, San Diego, California, U.S.). Statistical tests were also performed using GraphPad Prism. A one-way ANOVA was used to determine statistical differences among the means of two or more groups

with one factor. A two-way ANOVA was used to determine statistical differences among the means of two or more groups with two factors.

2.8. *In vivo* assays in bacteria

2.8.1. Casposon excision assay

E. coli strain EB377 was transformed with pCHL41 that contained a mini-casposon, an amp^R gene flanked by *A. boonei* casposon TIRs. The pCHL41 also encoded a *casposase* gene and cm^R gene. The transformed cells were plated on a LB agar plate containing 25 μg/ml chloramphenicol, 0.1% L-arabinose and 0.5 mM IPTG. The plate was incubated at 37°C for 2 days and then colony PCR was carried out to detect excision of mini-casposon.

2.8.2. Miller assay

For *in vivo* targeting *lacZ* gene on a plasmid, *E. coli* strain CL003 was transformed with a pRC7 plasmid containing *lac* operon and a plasmid encoding a sglacZ RNA and a protein. For *in vivo* targeting lacZ gene on a chromosome, *E. coli* strain EB377 was transformed with a plasmid encoding a sglacZ RNA and a plasmid encoding a protein. A single colony was inoculated into 5mL LB containing antibiotics to establish an overnight culture. The overnight culture was diluted 100-fold in 5 mL fresh LB containing antibiotics and grown at 37°C until OD_{600} reached 0.2. Then inducers (0.1% L-arabinose and 0.5 mM IPTG final concentration) were added to the culture to induce protein expression and the culture was grown until OD_{600} reached 0.6.

Cells were pelleted and resuspended in Z buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol). The cells were diluted 10^6, 10^7,

10^8-fold using Z buffer and 10μL of diluted cells were spotted on agar plates containing antibiotics to check cell viability.

The protocol of Miller assay was modified from the original protocol (Miller, 1972). Cells were pelleted and resuspended in the same volume of cold Z buffer. The OD_{600} of cells in Z buffer was recorded. Cells were diluted two-fold in Z buffer to a final volume of 1 mL and the diluted cells were permeabilised with 100 μL chloroform and 50 μL 0.1% SDS. Then 200 μL of 4 mg/mL ONPG in phosphate buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, pH 7.0) was added to the cells and incubated at 28°C for 12 min. Reactions were stopped by adding 0.5 mL 1 M Na_2CO_3. Reactions were centrifuged to remove debris and chloroform and OD_{420} and OD_{550} of the supernatant was recorded. The Miller unit was calculated as shown in the equation below.

$$\text{Miller units} = 1000 \times \frac{\text{OD}_{420} - 1.75 \times \text{OD}_{550}}{\text{T} \times \text{V} \times \text{OD}_{600}}$$

Where T is the reaction time in min, 12 in this case and V is the volume of culture used in the assay in mL, 0.5 in this case.

2.8.3. DNA integration assays

For casposase *in vivo* DNA integration, EB377 was transformed with pCHL2. A single colony was picked to establish an overnight culture. A fresh culture was set up using the overnight culture and grown to OD_{600}=0.6. Inducers at final concentrations of 0.5 mM IPTG and 0.1 % L-arabinose were added to the cells for protein expression and the culture was grown for three hours. The cells were then made electrocompetent as described in 2.1.6.2 and 150 ng of mini-casposon was electroporated into cells.

After one-hour recovery in SOC, cells were plated on LB agar plate containing 50 μg/ml ampicillin and 25 μg/ml chloramphenicol.

For Casp-Cas9 fusion proteins, the cells were cultured and induced protein expression as in Miller assay (section 2.8.2). The cells were made electrocompetent at OD_{600}=0.6 (section 2.1.6.2). For short oligo duplex integration, 3 μM final concentration of duplex DNA was electroporated into cells. For mini-casposon integration, 150 ng of mini-casposon was electroporated into cells. After one-hour recovery in SOC, cells were diluted 10^7-fold if the integrated product could not be selected for and cells were spread on plates containing the same antibiotics. For mini-casposon integration, all cells were spread on plates containing the same antibiotics plus 50 μg/ml ampicillin.

2.9. Experiments using human cells

2.9.1. Preparation of human cell-free extracts

Human osteosarcoma U2OS cells were seeded on four 10 cm petri dishes in Dulbecco's Modified Eaglr Medium (DMEM, Lonza, Basel, Switzerland) supplemented with 2 mM L-glutamine, 10% fetal bovine serum (FBS), 1x pen/strep (10 U/mL penicillin and 10 μg/mL streptomycin) and incubated at 37°C, 5% CO_2. When the cell density reached 70-80% confluency, the cells were washed with phosphate buffered saline (PBS, Sigma-Aldrich) and incubated with trypsin-EDTA solution (0.5 g/L trypsin, 0.2 g/L EDTA·4Na, Sigma) at 37°C for 5 min to detach cells. The DMEM complete medium was added to inhibit trypsin reaction and cells from 4 dishes were pooled together and centrifuged at 300 g for 10 min. The cell pellet was resuspended in hypotonic buffer (20 mM HEPES pH7.5, 5 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT) containing 1x protease inhibitor cocktail. Cells were pelleted again and resuspended in 500 μL of hypotonic

buffer and incubated on ice for 15 min. A 1 mL syringe and a 25G needle were used to homogenise cells with 20 strokes. The cell lysate was left on ice for 60 min to allow release of nuclear proteins and then centrifuged at 12,000 g for 15 min. The supernatant containing proteins was aliquoted and stored at −80°C.

2.9.2. Generation of a GFP expressing cell line

U2OS cells at low passage number was seeded in a 6-well plate. When the cell confluency reached 70-90%, the medium was changed to fresh one. pEGFP-c1 plasmid was a gift from Ronald Chalmers's lab and it was transfected into U2OS cells using lipofectamine 3000 (Invitrogen). In brief, 2.5 μg pEGFP-c1 diluted in Opti-MEM medium (Gibco, Thermo Fisher, Waltham, Massachusetts, U.S.) and 3.75μL of lipofectamine 3000 reagent diluted in Opti-MEM medium were mixed together and incubated at RT for 10 min. Then the lipid-DNA complex was added to cells dropwise and incubated at 37°C, 5% CO_2. After 2 days incubation, successful transfection and GFP expression was inspected under a fluorescent microscope. The medium was changed to DMEM complete medium containing 400 μg/mL G418 (geneticin, Sigma) to select for the plasmid. The cells were transferred in a T25 flask after reaching 70-90% confluency and subsequently transferred to a T75 flask.

2.9.3. Ribonucleoprotein (RNP) transfection into cells

2.9.3.1. Neon transfection system

RNP complexes were electroporated into U2OS cells by Neon transfection system using a 10 μL kit (Thermo Fisher) and following manufacturer's instruction. NLS-Cas9 at 13.7 pmoles was incubated with 15 pmoles of sgEGFP HDR RNA at RT for 10 min.

Then 1 μg of single-stranded oligodeoxynucleotide (ssODN) for Cas9 was added to the RNP complex and reaction volume was filled up to 7.5 μL using resuspension buffer R from the 10 μL kit. For NLS-Casp-hfnCas9 and NLS-Casp-hfdCas9 proteins, 1 pmol of proteins were incubated with 1 pmol of sgEGFP CaspoR RNA. The hfnCas9 fusion protein RNP complex was added with 0.5 μg of EGFP oligo for nCaspoR. The hfdCas9 fusion protein RNP complex was added with 0.5 μg of TK2425.

GFP expressing U2OS cells at 70-90% confluency was trypsinised and resuspended in DMEM complete medium. Cell number was counted by using a haemocytometer and 2 million cells were transferred into an Eppendorf tube. Cells were centrifuged at 300 g for 10 min to remove medium and washed with PBS. Cells were spun again to remove PBS and resuspended in resuspension buffer R. Then $2x10^5$ cells in 5 μL were mixed with 7.5 μL of RNP complexes and 10 μL of each RNP-cell mix was electroporated in a 10 μL Neon tip at 1230 V, 10 ms, 4 pulses. The electroporated cells were transferred into a well in a 24-well plate and cultured in DMEM complete medium without antibiotics pen/strep and G418. The cells were inspected 24 hours post-transfection.

2.9.3.2. jetCRISPR® transfection reagent

RNP complex at 330 nM was prepared as described in 2.9.3.1 at RT for 10 min in 50 μL of serum-free DMEM medium. Then 1 μg of ssODN for Cas9, 0.5 μg of EGFP oligo for nCaspoR, 0.5 μg of TK2425 were added to NLS-Cas9, NLS-Casp-hfnCas9, NLS-Casp-hfdCas9 RNPs respectively. JetCRISPR® reagent in 1.2 μL (Polyplus, New York, U.S.) was added to the RNPs and incubated at RT for 15 min. Each RNP sample was dispensed in a well in a 24-well plate and then $1.5x10^5$ GFP expressing U2OS cells in DMEM

complete medium were added into wells and mixed. The plate was incubated at 37°C and analysed after 24 hours.

2.9.4. T7 endonuclease I assay

When post-transfected cells in a 24-well plate reached 60-80% confluency, pEGFP-c1 plasmid was purified from cells using QIAprep® Spin Miniprep Kit (Qiagen). A EGFP DNA fragment was PCR amplified from the purified pEGFP-c1. Then 200 ng of PCR products were heated at 95°C for 10 min and slowly cooled to room temperature. T7 endonuclease I (NEB) was added to the reannealed PCR products and incubated at 37°C for 30 min before running on a 0.8% agarose gel to detect cleavage of PCR products.

Chapter 3

3. Biochemical characterisation of *A. boonei* casposase

3.1. Introduction

Recent studies suggested that the CRISPR repeat and adaptation module originated from a novel type of DNA transposons called casposons (Krupovic et al., 2014). Casposons encode a Cas1 homologue called casposase that shows integrase activity and catalyses casposon transposition (Hickman and Dyda, 2015). Casposase integrates casposon in a similar way to CRISPR spacer integration carried out by the Cas1-Cas2 complex. The terminal inverted repeats (TIR) at both ends of a casposon allow casposase to bind and the 3' ends are used in nucleophilic attack during transposition. Family 2 casposons are frequently found within 3' end of tRNA genes which plays similar role as leader sequence in CRISPR locus. Recent biochemical studies showed that *A. boonei* and *M. mazei* casposase preferentially targets the 3' end of tRNA-Pro gene and tRNA-Leu gene respectively (Béguin et al., 2016; Hickman et al., 2020). If there is no preferred target site, *A. boonei* casposase will integrate randomly (Hickman and Dyda, 2015). Upon casposon integration, a 14-15 bp target site is duplicated similar to expansion of CRISPR repeat.

How Cas1-Cas2 complex mediates spacer acquisition has been extensively studied in the CRISPR immune systems (Ivančić-Baće et al., 2015; Modell et al., 2017; Staals et al., 2016), however, much remains unknown for casposons as only *A. boonei* and *M. mazei* casposases have been biochemically characterised. There is no *in vivo* study of casposons thus far but recent mobility of casposons was detected by analysing the genome of 62 *M. mazei* strains (Krupovic et al., 2016). Here we studied *A. boonei* casposase in more depth via biochemical assays and expressed *A. boonei* casposase in

recombinant *E. coli* to try to establish an understanding about how casposase mediates integration *in vivo*. From this, we can understand more about the evolutionary relationship of casposase and CRISPR Cas1.

3.2. Results

3.2.1. *A. boonei* casposase structure and purification

A. boonei casposase was the first casposase biochemically characterised but there is no crystal structure of *A. boonei* casposase to date (Hickman and Dyda, 2015). Thus its protein monomer structure was predicted by RaptorX (Källberg et al., 2012). The predicted structure was largely similar to the published *M. mazei* casposase structure (Hickman et al., 2020) (Figure 17A). The misalignment at the N-terminal domain might be owing to conformational change induced when *M. mazei* casposase bound to a DNA substrate. The C-terminal domain containing the active site of the two proteins aligned very well, but *M. mazei* casposase structure lacked a C-terminal HTH domain, because the loop containing the HTH domain was disordered. The *A. boonei* casposase predicted structure was also aligned to CRISPR type I-E Cas1 from *E. coli* for comparison (Figure 17B). Overall, the predicted structure of *A. boonei* casposase was very similar to CRISPR Cas1 which also contains a N-terminal β-strand domain and a C-terminal α-helical domain. In addition, casposase contains all four highly conserved, previously identified Cas1 catalytic residues (Nuñez et al., 2014) and these four catalytic residues E168, H242, D254, E257 are in the same position superimposed on *E. coli* Cas1 structure (Figure 17B and Appendix 2). However, casposases from family 2 casposons contain an additional long disordered loop of about 70 amino acids at the C-terminal compared with CRISPR Cas1 (Hickman and Dyda, 2015; Krupovic et al.,

2014). This loop contains a helix-turn-helix (HTH) domain that might play a role in integration (Hickman et al., 2020).

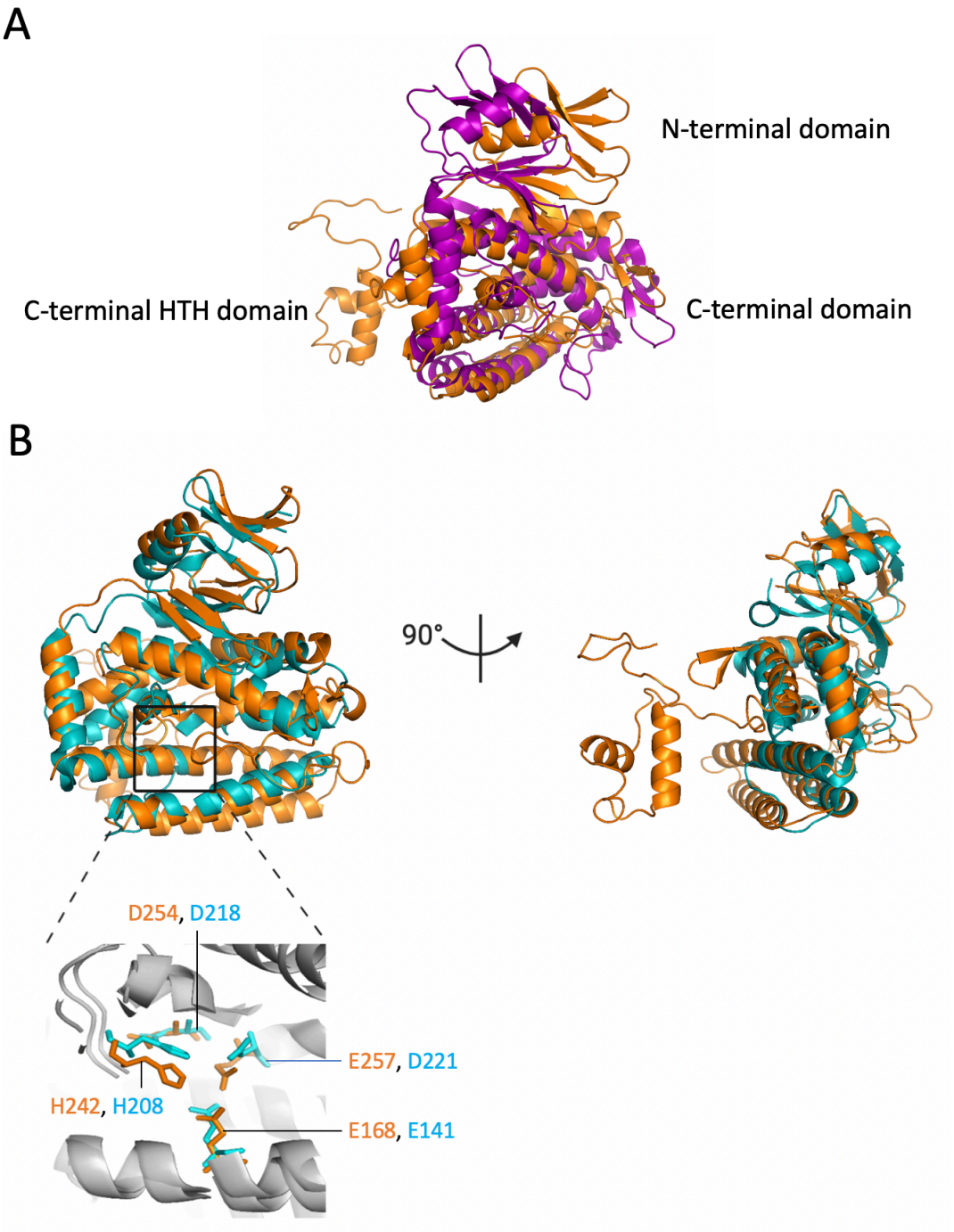


Figure 17. Comparison of A. boonei predicted structure to published M. mazei casposase and E. coli Cas1 structures.

(A) RaptorX predicted structure of *A. boonei* casposase (orange) was aligned to published *M. mazei* casposase structure (purple, PDB: 6OPM). (B) RaptorX predicted structure of *A. boonei* casposase (orange) was aligned to the structure of *E. coli* Cas1 monomer(Cyan, PDB: 3NKD). The inset showed enlarged view superimposing casposase active site residues (orange) with those in *E. coli* Cas1 (blue). The images were created using PyMOL.

His-tagged wild-type *A. boonei* casposase and its active site mutants H242A and D254A were purified as described in section 2.2.3. Briefly, the protein was purified by Ni^{2+} charged column as shown in figure 18A, followed by elution from a heparin column. The protein containing fractions were pooled and concentrated. Western blot of anti-6xHis proteins confirmed the purified protein was His-tagged casposase and the casposase comprised 85% of the total protein purified (Figure 18B and 18C).

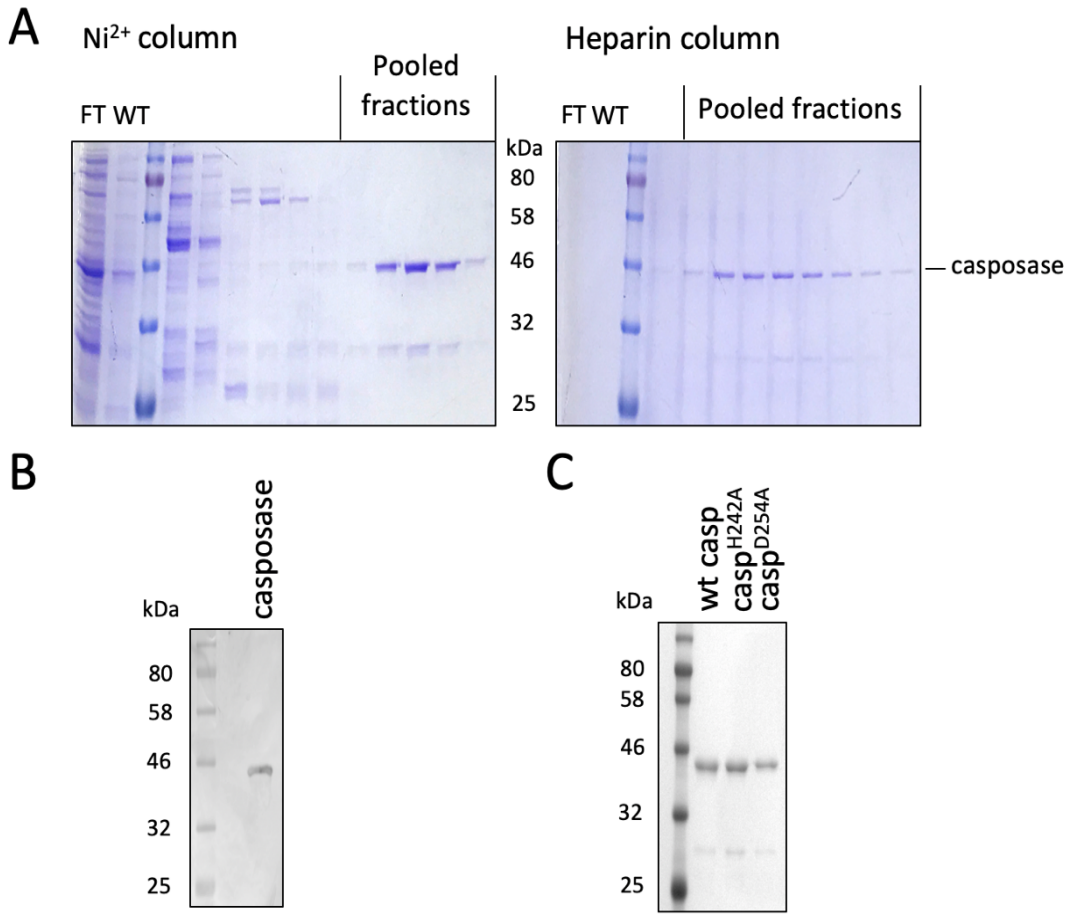


Figure 18. Purification of casposase and active site mutants.
(A) *A. boonei* casposase expressed in *E. coli* BL21-AI strain was purified through Ni^{2+} charged column (left) and then heparin column (right). Fractions containing mostly the target protein were pooled together and spin concentrated after the heparin column. FT: flow through. WT: wash through. (B) Western blot of anti-6xHis proteins confirmed the purified protein was His-tagged casposase. (C) Casposase^{H242A} and casposase^{D254A} active site mutants were purified the same as the wild-type. The concentrated proteins were run on a SDS gel to check purity.

3.2.2. Casposase-catalysed DNA disintegration

CRISPR Cas1 is an integrase and it can perform disintegration, the reverse reaction of integration, of a fork substrate (Rollie et al., 2015). Giving that casposase is homologous to Cas1 and itself is an integrase (Hickman and Dyda, 2015), it was expected that *A. boonei* casposase can carry out disintegration of a DNA substrate (fork 3: MW14+MW12+PM16). During disintegration, the -OH group at the 3' end of the oligonucleotide PM16 bound to the fork initiated a nucleophilic attack to the phosphodiester backbone of MW14 resulting in the release of 5' end of MW14 (Figure 19A). Casposase was incubated with fork 3 and two buffers containing different divalent metal ions were tested (Figure 19B). The result showed that casposase catalysed disintegration much more effectively in the presence of 5 mM Mn^{2+} than in the presence of 5 mM Mg^{2+} and disintegration product was seen at 25 nM of casposase. Therefore this Mn^{2+} buffer was used in subsequent assays for casposase in this study. This result was consistent with previous studies which showed higher enzymatic activity in Mn^{2+} for *A. boonei* casposase and *Pseudomonas aeruginosa* Cas1 (Hickman and Dyda, 2015; Wiedenheft et al., 2009). On the other hand, casposase^{H242A} and casposase^{D254A} active site mutants cannot carry out disintegration confirming the role of these catalytic residues (Figure 19C).

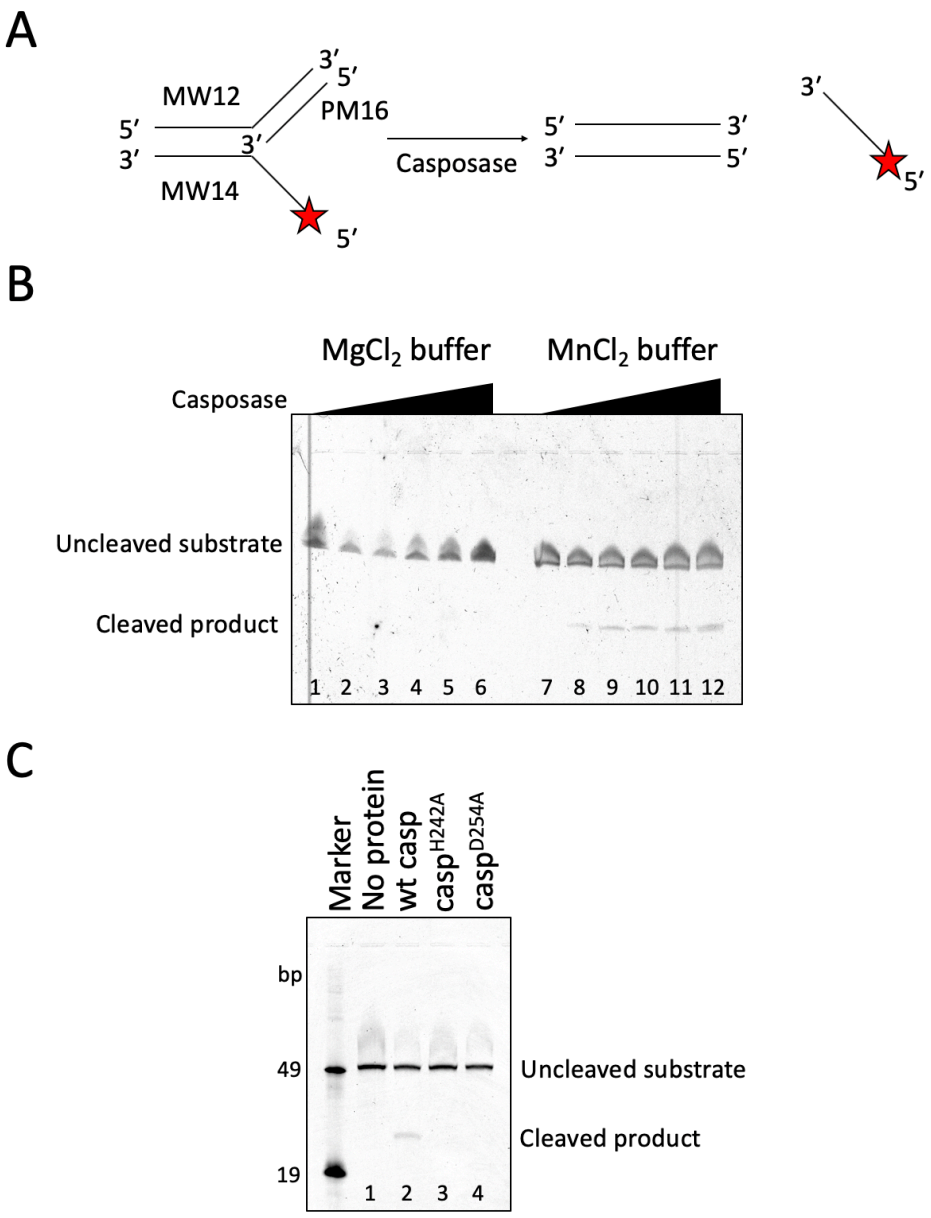
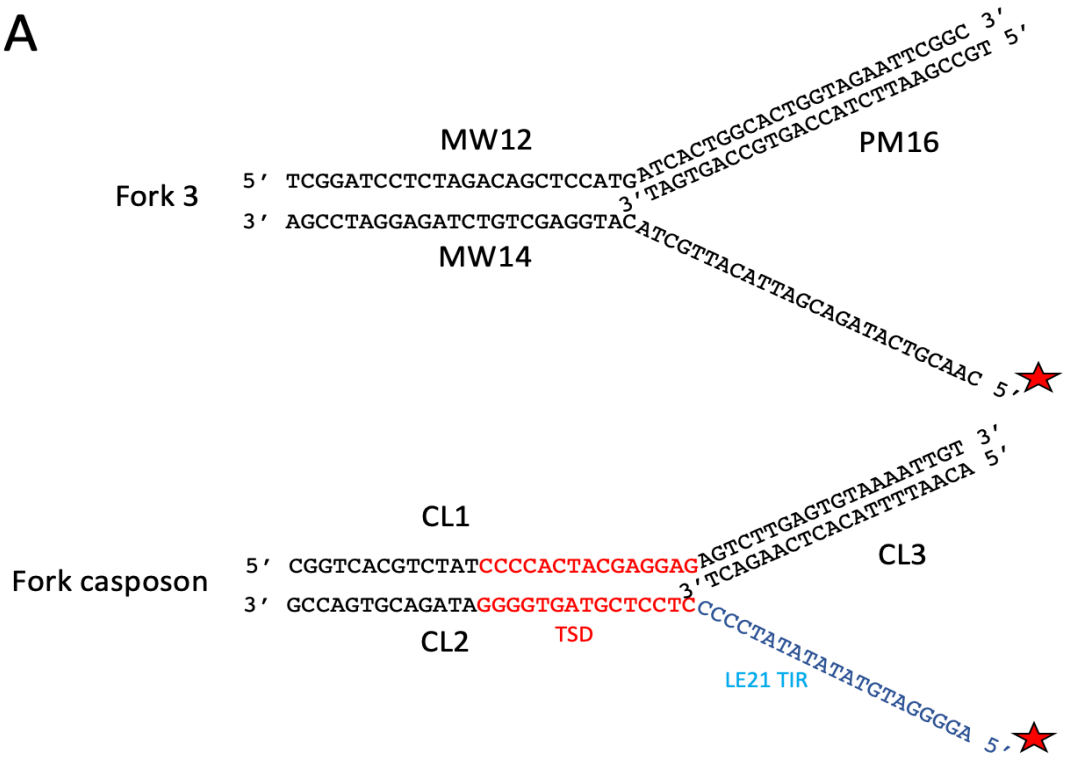
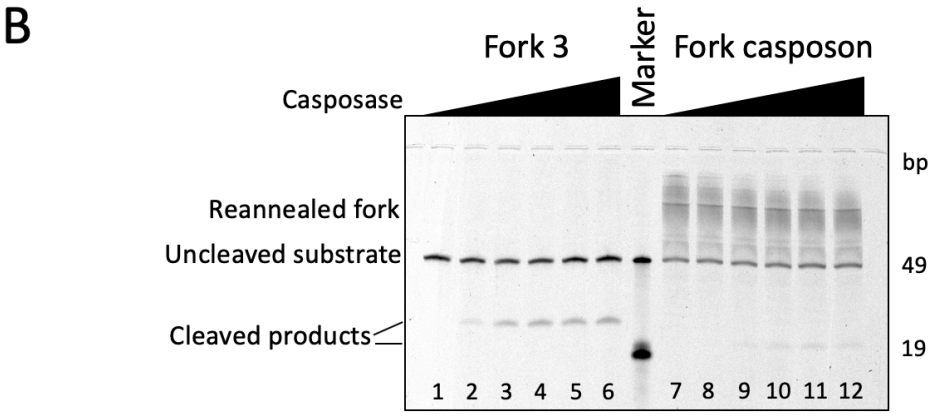


Figure 19. A. boonei casposase catalyses disintegration.
(A) Schematic diagram showing the disintegration reaction of the DNA substrate fork 3. (B) Analysis of divalent metal ion usage by *A. boonie* casposase. Casposase at concentrations of 0, 25, 50, 75, 150, 250 nM was incubated with 25 nM fork 3 substrate either in $MgCl_2$ buffer (10 mM Tris pH 7.5, 10% glycerol, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 50 μg/ml BSA) or $MnCl_2$ buffer (25 mM Tris pH 7.5, 9% glycerol, 106 mM KCl, 5 mM $MnCl_2$, 2 mM DTT, 50 μg/ml BSA) at 37°C for one hour. Disintegration activity was observed by the production of a shorter Cy5 labelled product. (C) Disintegration reaction carried out by 150 nM wild-type casposase and active site mutants casposase^{H242A} and casposase^{D254A} incubating with 25 nM fork 3 DNA substrate in integration buffer ($MnCl_2$ buffer in (B)).

Fork 3 comprising of random sequences was successfully disintegrated by casposase, this prompted us to compare the disintegration efficiency of fork 3 to a new DNA substrate, fork casposon containing sequences from *A. boonei* casposon (Figure 20A). The duplex region of the fork casposon contained a 15 bp target site duplication (TSD) found in *A. boonei* casposon and the single stranded region contained 21 nucleotides of the left end terminal inverted repeat (LE TIR) of native casposon. The result showed fork casposon was significantly poorer disintegrated by casposase compared with fork 3 (Figure 20B). The smearing above the uncleaved CL2 substrate from the gel was reannealed fork because there was no unlabelled CL2 added to compete with the Cy5 labelled CL2. Quantification of disintegration reactions illustrated that casposase activity reached a plateau at 75 nM of protein concentration (Figure 20C). It was observed that the maximum percentage of fork 3 disintegration was 30% whereas only 6% for fork casposon. This discrepancy might arise from the sequence specificity of casposase for nucleotides at the 3' end of the attacking strand and nucleotides at the target sites. Similar observations were made in CRISPR Cas1 catalysed disintegration (Rollie et al., 2015).


A
MW12
PM16
Fork 3
5′ TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC 3′
3′ AGCCTAGGAGATCTGTCGAGGTACATCGTTACATTAGCAGATACTGCAAC 5′
3′ TAGTGACCGTGACCATCTTAAGCCGT 5′
MW14
CL1
CL3
Fork casposon
5′ CGGTCACGTCTATCCCCACTACGAGGAGAGTCTTGAGTGTAAAATTGT 3′
3′ GCCAGTGCAGATAGGGGTGATGCTCCTCCCCCTATATATATGTAGGGGA 5′
3′ TCAGAACTCACATTTTAACA 5′
CL2
TSD
LE21 TIR


B
Fork 3
Marker
Fork casposon
Casposase
Reannealed fork
Uncleaved substrate
Cleaved products
bp
49
19
1 2 3 4 5 6 7 8 9 10 11 12

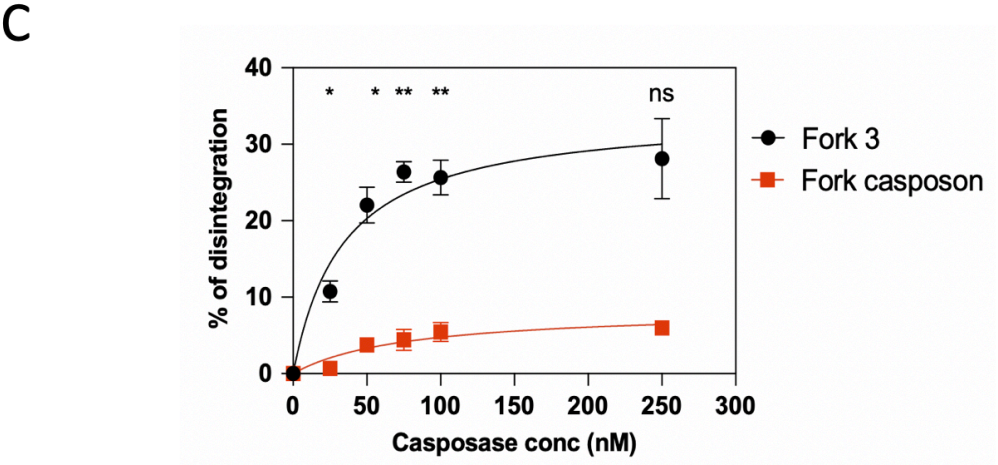

C
% of disintegration
Casposase conc (nM)
0
10
20
30
40
0 50 100 150 200 250 300
* * ** **
ns
Fork 3
Fork casposon

Figure 20. Disintegration of two different DNA fork substrates.
(A) Structure of fork 3 (top) and fork casposon (bottom). For fork casposon, the target site duplication (TSD) was highlighted in red and the 21 nucleotides left end terminal inverted repeat (LE TIR) was highlighted in blue. The red star represented the Cy5 end label. (B) Casposase at concentrations of 0, 25, 50, 75, 150, 250 nM was incubated with 25 nM fork 3 and fork casposon in integration buffer at 37°C for one hour. (C) Quantification of casposase disintegration of fork 3 to fork casposon from part (B). N=2. Error bars represent standard error of the mean. Two-way ANOVA was performed to determine statistical significance between fork 3 integration and fork casposon. *: p-value <0.05, **: p-value<0.01, ns: non-significant.

To explain the difference in disintegration of fork 3 and fork casposon, casposase bound to the two fork substrates were loaded on native polyacrylamide gels to detect whether there was a difference in binding affinity (Figure 21). However, no complex of casposase bound to DNA migrated into the gel for both substrates and the fork substrates were trapped in wells at high concentration of casposase. Therefore, it was unclear why fork casposon gave less product and it may be that the TSD and LE TIR sequence were in favour of integration rather than disintegration thus shifting the equilibrium.

Overall, *A. boonei* casposase showed better disintegration when using fork 3 as a substrate than fork casposon. As there is no known sequence motif on fork 3 for casposase binding, we next reasoned that casposase may be able to bind to random sequences to perform integration similar to CRISPR Cas1.

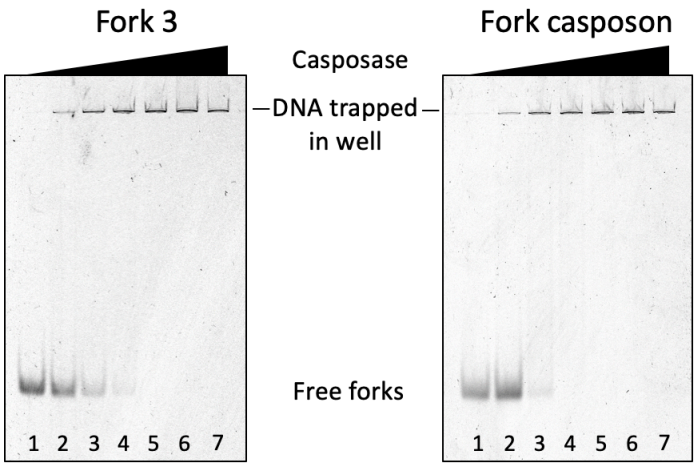


Figure 21. Electrophoretic mobility shift assay (EMSA) of casposase binding to fork 3 and fork casposon.
Casposase at concentrations of 0, 25, 50, 75, 100, 250, 500 nM was incubated with 25 nM fork 3 (left) and 25 nM fork casposon (right) respectively in binding buffer containing EDTA. After 20 min at 37°C, the reaction was loaded on 5% native PAGE gel and binding was observed by a band shift.

3.2.3. Casposase-catalysed short oligonucleotides integration

A. boonei casposase is a DNA integrase and it was shown integrating oligonucleotides into pUC19 (Hickman and Dyda, 2015). Integration of oligonucleotides generates a plasmid product that is nicked/relaxed from single-end integration, compared with supercoiled plasmid prior to integration, or linearised product from double-end integration at adjacent locations (Figure 22).

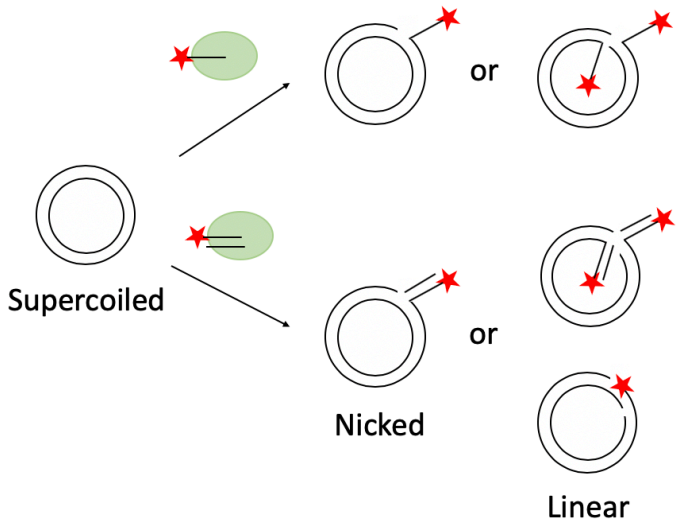


Figure 22. A cartoon shows casposase integrated ssDNA and dsDNA oligos into a plasmid leading to different conformations of the plasmid.
Green circle represented casposase and red star represented the Cy5 label at the 5' end of oligos.

Previous studies showed *A. boonei* casposase integrated both ssDNA and dsDNA derived from the left end terminal inverted repeat (LE TIR) of the casposon into a plasmid (Béguin et al., 2019; Hickman and Dyda, 2015). By combining this and disintegration results that showed casposase catalysed disintegration at random sequences, we used a 30-bp sequence derived from the LE TIR of the casposon (dsLE30) and its equivalent ssDNA (LE30 atk) as a 'standard' to compare casposase catalysed integration of dsDNA and ssDNA oligos with and without TIR sequences. Casposase but not the active site mutant catalysed insertion of most of these DNAs, with or without TIR, into the pACYC-Duet plasmid identified by the Cy5 label of each oligo being detectable on agarose gels at positions corresponding to nicked and linearised plasmid (Figure 23). The result showed that casposase can integrate single-stranded LE30 atk (Figure 23 lane 7) and this was consistent with previous study which used single-stranded LE26 (Béguin et al., 2019). However, the complementary strand to

LE30 atk, LE30 top was not integrated (Figure 23 lane 6). A 19-nucleotide (nt) PM32 was a poor substrate for integration compared with LE30 atk whereas a 28-nt TK24 and a 50-nt MW14 were integrated well (Figure 23 lanes 3, 4, 7, 9). Quantification of band intensity corresponding to Cy5 inserted plasmid showed two-fold increase in integration efficiency for TK24 and MW14 by compared with LE30 atk (Figure 24A). However, one-way ANOVA with multiple comparison to LE30 atk showed no statistical significance.

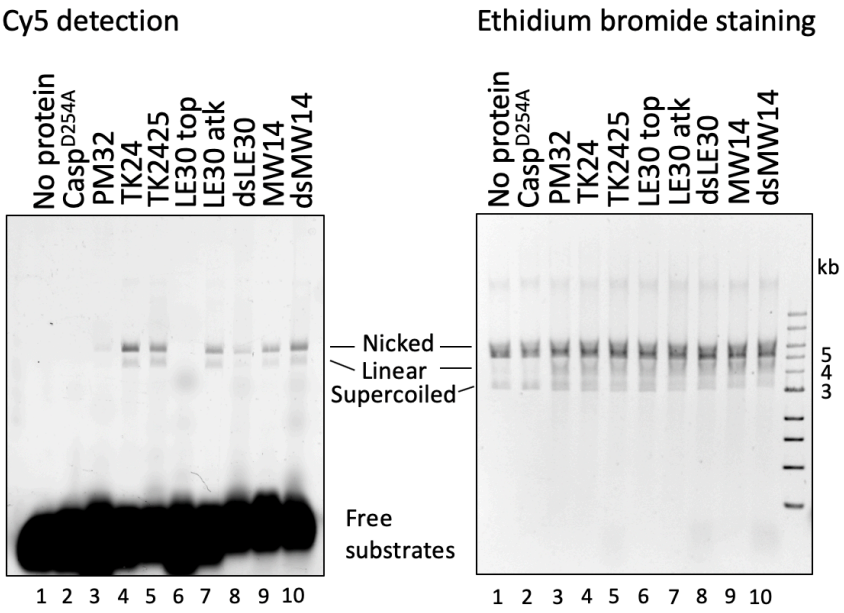
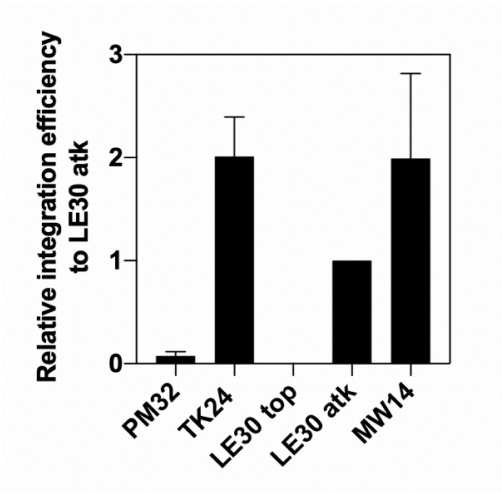


Figure 23. Casposase-catalysed ssDNA and dsDNA oligonucleotides integration into pACYC-Duet.
The reaction contained 150 nM casposase incubating with 100 nM of Cy5 labelled oligos and 150 ng of pACYC-Duet for one hour at 37°C. For the no protein control and casposase^{D254A} mutant, TK24 ssDNA was used. The gel was scanned to visualize integration of the fluorescent oligos (left) followed by staining with ethidium bromide (right).

The ssDNA oligos were aligned to see if there were any nucleotides playing a role in substrate recognition and integration by casposase (Figure 24B). As the nucleophilic attack occurred at the 3' end of the oligos, we reasoned that the important nucleotide

lay at the 3’ end. The alignment showed the terminal nucleotide at the 3’ end (-1 position) can be variable but nucleotides at -2 and -3 position from the 3’ end might involve in substrate recognition. The incompetent oligos PM32 and LE30 top both lacks a cytosine at -3 position and they both contained thymidine at the -2 position.

A



B

ssDNA substrates	sequence 5’ to 3’	-3-2
PM32	.....	TGTAATCGTCTATGACGTT
TK24	.....	GCAGTCCCCTCGCCTCAGCTACGCTCGT
LE30 top	.....	GGGGATATATATACATCCCCTCTTAAGTTC
LE30 atk	.....	GAACTTAAGAGGGGATGTATATATATCCCC
MW14	CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA	

Figure 24. Analysis of casposase-catalysed ssDNA oligos integration.
(A) Quantification of casposase catalysed ssDNA integration based on the gel shown in figure 23 and the band intensity was normalised to LE30 atk to obtain relative integration efficiency. Error bars represent standard error of the mean. N=3. (B) Sequence alignment of different ssDNA oligos. The red box highlighted the nucleotides at -2, -3 position from the 3’ end which may play a role in substrate binding.

In nature, ssDNAs are usually bound by single-stranded DNA binding protein (SSB), which protects ssDNAs from nuclease degradation (Pal and Levy, 2019). We tested if casposase can displace SSB from ssDNA oligo and perform integration (Figure 25).

TK24 pre-bound with SSB was incubated with casposase and pACYC-Duet. From the Cy5 scanning gel, SSB did not impede casposase ssDNA integration.

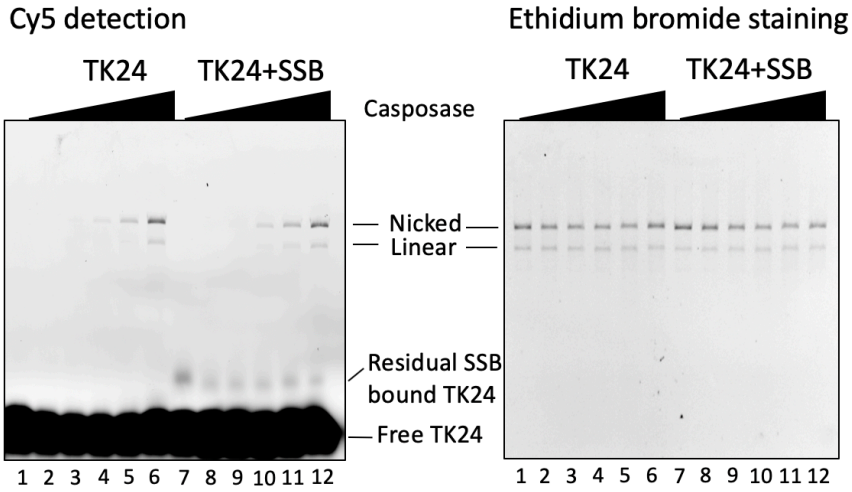
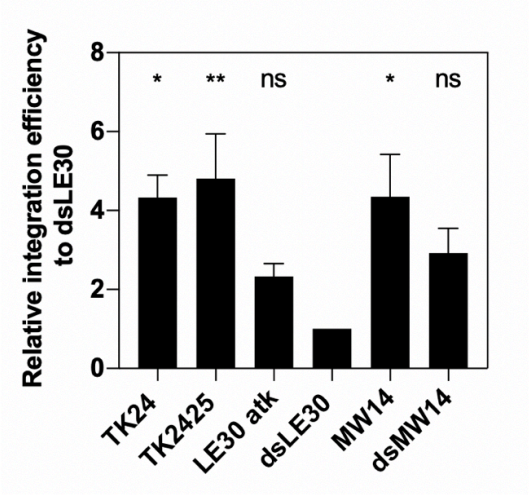


Figure 25. Casposase catalyses ssDNA integration in the presence of SSB.
Different concentrations of casposase at 0, 20, 40, 75, 150, 300 nM were incubated with either 100 nM of TK24 ssDNA (lanes 1-6) or 100 nM TK24 pre-bound with 500 nM SSB (lanes 7-12) and 150 ng pACYC-Duet for one hour at 37°C. SDS at 1% final concentration was added to the proteinase K stop buffer to release TK24 from SSB. The gel was scanned to visualize integration of the fluorescent TK24 (left) followed by staining with ethidium bromide (right).

For casposase-catalysed dsDNA oligos integration, dsLE30 was used as a 'standard' and its integration efficiency was compared with other dsDNA substrates (Figure 23 lanes 5, 8, 10). TK2425 (TK24 annealed to TK25) contained 5-nt 3' overhangs at both ends and its integration efficiency was significantly higher than dsLE30 with a four-fold increase (Figure 26A and 26B). Another dsDNA substrate, dsMW14 (MW14 annealed to EW3) was integrated more efficiently than dsLE30 showing a two-fold increase in efficiency, albeit with no statistical significance.

A



B

dsDNA oligos

TK2425
GCAGTCCCCTCGCCTCAGCTACGCTCGT 3′
3′ CGGGTCGTCAGGGGAGCGGAGTCGATGC

dsLE30
GAACTTAAGAGGGGATGTATATATATCCCC 3′
3′ CTTGAATTCTCCCCTACATATATATAGGGG

dsMW14
CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA 3′
3′ GTTGCAGTATCTGCTAATGTAACGATGTACCTCGACAGATCTCCTAGGCT

Figure 26. Analysis of casposase-catalysed ssDNA and dsDNA oligos integration.
(A) Quantification of ssDNA and dsDNA integration by casposase based on the gel shown in figure 23 and the band intensity was normalised to dsLE30 to obtain relative integration efficiency. Error bars represent standard error of the mean. N=4. One-way ANOVA with multiple comparison was performed to determine statistical significance. *: p-value <0.05, **: p-value<0.01, ns: non-significant. (B) Sequence of different dsDNA oligos.

When the integration efficiency of ssDNA was compared with equivalent dsDNA, the efficiency of blunt ended dsDNA oligo products was decreased, consistent with previous study (compare Figure 23 lanes 7 with 8, 9 to 10 and Figure 26) (Béguin et al., 2019). TK2425 was integrated as effectively as TK24 because the dsDNA substrate contains 5-nt 3' overhangs which is favoured by casposase (compare Figure 23 lanes 4 with 5 and Figure 26) (Hickman et al., 2020). The integration results of TK24 and TK2425 were consistent with anisotropy measurements that showed similar binding

affinities of casposase to ssDNA and dsDNA versions of TK2425 (K_d=49 nM, Figure 27) suggesting the 3' overhangs might help the protein binding to the substrate.

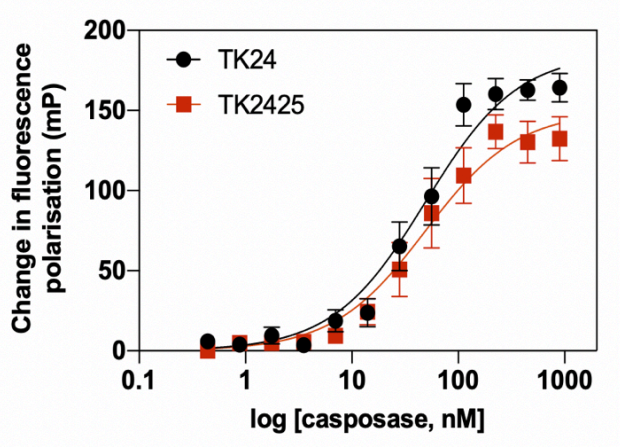


Figure 27. Fluorescence anisotropy data showing casposase bound to TK24 as effective as TK2425.
The dissociation constant K_d for both substrates was 49 nM. Error bars represent standard error of the mean. N=4.

3.2.4. Casposase-catalysed Long DNA integration

Previous studies showed *A. boonei* casposase integrates the kanamycin resistance gene into a plasmid when it is flanked by TIR sequences from the casposon (Béguin et al., 2016; Hickman and Dyda, 2015). Here, a similar assay was carried out to test if casposase could integrate a 1.2 kb ampicillin resistance gene (*amp*^{R}) encoding β-lactamase as part of a mini-casposon with TIR ends. From the short oligonucleotides integration assay, it was observed that casposase could catalyse integration of random sequences into a plasmid and it preferred duplex oligo containing 3' overhangs over blunt-ended duplex. Therefore, *amp*^{R} with 4-nt 3' overhangs and blunt-ended *amp*^{R} were also tested. Casposase at 150 nM was incubated with 150 ng pACYC-Duet and 45 ng of linear *amp*^{R} gene with different 3' ends (flanked by TIRs, blunt-ended without

TIRs, 4-nt 3′ overhangs) overnight. Product plasmids were ethanol precipitated and PCR amplified to detect integration product (Figure 28A and 28B). The result showed casposase could integrate all three long DNA substrates as determined by PCR analysis of the products. The PCR products identified on the agarose gel differed because the *amp*^{R} without TIRs (Figure 28B lanes 3 and 4) was the shorter version of mini-casposon (lane 2) and had 100 bp deletions at both ends.

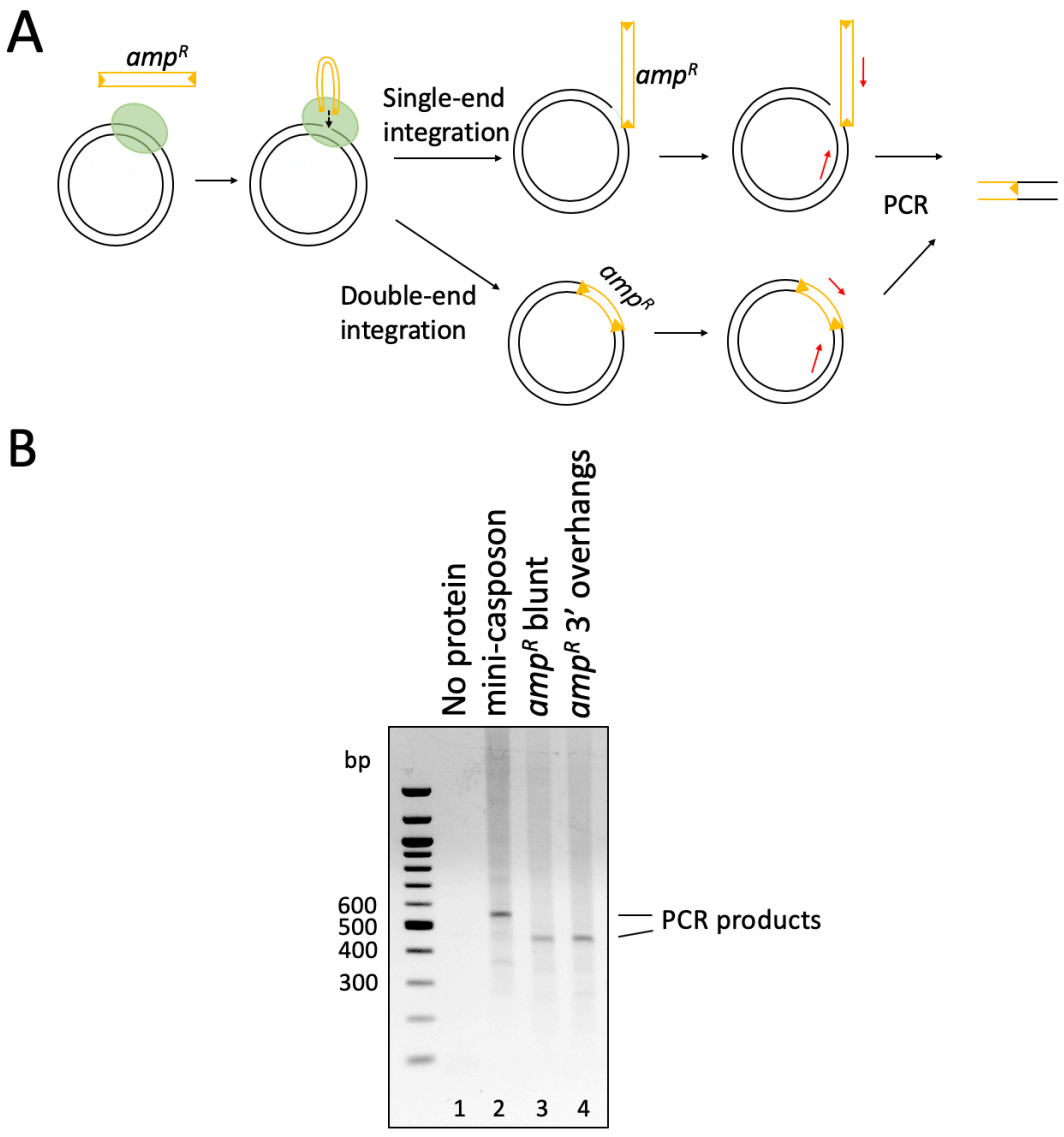
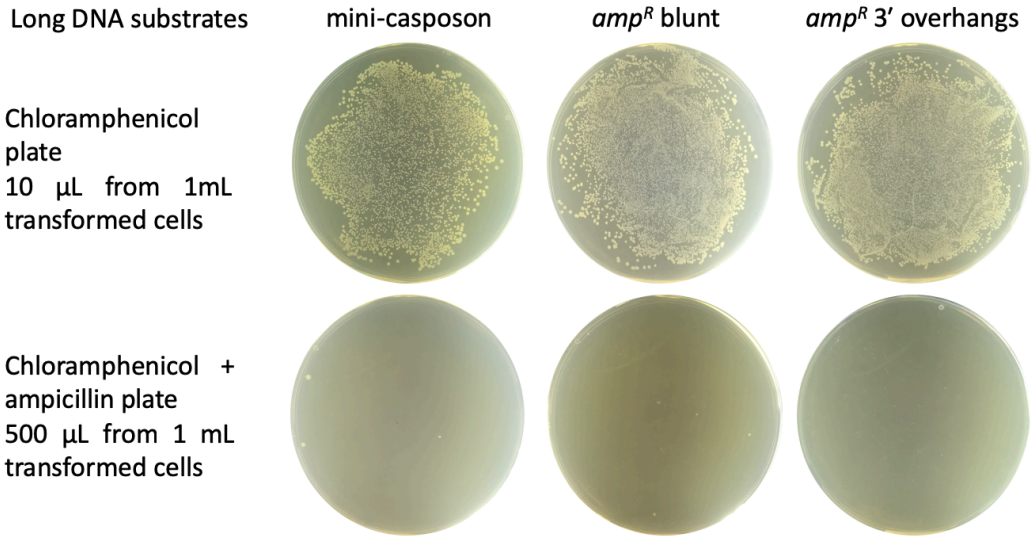


Figure 28. Casposase-catalysed Integration of linear ampicillin resistance gene into a plasmid.
(A) A cartoon showing the integration reaction and subsequent PCR reaction for detecting integration product. Green circle represented casposase and red arrows

represented PCR primers. (B) PCR reactions showing the linear DNA with or without TIR was integrated into the plasmid. The band size of *amp*^{R} no TIR (*amp*^{R} blunt and *amp*^{R} 3' overhangs) was smaller because it was the shorter version of *amp*^{R} with TIR (mini-casposon).

The PCR reaction in figure 28B amplified both single-end integration product and double-end integration product. As only double-end integration would result in a complete plasmid, plasmids purified from the integration were electroporated into *E. coli* DH5α to select for full-site integration (Figure 29A). Integration efficiency of *amp*^{R} gene was determined by the number of ampicillin and chloramphenicol resistant colonies divided by the number of chloramphenicol resistant colonies which assumed all plasmids were transformed (Figure 29B). Although PCR result showed *amp*^{R} with 3' overhangs was covalently attached to the plasmid, there was no ampicillin resistant colony observed after transforming into DH5α (Figure 29, Table 3.1 and Appendix 3). This suggested integration of *amp*^{R} with 3' overhangs was single-end integration. For the long DNA substrates mini-casposon and blunt-ended *amp*^{R} without TIRs, it was observed that mini-casposon with TIRs flanking was integrated five times more efficiently than *amp*^{R} without TIRs (Table 3.1 and Appendix 3). This suggested the TIRs at both ends might help casposase to capture the DNA substrates for integration.

A



B

Long DNA substrates	Colony number on cm plate for 10 µL of cells	Total transformed cells per mL	Colony number on cm and amp plate for 1 mL of cells	Integration efficiency
mini-casposon	3200	3.20×10^{5}	4	1.25×10^{-5}
***ampR* blunt**	6560	6.56×10^{5}	2	3.05×10^{-6}
***ampR* 3′ overhangs**	6912	6.91×10^{5}	0	0

Figure 29. Transformation of long DNA inserted plasmids into DH5α gives colonies. (A) Plasmids resulting from casposase-catalysed long DNA integration were electroporated into *E. coli* DH5α and transformed cells were spread on chloramphenicol plates to calculate total successfully transformed cells and on chloramphenicol and ampicillin plates to detect full-site integration products. (B) The colony number from (A) on each plate was used to calculate integration efficiency of each long DNA substrate. Cm: chloramphenicol. Amp: ampicillin.

Table 3.1. Integration efficiency of each long DNA substrate.

Integration efficiency / Substrates	Repeat 1	Repeat 2	Repeat 3	Average
mini-casposon	1.25×10^{-5}	7.30×10^{-6}	8.65×10^{-6}	9.48×10^{-6}
***ampR* blunt**	3.05×10^{-6}	0	2.31×10^{-6}	1.79×10^{-6}
***ampR* 3′ overhangs**	0	0	0	0

The plasmids pACYC-Duet with mini-casposon or *amp*^{R} blunt insertion were purified from ampicillin and chloramphenicol resistant colonies and were digested by a unique restriction cutter BamHI to check the plasmid size (Figure 30A). The gel illustrated that the plasmid size increased from 4 kb to around 5 kb and this increase in size matched to the size of the mini-casposon and *amp*^{R} blunt thus confirming the insertion of long DNA substrates. Next, the plasmids inserted with mini-casposon and *amp*^{R} blunt were sequenced from the integrated *amp*^{R} gene across both ends to determine the precise integration sites and target site duplications (Figure 30B-D). The sequencing data confirmed casposase did not require TIRs for integrating long DNA substrates but integrates the substrates into different locations. After the integration, a 15-bp target site was duplicated as observed in previous studies (Béguin et al., 2016; Hickman and Dyda, 2015). From 10 sequenced mini-casposon integrated pACYC-Duet (Appendix 4), it was observed that the mini-casposon was integrated into different locations (Figure 31A). Therefore, a DNA logo plot was generated to identify if there was a sequence motif for casposase target site recognition (Figure 31B). The plot revealed that the first nucleotide of the target site (nucleotide 16 in the DNA logo plot) and the second nucleotide upstream of the target site (nucleotide 14) were conserved that it can only be either a G or a T.

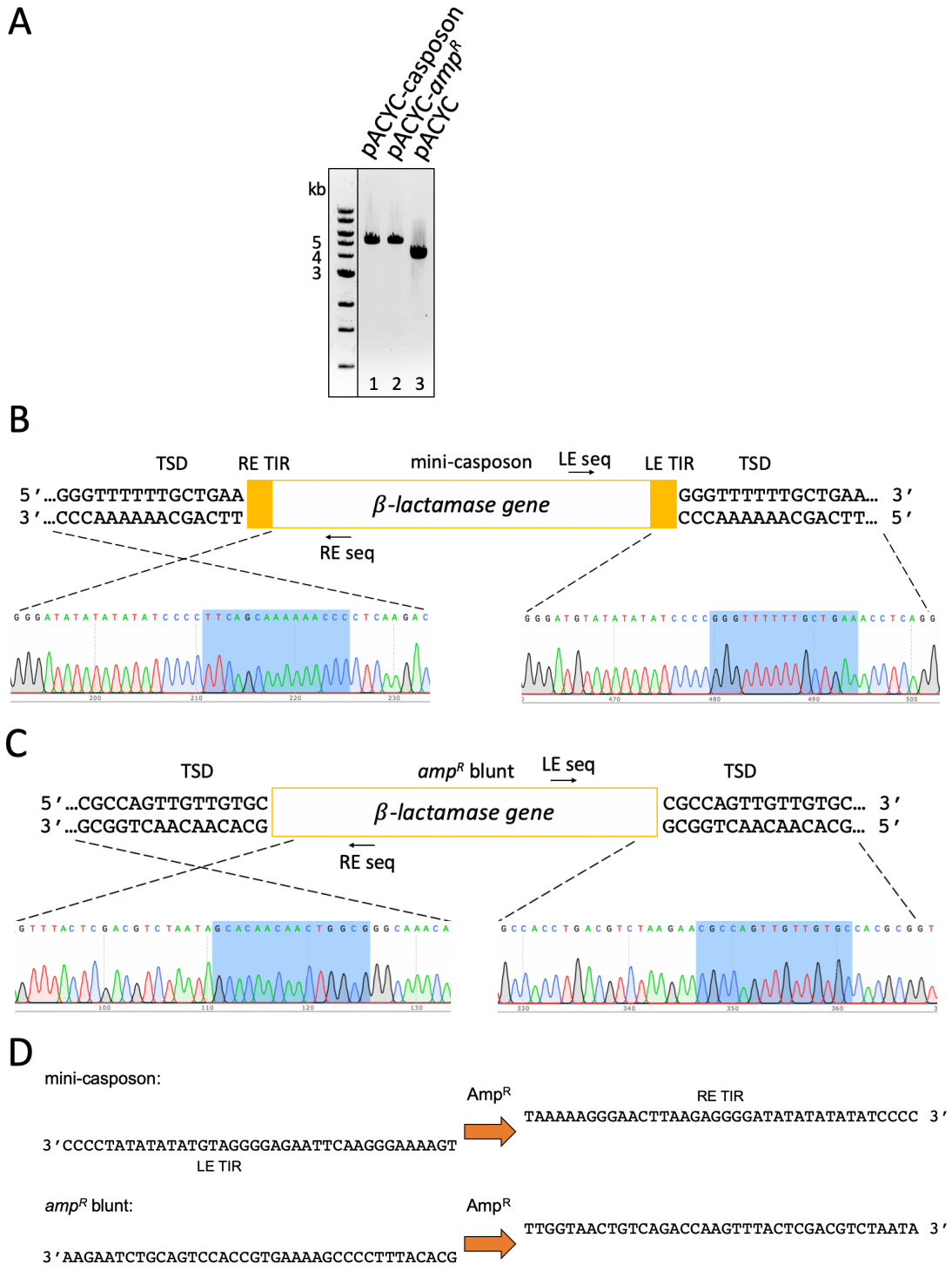


Figure 30. Sequencing results verified the insertion of ampR with or without TIR into pACYC-Duet after transformation and purification of the integrated plasmid.

After integration of mini-casposon and *amp*^{R} blunt into pACYC-Duet, the plasmids were ethanol precipitated and resuspended in dH_2O. The plasmids were transformed into *E. coli* DH5α and plasmids were purified from ampicillin resistant colonies. (A) The purified plasmids were cut by a unique cutter BamHI to check the size. The border at both ends of mini-casposon (B) and *amp*^{R} blunt (C) was sequenced across to determine the precise integration site and target site duplication (TSD, highlighted region). LE TIR: left end terminal inverted repeat. RE TIR: right end terminal inverted repeat. LE seq

and RE seq were sequencing primers used. (D) Comparison of 38 bp terminal sequence at both ends of mini-casposon with that of *amp*^{R} blunt. Only the 3′ nucleophilic attack strands are shown. (D) DNA logo plot of pre-integration site generated from 10 sequenced mini-casposon integration sites in pACYC-Duet. Nucleotides 1-15 were upstream sequence and nucleotides 16-30 were target site which would be duplicated after integration. Nucleotides 31-45 were downstream sequence. The sequence of individual site was listed in supplementary data 3.

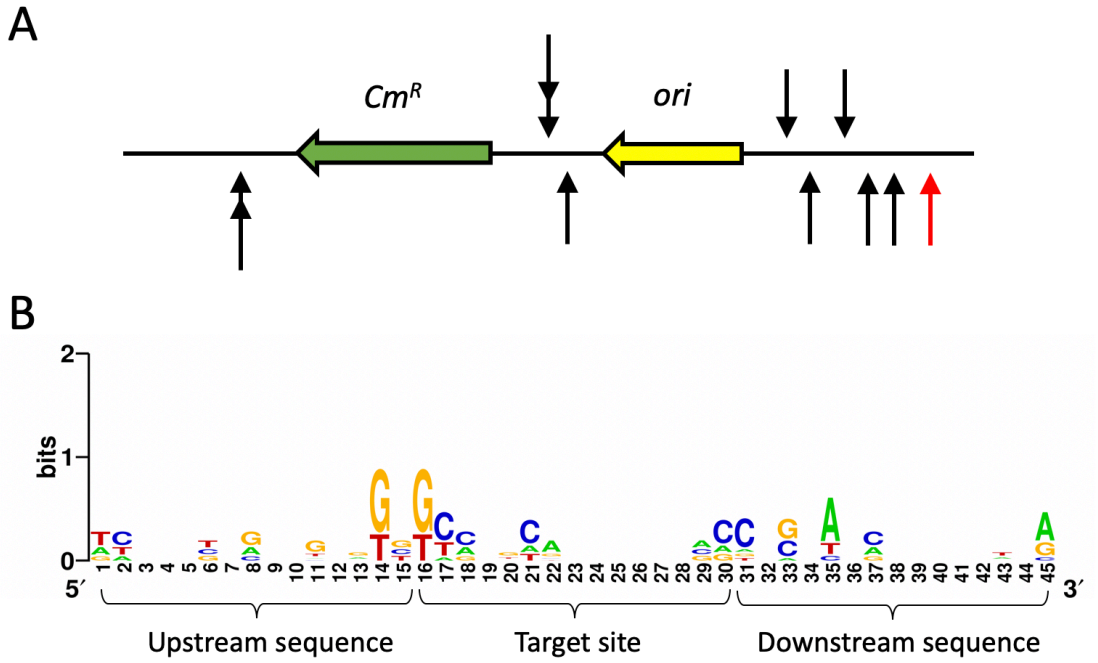


Figure 31. Amp^R was integrated into different locations of the plasmid by casposase. (A) Location of integration sites of mini-casposon (black arrows) and *amp*^{R} blunt without TIRs (red arrow) in pACYC-Duet. Ten mini-casposon integrated pACYC-Duet plasmids were sequenced and one for *amp*^{R} blunt. Direction of arrows indicates the direction of insertion of the 3′ -OH of the left end nucleophilic attack strand into the plus strand or the minus strand. (B) DNA logo plot of pre-integration site generated from 10 sequenced mini-casposon integration sites in pACYC-Duet. Nucleotides 1-15 were upstream sequence and nucleotides 16-30 were target site which would be duplicated after integration. Nucleotides 31-45 were downstream sequence.

3.2.5. Casposase cannot integrate casposon *in vivo* in *E. coli*

There has been no published *in vivo* study of casposase up to date. Comparative genomic analysis of 62 strains of *M. mazei* provided evidence of recent mobility of casposons thus showing some of these casposons are active (Krupovic et al., 2016).

However, there is no evidence showing whether casposon is mobilised via copy-and-paste or cut-and-paste mechanism. For cut-and-paste mechanism, casposase needs to excise casposon from one location first and insert it into another location. However, it was shown that casposase could not excise mini-casposon out of the plasmid both *in vitro* and *in vivo* (Figure 32A and 32B). As casposon encodes other proteins such as the uncharacterised endonuclease protein and family B polymerase, casposon transposition may require the endonuclease protein for cut-and-paste mechanism or require the family B polymerase for self-synthesising copy-and-paste mechanism.

To test if heterologously expressed casposase can catalyse *in vivo* DNA integration in *E. coli*, linear mini-casposon (amp^R gene flanked by TIRs) was electroporated into cells. EB377, MG1655 strain with an inducible T7 RNA polymerase, transformed with casposase plasmid pCHL2 was grown to mid-log phase (OD_{600}=0.6) and then the cells were induced to express casposase for three hours. Cells were then electroporated with 150 ng of mini-casposon but no colonies grew on ampicillin agar.

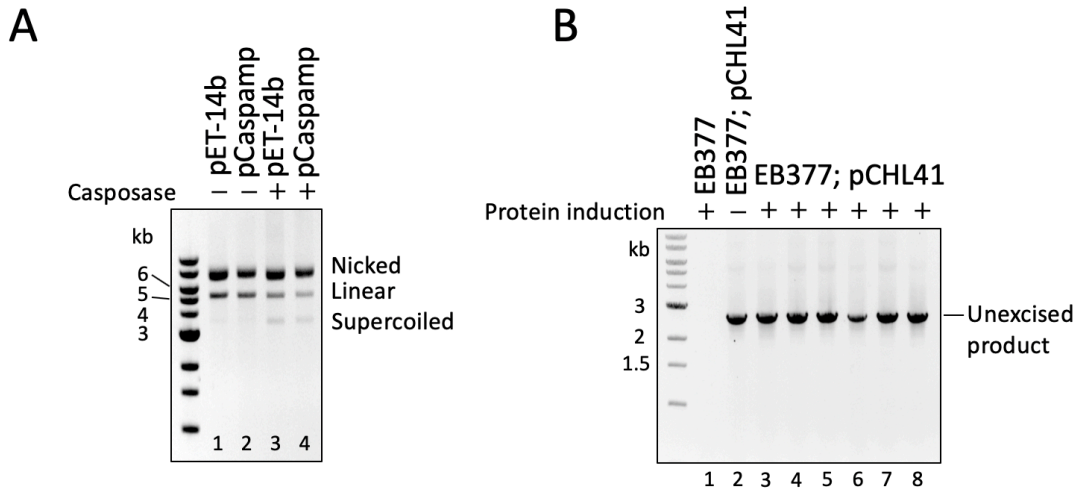


Figure 32. Casposase could not excise mini-casposon out of a vector.
(A) pCaspamp containing TIRs flanking the *amp^R* gene and its parental plasmid pET-14b were incubated with 150 nM casposase at 37°C for one hour. The plasmid was then run on agarose gel to detect release of mini-casposon at 1.2 kb. (B) *E. coli* EB377

strain was transformed with pCHL41 which contained casposase gene and mini-casposon in pACYC-Duet. The transformed cells were plated on chloramphenicol agar plate containing 0.5 mM IPTG and 0.1% L-arabinose for protein expression. After two days incubation, colony PCR was carried out to detect excision of mini-casposon. The 2.7 kb PCR band resulted from unexcised pCHL41. If mini-casposon was excised, a new PCR product at 1.4 kb would appear.

3.3. Discussion and conclusion

The adaptation complex of CRISPR-Cas systems builds the foundation of adaptive immune system as it generates a library of previously encountered invading MGEs. Cas1 protein is an important part of the adaptation complex as it catalyses spacer integration into a CRISPR locus (Nuñez et al., 2014). Phylogenetic studies suggested that CRISPR adaptation module has arisen from casposon and the casposon encoded Cas1 homologue, casposase may be ancestral to Cas1 proteins (Krupovic et al., 2014; Makarova et al., 2018). Indeed, it was shown that casposase catalyses substrate integration in a similar way to CRISPR spacer acquisition (Béguin et al., 2016, 2019). The casposase from family 2 casposons in *A. boonei* and *M. mazei* was shown to have a target site preference for the very 3' end of host tRNA gene. The 3' end of tRNA gene contains sequence motif for recruiting casposase and the first four nucleotides of the target site allows the casposase to 'see' the border and integrate substrates precisely. If the 3' end of tRNA gene and target site are absent, the substrates were integrated into random locations (Béguin et al., 2019; Hickman et al., 2020). These observations are consistent with CRISPR adaption in which a new spacer is predominantly integrated into the leader-proximal repeat. The leader sequence of a CRISPR locus contains sequence motif or is bound by a host protein to form a structure to recruit the adaptation complex (Kim et al., 2019b; Nuñez et al., 2016). Although structure analysis of casposase and Cas1 monomer showed similar domain organisation and

high degree of overlapping of active site residues (Figure 17B), the two proteins form different complexes. Casposase functions on its own by forming a homotetramer during integration (Hickman et al., 2020) and our results showed casposase can integrate DNA substrates with different lengths. On the other hand, a Cas2 dimer is required to bridge two Cas1 dimers for efficient spacer integration and the Cas2 dimer acts as a molecular ruler for precise spacer length (Nuñez et al., 2015c). This suggests during evolutionary transformation from casposase to CRISPR Cas1, Cas1 has acquired the ability to bind to Cas2 meanwhile lost the ability to bind to long DNA substrates. A recently characterised CRISPR type V-C Cas1 may provide an evolutionary intermediate step between casposase and Cas1 because most type V-C systems lack a *cas2* gene and their Cas1 forms a homotetramer for integration (Wright et al., 2019). The Cas1 phylogenetic analysis also showed type V-C Cas1 formed a branch rooted near casposase branch (Makarova et al., 2018). Experimental evidence illustrated that the V-C Cas1 could integrate spacers into the CRISPR locus *in vitro* but many off-target sites were seen (Wright et al., 2019).

Casposons are a new superfamily of DNA transposons which are flanked by TIRs at both ends (Krupovic et al., 2014). The N-terminal HTH domain of transposases specifically binds to an internal region of TIRs to initiate transposition (Munoz-Lopez and Garcia-Perez, 2010; Ramakrishnan et al., 2019). Unlike most transposases, the casposase encoded by *A. boonei* casposon does not strictly rely on TIR for integration and can integrate random sequences, especially for short oligonucleotides (Figure 23). This promiscuous substrate binding property has passed to Cas1 and laid the foundation of CRISPR adaptation. From the ssDNA integration, the results showed the -2 and -3 nucleotides from the 3’ end might play a role in casposase substrate

recognition (Figure 24B). This observation was consistent with a previous study that showed anything other than cytosine in the -3 position abolished substrate integration (Béguin et al., 2019). For short dsDNA integration, it was observed that a substrate with 3' overhangs was better incorporated into the plasmid (Figure 26) and this was also observed for the CRISPR Cas1-Cas2 adaptation complex (Nuñez et al., 2015a). For long DNA integration, casposase resembles transposases showing higher integration efficiency for DNA flanked by TIRs at both ends (Table 3.1). Because there was no native target site for casposase in the plasmid used in this study, casposase integrated substrate randomly into the plasmid. This level of target site flexibility may help the spreading of casposon during horizontal gene transfer when its preferred integration site is absent or mutated. Our results showed casposase was not very active in integrating mini-casposon into pACYC-Duet and this was due to the lack of casposase native target site. A recent study showed *A. boonei* casposase activity dropped to 10% when casposase integrated a TIR derived dsDNA oligonucleotide into a plasmid lacking a casposase target site (Béguin et al., 2019). From 10 mini-casposon integration sites, we identified two nucleotides at the upstream-target site border might play a role in target site recognition of casposase in the absence of native target site while there is no high sequence specificity at the target site-downstream border (Figure 31B). This is consistent with CRISPR adaptation in which the leader-repeat border dictates the first nucleophilic attack and the second nucleophilic attack is dictated by a molecular ruler mechanism (Goren et al., 2016). A previous study also generated a DNA logo blot from 20 mini-casposon integration sites in pUC19 by *A. boonei* casposase (Hickman and Dyda, 2015). This DNA logo plot showed no pattern of the nucleotide frequency at each position, whereas our DNA logo plot showed a G or T is frequently found at

the upstream-target site border. To better understand the sequence motif at the upstream-target site border recognised by casposase, a larger sample size will be needed to draw a more accurate conclusion.

Our results showed casposase alone was not able to excise mini-casposon from a vector. To understand how casposon is mobilised, other common proteins encoded by casposons should be studied. The question remains as to whether the family B polymerase encoded by casposons can replicate the casposons and what the roles of the casposon-associated endonuclease and putative DNA binding proteins are. To test whether casposase was active *in vivo*, integration experiments were performed *in vivo* by transforming linear mini-casposon into *E. coli*. Although electroporation was chosen as a high transformation efficacy method. The resulting lack of *in vivo* integration may well be due to rapid degradation of linear DNA inside the cell due to the RecBCD nuclease (Wiktor et al., 2018). Expression of a λ bacteriophage protein Gam which inhibits RecBCD complex may allow casposase *in vivo* integration to be seen (Wilkinson et al., 2016).

Chapter 4

4. Biochemical characterisation of casposase-Cas9 fusion proteins

4.1. Introduction

The programmable targeting capacity of the CRISPR-Cas9 system and its ease of use has made it a useful tool for genetic modification of genomes. Cas9 generates a site-specific DNA double-strand break (DSB) at the R-loop it forms during an interference reaction, initiating insertion of user-defined DNA sequence by homology-directed repair (HDR) catalysed by host cell's enzymes (Jiang et al., 2013; Jinek et al., 2012). Although targeted DNA insertion by CRISPR-Cas9 and HDR is achievable, this approach has several limitations. Firstly, HDR is at low efficacy and can sometimes trigger unpredictable genome rearrangement (Guirouilh-Barbat et al., 2014). Secondly, DSBs can be repaired by competing pathways other than HDR such as non-homologous end joining (NHEJ) and alternative end joining, resulting in various mutations at the site (Jiang et al., 2015b). Thirdly, in eukaryotic cells HDR enzymes is confined to S and G2 phases of the cell cycle meaning the approach is limited in non-dividing cells (Nami et al., 2018). Strategies have sought to overcome these HDR related limitations utilising small molecules to inhibit NHEJ pathway or upregulate HDR pathway (Chu et al., 2015; Cubbon et al., 2018; Song et al., 2016). Other strategies have been developed due to the fact that RNA-guided DNA binding capacity remains functional when Cas9 and its mutants are fused to other proteins. These include covalently tethering DNA repair template to Cas9 or fusing Cas9 to CtIP, a protein involved in early step of HDR (Aird et al., 2018; Charpentier et al., 2018).

Targeted gene insertion can be also achieved in a HDR independent manner by exploiting targeted integration of transposon (Feng et al., 2010). Recent studies reported a dCas9-transposase fusion protein and two natural systems in which Tn7-like transposons encode CRISPR-Cas proteins (Bhatt and Chalmers, 2019; Klompe et al., 2019; Strecker et al., 2019). The CRISPR ribonucleoprotein (RNP) complex forms a R-loop with target sequence then recruits Tn7 transposase to direct transposon integration at a fixed distance downstream of the R-loop. From chapter 3, we showed *A. boonei* casposase integrated a wide range of DNA substrates without TIRs and it was promiscuous in target site recognition in the absence of native target site. This makes casposase a potential DNA integration tool to by-pass HDR if the integration can be programmable. By fusing casposase to Cas9 protein, we reasoned that casposase integrates DNA substrates at a defined site near the R-loop formed by sgRNA-Cas9 RNP complex.

4.2. Results

4.2.1. Cloning of fusion protein genes

Cas9 protein is a versatile tool in protein engineering as different effector proteins fused to Cas9 or its mutants, nCas9 and dCas9, could bind to DNA in a sgRNA-guided manner. These proteins can be fused to Cas9 at the N-terminal or C-terminal (Bhatt and Chalmers, 2019; Guilinger et al., 2014; Komor et al., 2016). We therefore fused casposase to the N-terminal of *S. pyogenes* Cas9 (Figure 33). Three flexible linker sequences were tested: 18 amino acids long $(GGS)_6$, 24 amino acids long $(GGS)_8$ and the 16 amino acids SGSETPGTSESATPES sequence called ('XTEN', Guilinger et al., 2014).

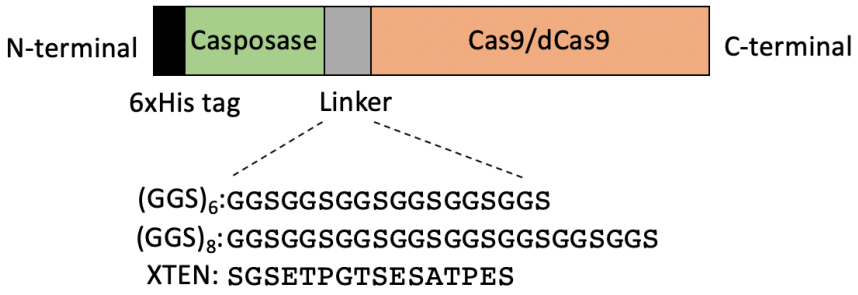


Figure 33. Schematic structure of casposase-Cas9/dCas9 fusion protein.
The sequence of three different linkers were shown.

To generate the fusion protein gene construct, we first utilised overlap extension PCR as described in 2.1.1 (Figure 34). In brief, the *casposase* gene and *cas9* gene were amplified separately with the *casposase* reverse primer containing 20 base pairs (bp) complementary sequence to the *cas9* forward primer. For the second round of PCR, the *casposase* and *cas9* PCR fragments were used as template. The two fragments were annealed by the complementary sequence during denaturing step in the first cycle and then the 3' end was extended by the polymerase to generate a fusion construct which can be amplified by the *casposase* forward primer and *cas9* reverse primer.

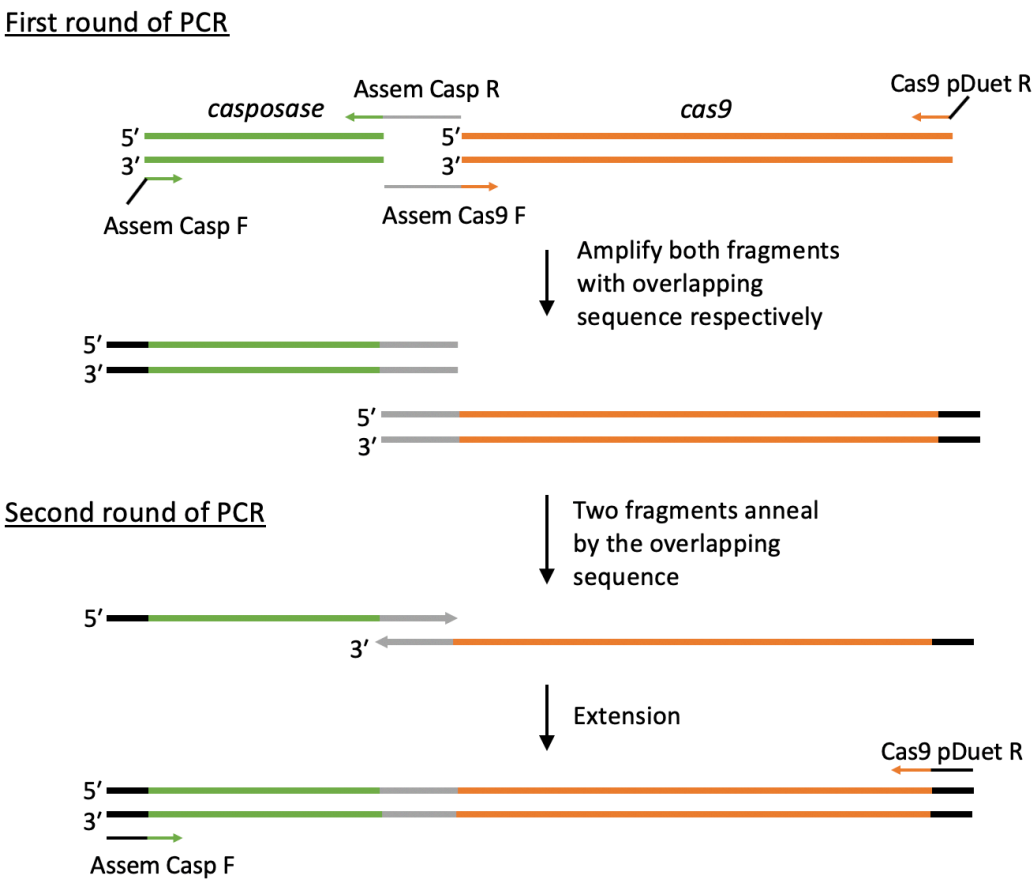


Figure 34. Schematic diagram showing generation of casposase-cas9 protein fusion DNA construct by overlap extension PCR.

Figure 35A showed amplification of the 1.2 kb *casposase* and 4.1 kb *cas9* gene. The correct gene fragments were gel purified and used as template in overlap extension PCR using *casposase* forward primer Assem Casp F and *cas9* reverse primer Cas9 pDuet R (Figure 35B). The fusion construct is 5.3 kb and a band around 5 kb was observed after analysis using agarose gel electrophoresis. The correct sized DNA fragment was gel purified and subject to another round of PCR to obtain more DNA for restriction digestion and subsequent molecular cloning steps. However, re-amplification of the fusion construct gave rise to smearing and no specific band was observed (Figure 35C). This suggested the 5 kb band from figure 19B might be non-specific product. The overlapping region between *casposase* and *cas9* fragments was

only 20 bp and this might be too short for effective annealing of the two fragments. The second round of PCR illustrated in figure 34 would be impeded by the dsDNA fragment formed by ssDNA reannealing to the fully complementary strand. To solve this, Gibson Assembly method could be used. It involves addition of a 5' to 3' exonuclease that creates 'sticky' ends at the overlapping region of the two double-stranded fragments, thus helping the annealing of the two gene fragments.

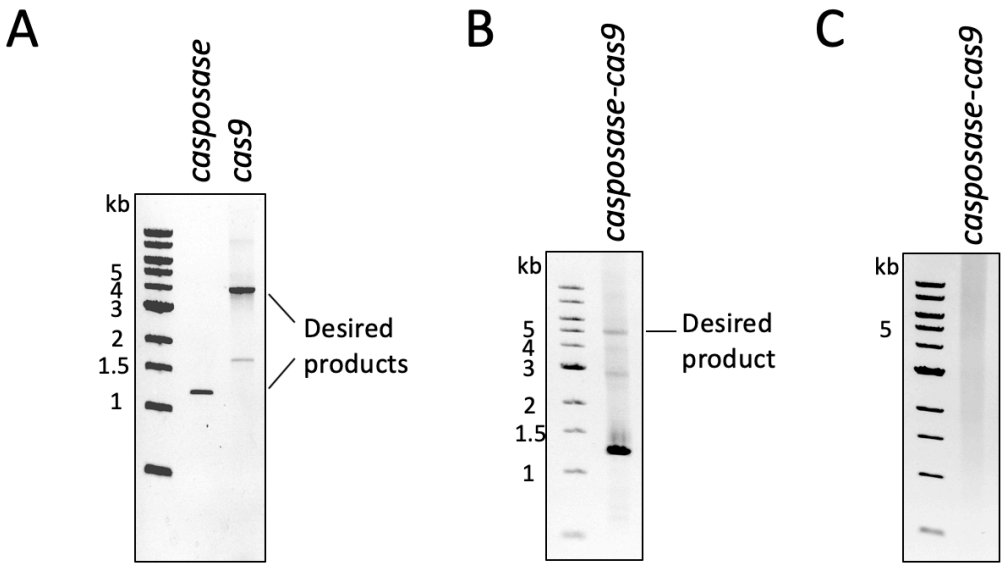


Figure 35. Construction of casposase-cas9 fusion by overlap extension PCR.
(A) Amplification of *casposase* and *cas9* genes. (B) Amplification of casposase-cas9 by overlap extension PCR using *casposase* and *cas9* PCR product as template. (C) Re-amplification of *casposase-cas9* from purified DNA from (B).

To construct fusion gene of *casposase* and *cas9*, we switched the strategy to multistep cloning as described in 2.1.5. The *casposase* gene was first inserted into pACYC-Duet to generate pCHL2 (Figure 36A). The *cas9* or *dcas9* gene was then inserted immediately downstream of the *casposase* gene (Figure 36B). The linker sequence was inserted between *casposase* and *cas9*/*dcas9* by Q5 site-directed mutagenesis PCR.

The plasmid at the end of each cloning step was sequenced to verify that genes were in-frame without mutation.

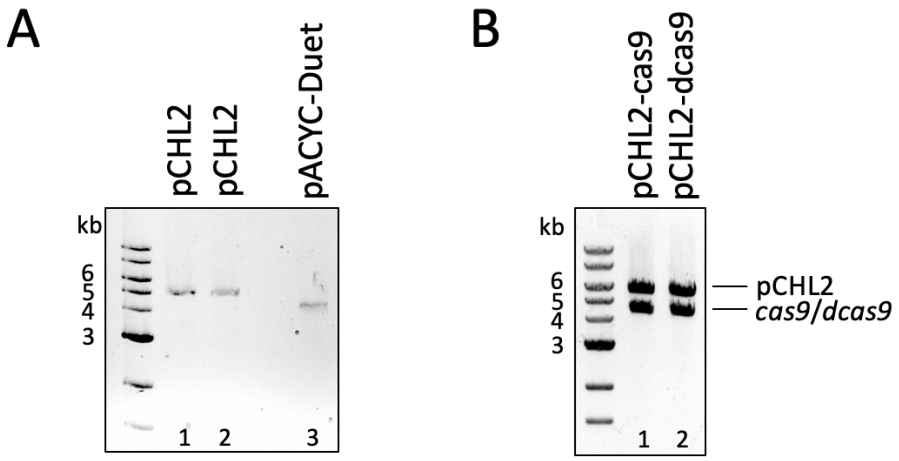


Figure 36. Insertion of casposase and cas9/dcas9 genes into pACYC-Duet in two cloning steps.
(A) pCHL2 (pACYC-Duet with *casposase* insertion) and pACYC-Duet were linearised by BamHI digestion. (B) pCHL2 with *cas9* or *dcas9* insertion were cut by PstI and NotI to release the *cas9* or *dcas9* fragment.

4.2.2. Fusion protein expression and purification

The casposase-Cas9 fusion protein comprises 1811 amino acids with a molecular weight (MW) estimated to be 210 kDa. In general, the proportion of soluble protein decreases dramatically when *E. coli* expresses a recombinant protein at MW above 60 kDa at 37°C (Susanne Gräslund et al., 2008). Protein mis-folding and precipitation result in protein expressed in insoluble form thus reducing the yield. To improve the yield of protein, protein induction should be carried out at lower temperature to achieve slower protein synthesis which allows more time for new translated proteins to fold properly (San-Miguel et al., 2013; Susanne Gräslund et al., 2008). Therefore, different induction protocols were tested to overexpress the fusion protein Casp-$(GGS)_8$-Cas9 (Figure 37). The pilot overexpression results showed auto-induction did

not work for the fusion protein while classical IPTG induction showed fusion protein expression. Overnight expression of Casp-$(GGS)_8$-Cas9 at 18°C by IPTG induction gave relatively good amount of protein so this induction protocol was chosen for expressing and purifying proteins.

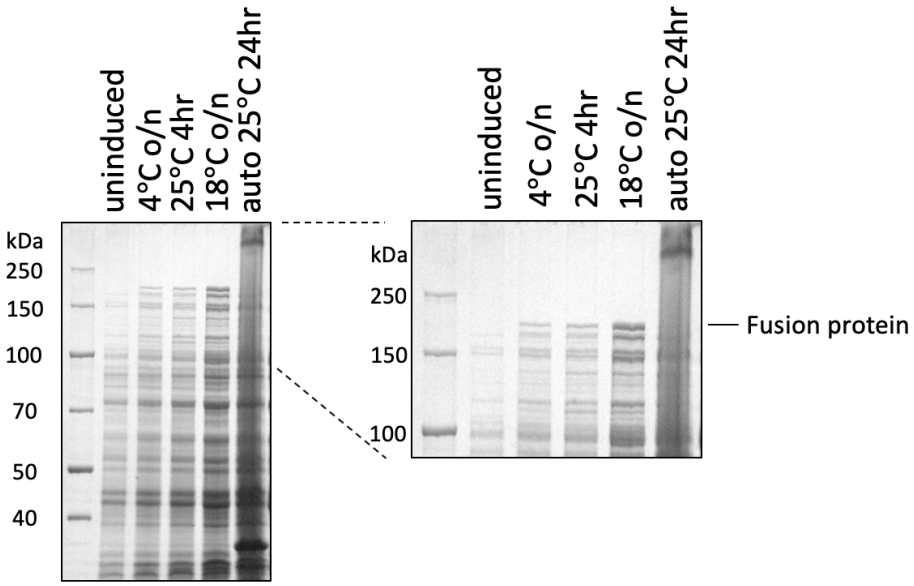


Figure 37. Optimisation of expression protocol of Casp-$(GGS)_8$-Cas9 fusion protein.
BL21-AI transformed with pCHL6 was cultured and induced protein expression using classical IPTG induction protocol as described in 2.2.1. After protein induction, the cultures were incubated at different temperatures for different lengths of time. The expression of Casp-$(GGS)_8$-Cas9 protein was auto-induced as described in 2.2.2. The culture was incubated at 37°C for eight hours in auto-induction medium and then incubated at 25°C for 24 hours. Fusion protein expression was analysed by running 6 µL of each induced sample on an 8% SDS gel. An enlarged view focusing on the 100-250 kDa region was shown on the right.

The Casp-Cas9/dCas9 fusion proteins were purified through Ni^{2+} column and heparin column as described in 2.2.3 (Figure 38A). The Cas9/dCas9 proteins with N-terminal maltose-binding protein (MBP) tag were first purified through Ni^{2+} column and heparin column. Then the MBP tag was removed by TEV protease and the de-tagged Cas9/dCas9 were purified through S-300 gel filtration column (Figure 38B). All purified and concentrated proteins were analysed by SDS-PAGE to check the protein purity (typified in figure 38C). Fusion proteins and Cas9/dCas9 showed obvious smaller contaminant bands while casposase did not. As these proteins were all purified the same way through Ni^{2+} column and heparin column and the larger contaminants above 46 kDa were absent in casposase, we believed the contaminant bands were protein degradation product arising from proteolysis (Wingfield, 2015).

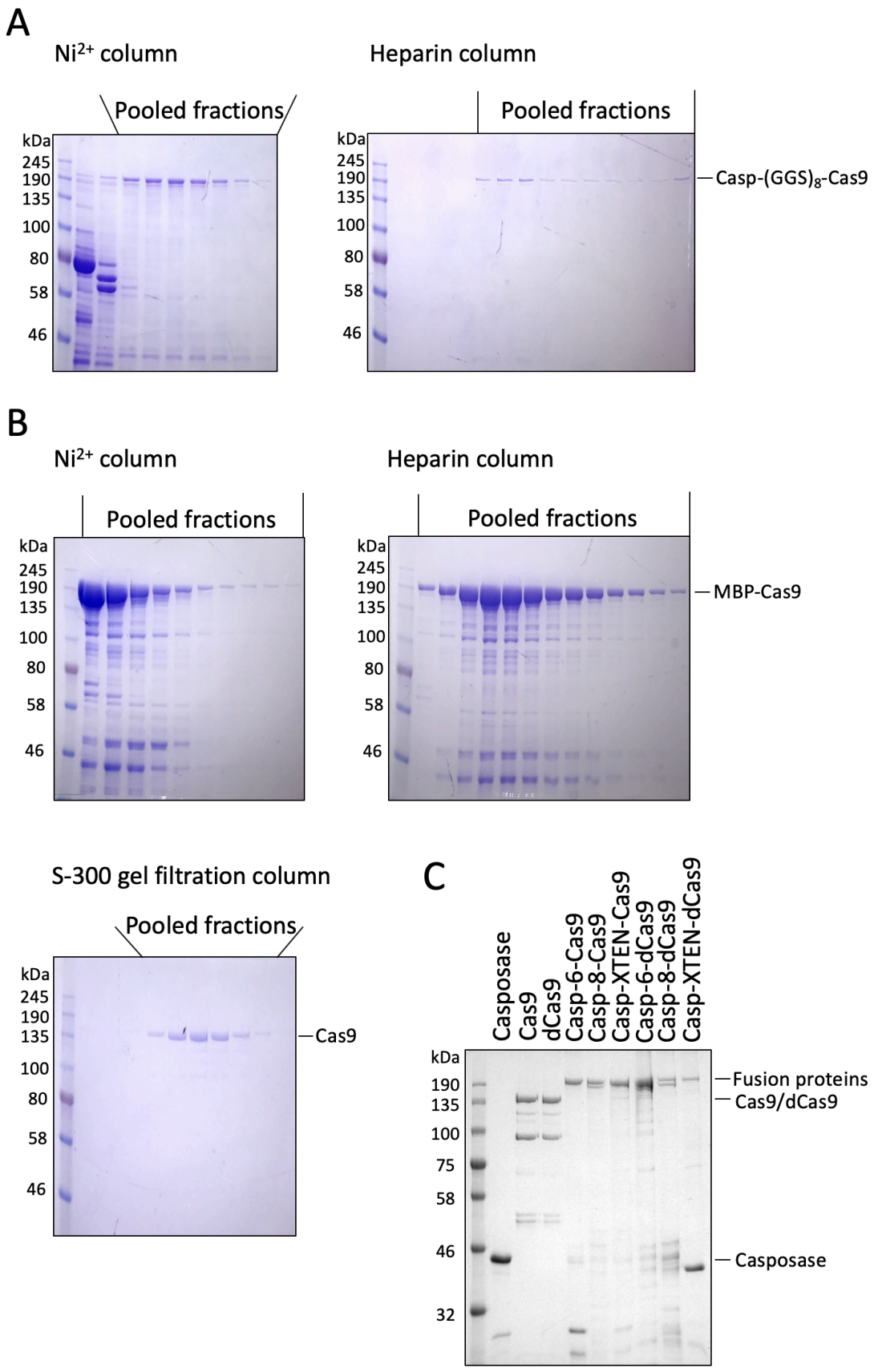

A
Ni2+ column
Heparin column
Pooled fractions
Pooled fractions
kDa
245
190
135
100
80
58
46
Casp-(GGS)8-Cas9
B
Ni2+ column
Heparin column
Pooled fractions
Pooled fractions
MBP-Cas9
S-300 gel filtration column
Pooled fractions
Cas9
C
Casposase
Cas9
dCas9
Casp-6-Cas9
Casp-8-Cas9
Casp-XTEN-Cas9
Casp-6-dCas9
Casp-8-dCas9
Casp-XTEN-dCas9
190
135
100
75
58
46
32
Fusion proteins
Cas9/dCas9
Casposase

Figure 38. Protein purification of fusion proteins and Cas9/dCas9.
(A) Purification of Casp-$(GGS)_8$-Cas9 through Ni^{2+} column and heparin column. Desired protein containing fractions were pooled and run through the next column or concentrated for storage. (B) Purification of Cas9. (Top) The MBP tagged Cas9 was purified Ni^{2+} column and heparin column. (Bottom) The de-tagged Cas9 was purified through S-300 gel filtration column. (C) SDS gel showing all purified proteins. The '6' and '8' between Casp and Cas9/dCas9 represents the $(GGS)_6$ and $(GGS)_8$ linkers.

4.2.3. Fusion proteins retain casposase activity

The functionality of Casp-Cas9/dCas9 fusions was first tested through disintegration catalysed by the casposase moiety, because disintegration activity was sensitive for detecting assays using 25 nM casposase compared with 75 nM casposase needed for integration (Figures 20 and 25). Disintegration activities of the purified fusion proteins with different linkers sequences were compared under the same conditions, described in section 3.2.2 (Figure 39). Reactions were carried out using excess amount of fusion proteins at 150 nM along with a positive control using casposase and negative controls using Cas9 and dCas9. While casposase cleaved fork 3 as expected, it was observed that apo-Cas9 cleaved the Cy5 labelled MW14 at multiple sites (Figure 39 lane 5). This was consistent with a previous study showing *S. pyogenes* Cas9 cleaves ssDNA in an RNA-independent manner (Sundaresan et al., 2018). dCas9 was inactive in DNA cleavage because the HNH and RuvC nuclease domains are inactive. All fusion proteins demonstrated disintegration activity proving that the casposase moiety in fusion proteins was functional. The low disintegration activity in Casp-$(GGS)_6$-Cas9 might be due to poor protein quality from purification (lane 7). Smaller cleavage products in Casp-$(GGS)_8$-Cas9 corresponded to Cas9 non-specific cleavage because of the same band pattern (lanes 5 and 8). However, Casp-XTEN-Cas9 and Casp-XTEN-dCas9 showed novel smaller cleaved products compared with casposase and Cas9 (lanes 4,

5, 10 and 11). This suggested the two proteins were contaminated by nuclease during purification. Casp-$(GGS)_8$-Cas9 and Casp-$(GGS)_8$-dCas9 contained the same linker sequence and exhibited similar disintegration activity comparable with casposase among all tested candidates (lanes 2, 3, 4 and 8). Comparison of disintegration activity on fork 3 over a range of protein concentrations further confirmed fully functional casposase activity from Casp-$(GGS)_8$-dCas9 (Figure 40A and 40B). Thus, all subsequent assays were done using fusion proteins containing the $(GGS)_8$ linker sequence and proteins were named Casp-Cas9 and Casp-dCas9. The protein sequences of the two fusion proteins were shown in Appendix 5.

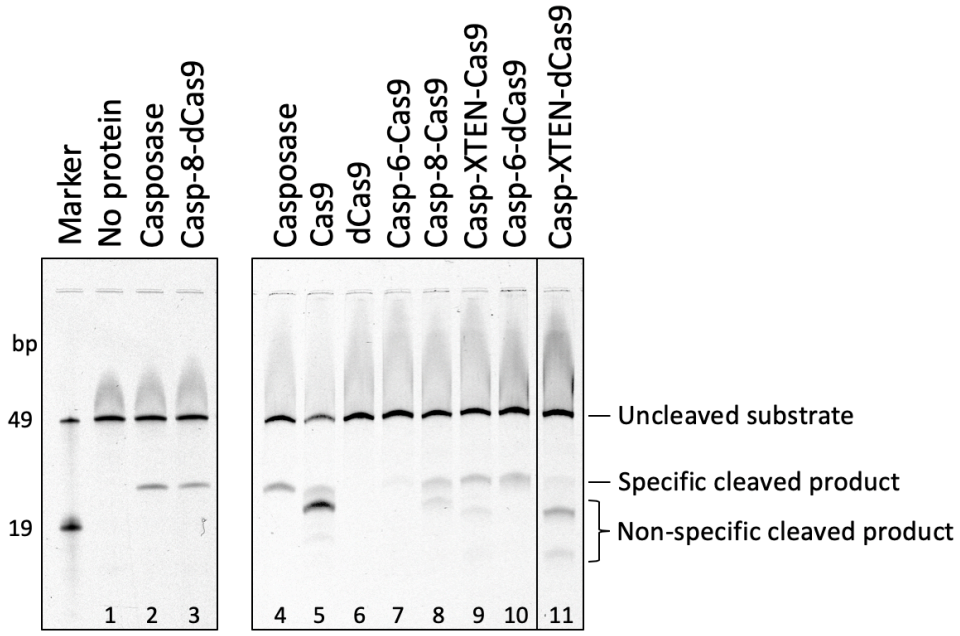


Figure 39. Fusion proteins-catalysed disintegration of a DNA fork substrate.
All proteins at 150 nM were incubated with 25 nM DNA fork 3 substrate at 37°C for one hour. The Casp-$(GGS)_8$-dCas9 reaction was run on a denaturing acrylamide gel (left, lanes 1-3) and the rest of fusion proteins were run on the other gel (right, lanes 4-11). An irrelevant lane was removed from the right gel, as indicated by a black line between lanes 10 and 11. The '6' and '8' between Casp and Cas9/dCas9 represents the $(GGS)_6$ and $(GGS)_8$ linkers.

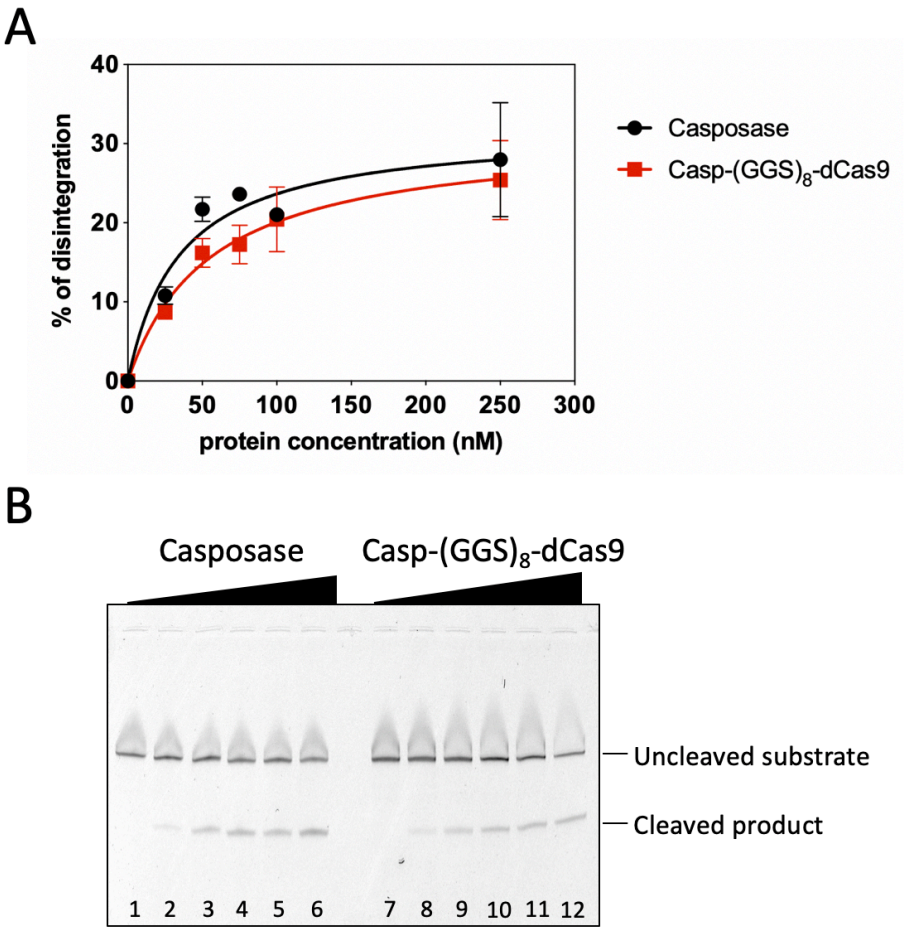


Figure 40. Casp-(GGS)$_8$-dCas9 catalysed disintegration as effective as casposase. (A) Quantification of disintegration activity of Casp-(GGS)$_8$-dCas9 compared with casposase. Proteins at 0, 25, 50, 75, 100, 250 nM were incubated with 25 nM fork 3 at 37°C for one hour. N=3. Error bars represents standard error of the mean. (B) A representative gel for quantification in (A).

4.2.4. Fusion proteins retain Cas9 activity

Cas9 functionality of fusion proteins was tested through R-loop formation by providing fusion proteins with sgRNA. A target DNA dsMW14 comprising MW14 and EW3 contains a PAM sequence 5'-TGG for the recognition of sgRNA-Cas9/dCas9 RNP complex (Figure 41A). R-loop formation between sgMW14 and dsMW14 by fusion proteins, Casp-Cas9 and Casp-dCas9, was readily detected after de-proteinising the reactions (Figure 41B lanes 10 and 12). The R-loops were confirmed to be genuine by treatment with RNase H, which degrades RNA in RNA-DNA hybrids (Figure 41B lanes

15 and 16). Measurement of R-loop formation over a range of protein concentrations identified that the efficiency of R-loop formation was similar for Cas9 compared with the fusion proteins (Figure 42). However, only 15% of total DNA substrate was accommodated into R-loop at the highest protein concentration; this is thought to be because of instability of R-loop structures after removal of proteins by proteinase K.

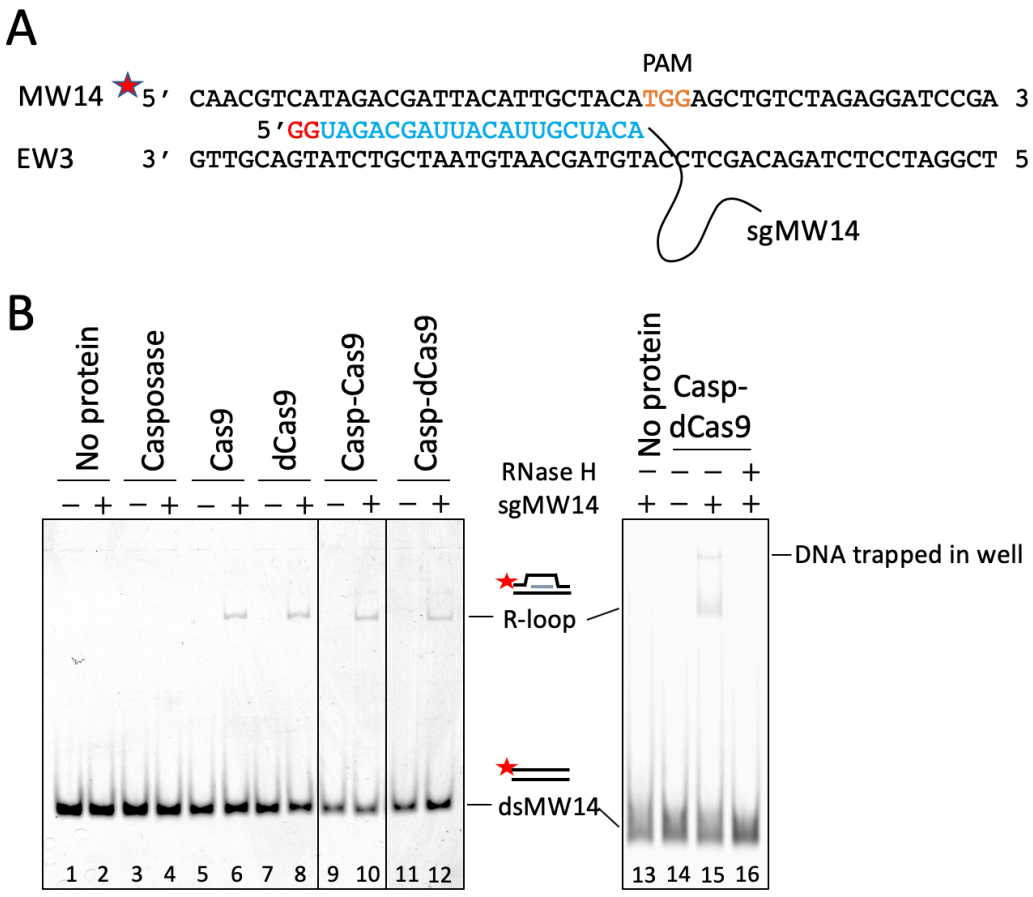


Figure 41. Fusion proteins formed R-loop with target DNA in the presence of sgRNA. (A) Schematic diagram showing the R-loop formed between dsMW14 and sgMW14 by Cas9 and fusion proteins. dsMW14 contains a PAM 5′-TGG labelled in orange. Nucleotides labelled in blue in the sgMW14 RNA are targeting sequence and two 5′ terminal guanines in red are by-product after *in vitro* transcription by T7 RNA polymerase. (B) A gel showing fusion proteins formed R-loop with dsMW14 in the presence of sgMW14. Proteins at 150 nM were first incubated with or without 450 nM sgMW14 and then incubated with 25 nM Cy5 labelled dsMW14 at 37°C for 20 min. Reactions were treated with proteinase K before running on gels. Irrelevant lanes were removed from gel on the left, as indicated by a black line between lanes. The gel on the right shows the R-loop formed by Casp-dCas9 was degraded by RNase H.

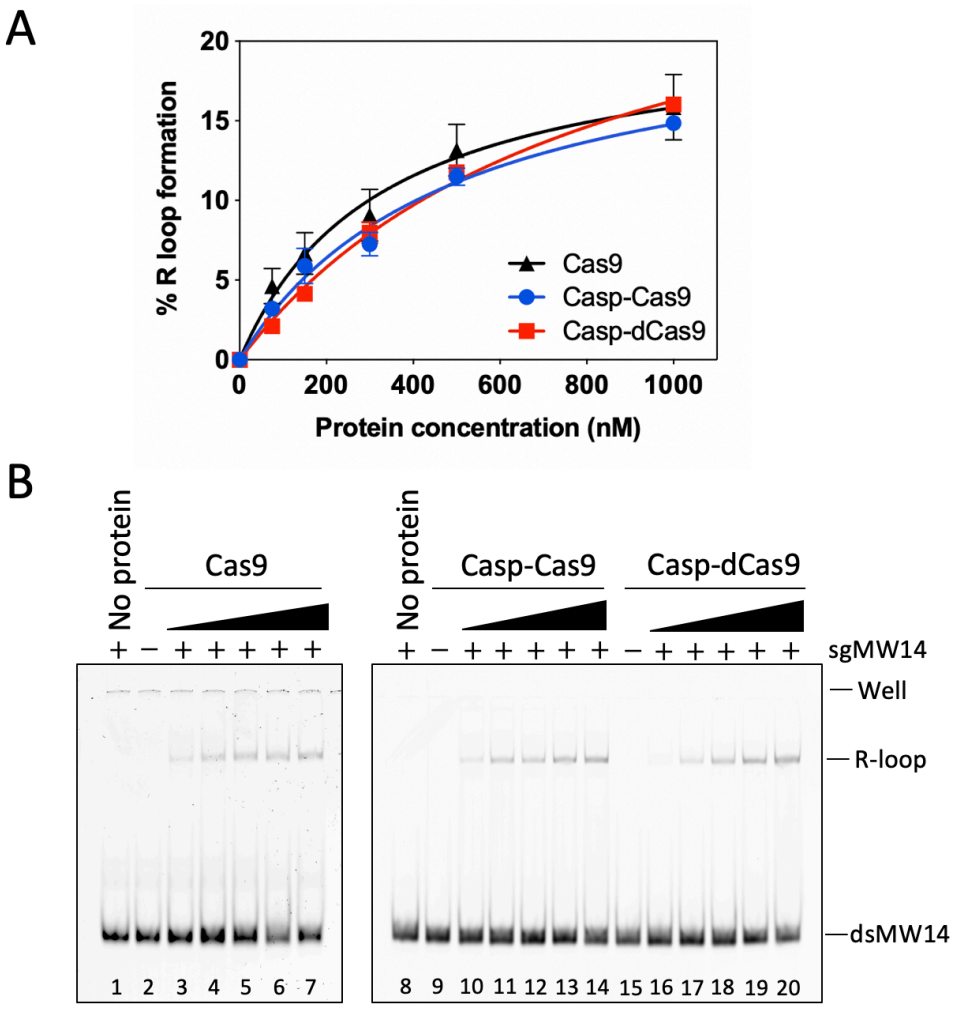
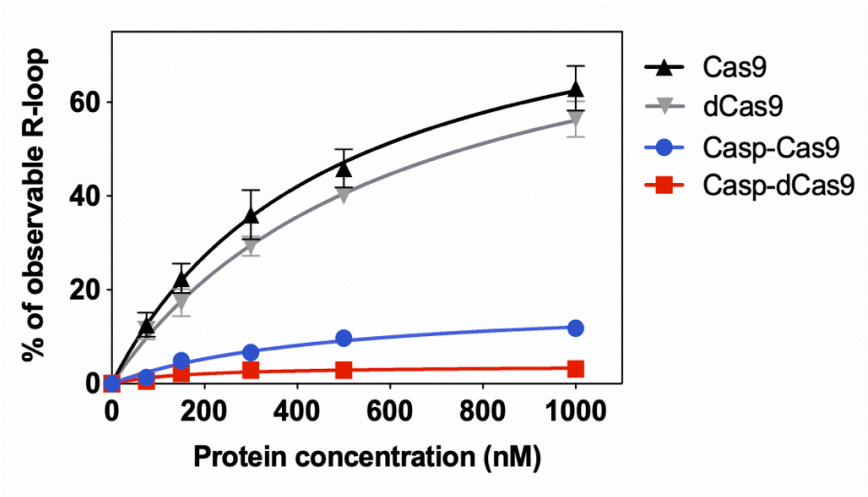


Figure 42. Fusion proteins formed R-loop as effective as Cas9.
(A) Quantification of R-loop formation by Casp-Cas9 and Casp-dCas9 compared with Cas9. Error bars represented standard error of the mean. N=3. (B) Representative gels showing R-loop formation, used in quantification in (A). Protein concentrations were at 75, 150, 300, 500 and 1000 nM. Protein concentration was at 300 nM in no sgRNA controls.

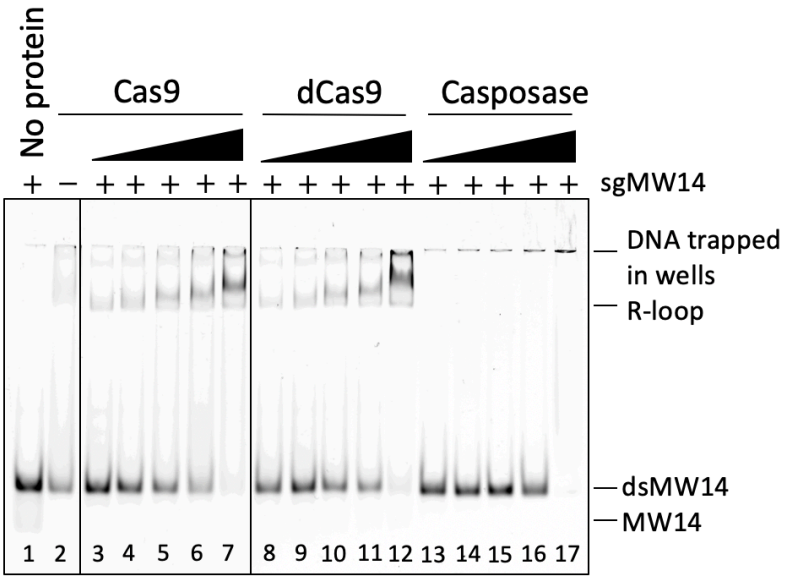
Next, we tested if fusion of casposase to Cas9/dCas9 might influence target DNA binding by the Cas9/dCas9 moiety. The R-loop formation assay was repeated but without proteinase K treatment. In contrast to similar R-loop formation efficiency for Cas9 and fusion proteins (Figure 42A), the migration of DNA-protein complexes into a polyacrylamide gel for Casp-Cas9 and Casp-dCas9 was significantly altered, compared with Cas9 and dCas9 (Figure 43). At 1000 nM protein concentration, only 3% and 11%

of target DNA formed stable R-loops with Casp-dCas9 and Casp-Cas9 RNPs respectively compared with more than 50% of stable R-loop formation in Cas9 and dCas9. The difference in DNA-protein complex formation between fusion proteins and Cas9/dCas9 might result from casposase binding to DNA at multiple sites leading to very large complexes observed as 'in well' aggregates (Figure 43B lanes 13-17, 20-24, 26-30). Because dsMW14 is small (50 bp), casposase binding to dsMW14 would destabilise the R-loop. The binding of casposase to dsMW14 might be transient because no DNA-casposase complex migrated into the gel at any protein concentrations (lanes 13-17).

A


% of observable R-loop
60
40
20
0
0
200
400
600
800
1000
Protein concentration (nM)
Cas9
dCas9
Casp-Cas9
Casp-dCas9

B


No protein
Cas9
dCas9
Casposase
sgMW14
DNA trapped in wells
R-loop
dsMW14
MW14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

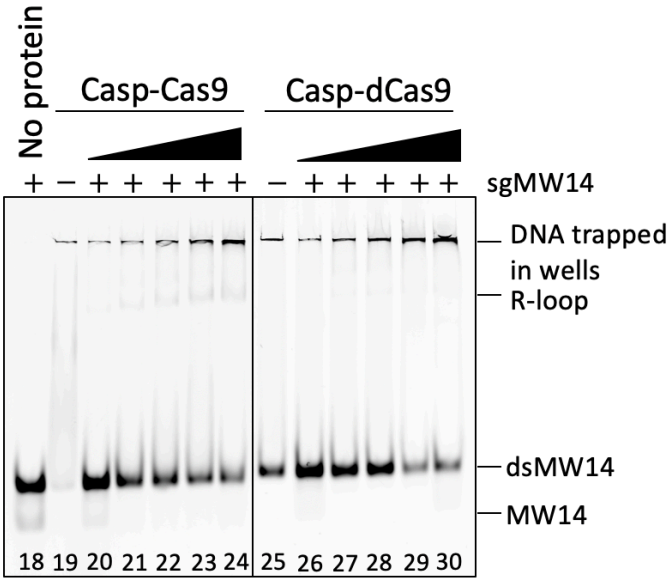

No protein
Casp-Cas9
Casp-dCas9
sgMW14
DNA trapped in wells
R-loop
dsMW14
MW14
18 19 20 21 22 23 24 25 26 27 28 29 30

Figure 43. DNA bound fusion protein complexes migrated differently compared with Cas9 and dCas9.
(A) Quantification of target DNA binding assay by Casp-Cas9 and Casp-dCas9 compared with Cas9 and dCas9. Error bars represented standard error of the mean. N=3. (B) Representative gels showing target DNA binding assay by different proteins, used in quantification in (A). Protein at 75, 150, 300, 500 and 1000 nM concentrations were incubated with three-fold molar excess of sgMW14 and then 25 nM Cy5 labelled dsMW14 was added at 37°C. Reactions were directly run on gels after 20 min incubation. Irrelevant lanes were removed from gel, as indicated by a black line between lanes. Protein concentration was at 300 nM in no sgRNA controls.

DNA target binding assays using the DNA oligo duplex dsMW14 might be interfered by non-specific binding of the casposase moiety in fusion proteins. We therefore tested R-loop formation using pUC19 plasmid as target DNA (Figure 44). Both sgRNA bound dCas9 and Casp-dCas9 were effective at forming R-loops in pUC19 DNA while casposase could not, as expected. We conclude that Cas9 and dCas9 were able to form R-loops with target DNA in the presence of sgRNA when fused to casposase.

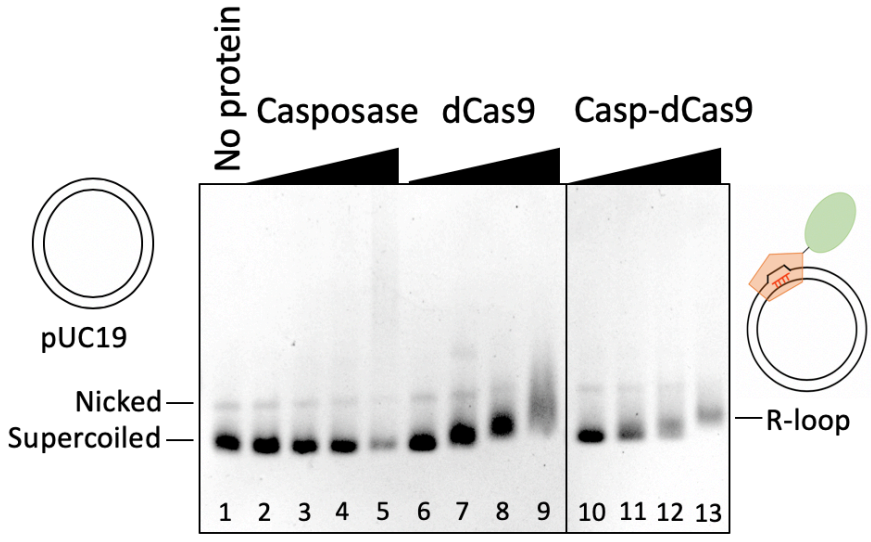


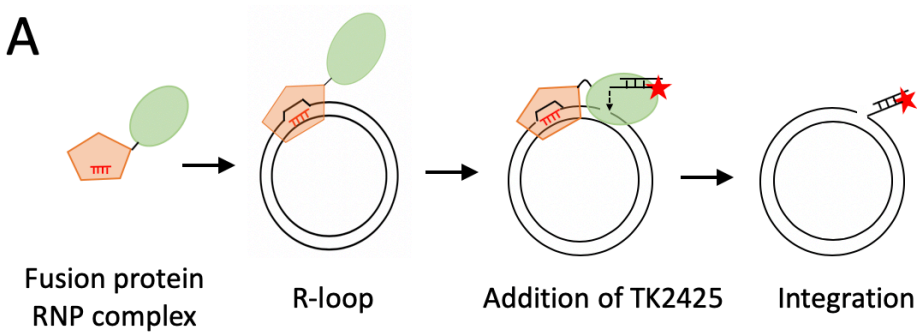
Figure 44. Casp-dCas9 and dCas9 formed R-loop with pUC19 plasmid in the presence of sgRNA.
Casposase, dCas9 and Casp-dCas9 at concentrations 40, 75, 150 and 300 nM were first incubated with three-fold molar excess of sgpUC19 RNA and then incubated with pUC19 at 37°C for 20 min. The dCas9 and casp-dCas9 RNP complexes formed R-loops

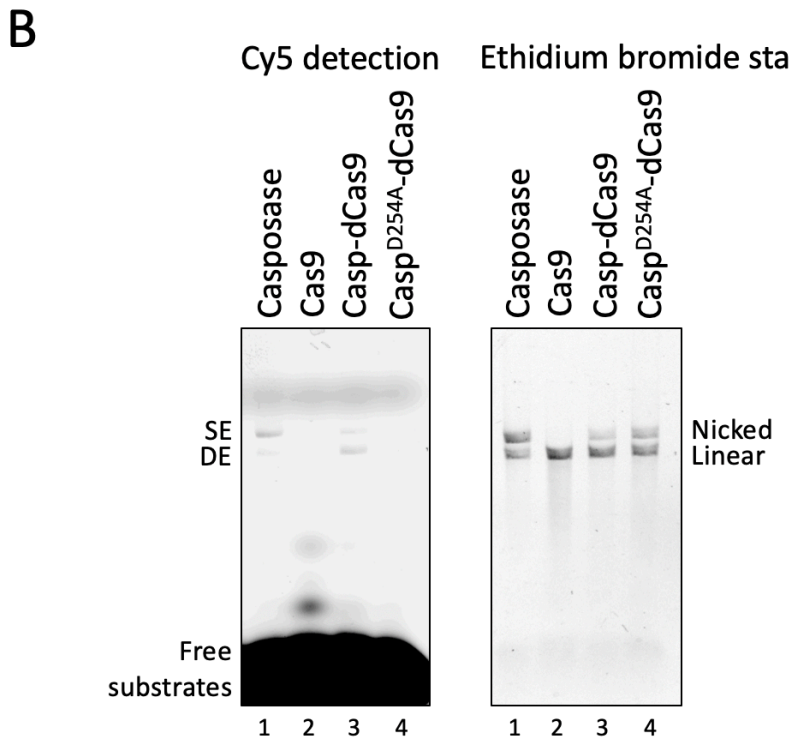
with pUC19 observable as plasmid up-shifts in ethidium bromide stained agarose gel. Irrelevant lanes were removed from gel, as indicated by a black line between lanes. R-loop formed by casp-dCas9 was shown as a cartoon on the right of the gel. The orange pentagon represents sgRNA-bound dCas9 and green sphere represents casposase.

4.2.5. *In vitro* sgRNA-guided short oligo integration

A. boonei casposase randomly integrates DNA into plasmids that lack its native target site, the 3' end of tRNA-Pro gene (Béguin et al., 2016; Hickman and Dyda, 2015). The above results demonstrated Casp-Cas9 and Casp-dCas9 fusion proteins contain functional casposase moiety and are able to form R-loop with target DNA in the presence of sgRNA (Figure 39 and 41B). Next the fusion proteins were tested if they can integrate DNA oligos guided by RNA when the R-loop is formed from sgRNA. The steps used to test this were shown in figure 45A. A sgRNA (sgpACYC) targeting to a plasmid pACYC-Duet was first incubated with fusion proteins to form RNP complexes followed by incubating with the target plasmid for specific binding and R-loop formation by the fusion proteins. DNA oligonucleotides TK24 and TK2425 were added for integration (Figure 45B and 45C). Specific binding and R-loop formation by fusion proteins was confirmed by ethidium bromide staining because only when Casp-Cas9 bound to sgRNA generated a DSB in the target plasmid and gave a linear product (Figure 45C lane 3). Fluorescence imaging to detect the Cy5 end-label of TK oligos showed fusion proteins integrated both TK24 ssDNA and TK2425 dsDNA into the plasmid forming single-end and double-end integration products (Figure 45B lane 3 and 45C lanes 2-5). DNA integration was catalysed by the casposase moiety in fusion proteins because an active site mutation D254A in the casposase moiety abolished DNA integration (Figure 45B lane 4). Fusion proteins catalysed TK2425 integration into the plasmid both in the presence and absence of sgRNA because Casp-Cas9 and Casp-

dCas9 retained casposase activity to integrate DNA into random sites. This integration activity into random sites is indistinguishable from sgRNA targeted integration by plasmid DNA electrophoresis.


A
Fusion protein
RNP complex
R-loop
Addition of TK2425
Integration


B
Cy5 detection
Ethidium bromide staining
Casposase
Cas9
Casp-dCas9
Casp^{D254A}-dCas9
SE
DE
Free
substrates
Nicked
Linear
1 2 3 4

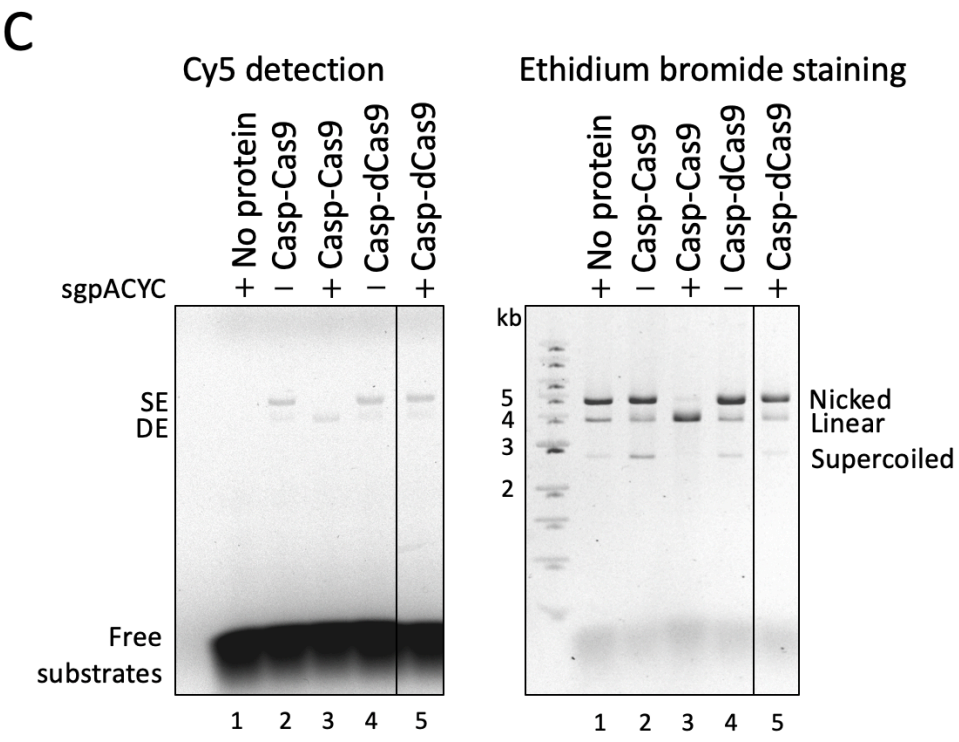

C
Cy5 detection
Ethidium bromide staining
No protein
Casp-Cas9
Casp-Cas9
Casp-dCas9
Casp-dCas9
+ − + − +
sgpACYC
SE
DE
Free
substrates
kb
5
4
3
2
Nicked
Linear
Supercoiled
1 2 3 4 5

Figure 45. Fusion proteins integrated DNA oligonucleotides into pACYC-Duet.
(A) Schematic diagram showing fusion protein catalysed targeted DNA integration. (B) Integration of 100 nM TK24 ssDNA into plasmid pACYC-Duet by 150 nM casposase and Casp-dCas9 at 37°C for one hour. Casp-dCas9 containing casposase active site mutation D254A could not perform integration. sgACYC was added in all samples. (C) Fusion proteins Casp-Cas9 and Casp-dCas9 at 150 nM were incubated with or without 450 nM sgpACYC RNA. A target plasmid pACYC-Duet was added for 20 min incubation and then 100 nM TK2425 was added for incubation at 37°C for one hour. Integration of dsTK24 into pACYC-Duet was assessed using ethidium bromide to indicate supercoiled, linear and nicked plasmids (right), and fluorescence to detect Cy5 labelled TK2425 that was single end (SE) or double end (DE) integrated (left).

To identify if sgRNA guided DNA integration for fusion proteins, plasmids purified from integration reactions were amplified across the targeting region using a primer annealing to pACYC-Duet close to the targeting region and a primer annealing to integrated substrate TK24 or TK25 (Figure 46A and 46B). PCR reactions using TK24 or TK25 gave the same band pattern and band intensity suggesting integration of TK2425 was not biased to either orientation. PCR reactions from DNA integration by fusion proteins without sgRNA gave a wide range of PCR products indicating random integration of the DNA substrate (Figure 46B lanes 3-4 and 7-8). The band intensity of a DNA band at about 400 bp was strong suggesting that it might correspond to a DNA integration hotspot recognised by casposase. Hereafter, this hotspot was referred to as ‘site A’. For reactions using Casp-Cas9 and sgpACYC, PCR products were confined to about 300 bp, consistent with expected product from targeted integration (lanes 5-6). This integration site corresponded to sgRNA targeting was referred as ‘site B’ hereafter. In reactions using Casp-dCas9, there was substantially more ‘site B’ PCR product in reactions containing sgpACYC than reactions without sgpACYC (lanes 7-10). However, Casp-dCas9 still randomly integrated DNA in the presence of sgpACYC depicted by a wide range of PCR products. The reason why PCR reactions of Casp-Cas9

with sgpACYC gave confined products is because the Cas9 moiety generated a DSB at the targeting region preventing extension after the break site in subsequent PCR.

DNA in bands corresponding to site A and site B integration was gel extracted and sequenced to determine exact sites of integration (Appendix 6). Sequencing results showed Casp-Cas9 and Casp-dCas9 bound to sgpACYC integrated DNA into site B located 26 bp downstream of the targeting region (Figure 46C). On the other hand, the DNA integration hotspot in the absence of sgpACYC, site A, was located 54 bp upstream of the targeting region.

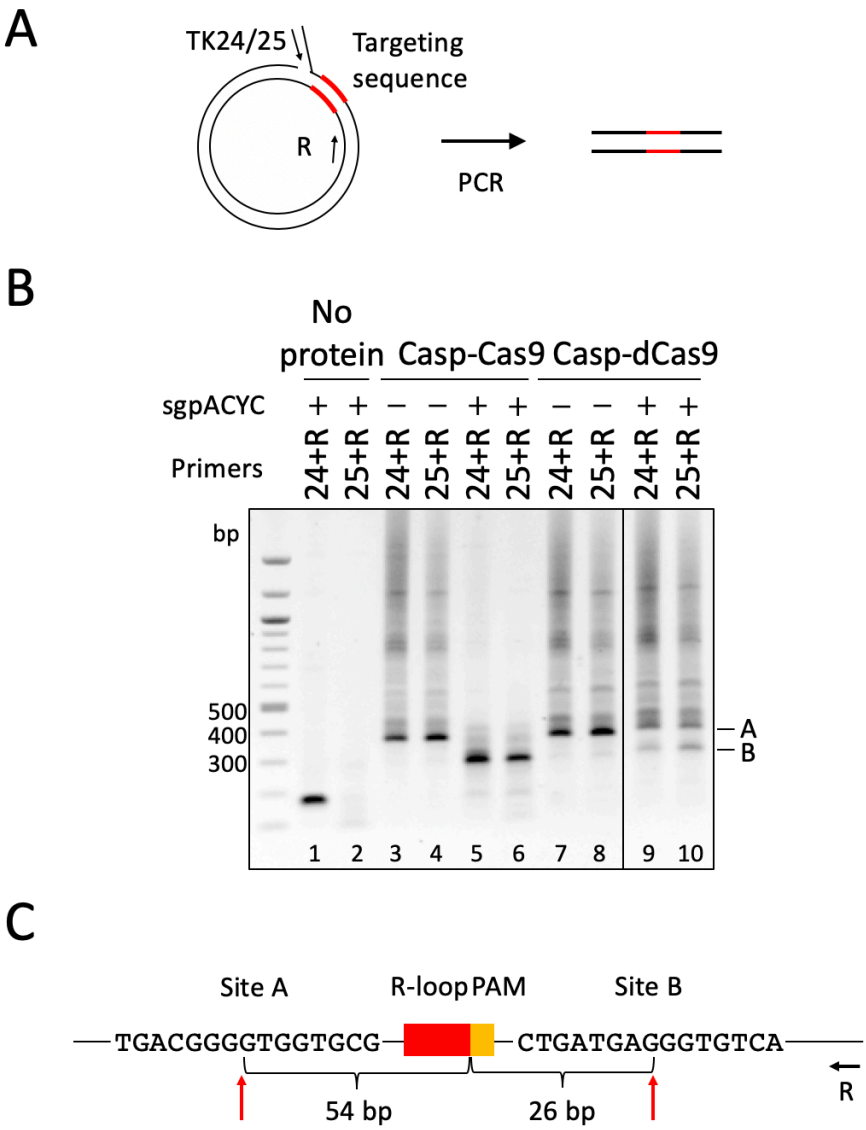


Figure 46. sgRNA-guided DNA integration by fusion proteins into pACYC-Duet.
(A) A schematic diagram showing PCR reactions amplifying integrated plasmid for detection of targeted integration. PCR reactions used a primer TK24 or TK25 to anneal integrated DNA, and a primer R to anneal pACYC-Duet. (B) Integrated plasmid products were PCR amplified to determine integration sites of TK2425. DNA bands, referred to as A and B, were gel extracted and sequenced (data shown in Appendix 6). (C) A schematic diagram showing TK2425 integration sites close to the targeting region of sgRNA in pACYC-Duet. Red arrows indicate the insertion of TK24 or TK25 between the two bases. R represents the PCR primer p15A ori R or sequencing primer SPIN seq.

To ensure integration into site B was guided by sgRNA rather than cryptic sequence preference of casposase, DNA sequence flanking site B was mutated. Site A was also mutated to act as a control to see if mutation can abolish DNA integration hotspot. pAB plasmid was synthesised comprising 250 bp DNA sequence from pACYC-Duet containing sgpACYC targeting site, site A and site B. Another plasmid pABmut was also synthesised being identical to pAB except for randomly mutated 15 bp DNA sequences flanking sites A and B. pAB and pABmut were used as target plasmids to compare integration of TK2425 by Casp-Cas9 in the presence and absence of sgpACYC (Figure 47A). Results of reactions using pAB were consistent with reactions using pACYC-Duet. TK2425 DNA substrate was largely integrated into the DNA hotspot site A by Casp-Cas9 in the absence of sgpACYC but integration site of TK2425 shifted to site B in the presence of sgpACYC (Figure 47A lanes 1-4). In reactions using pABmut plasmid, integration into mutated site A was barely seen in the absence of sgpACYC and was totally lost in the presence of sgpACYC (lanes 5-8). However, TK2425 was still integrated into mutated site B, indicated as B' on the gel, in the presence of sgpACYC. Sequencing of the integration product into mutated site B confirmed DNA inserted 26 bp downstream of the targeting region, the same distance as insertion into site B in pACYC-Duet (Figure 47B and Appendix 7). These results suggested the site A in pACYC-Duet was a casposase preference site and DNA integration into site B was governed by sgRNA targeting. Therefore, we conclude that casposase-catalysed DNA integration can be guided by sgRNA targeting when tethered to Cas9 or dCas9, despite the accompanying random integration from the casposase moiety.

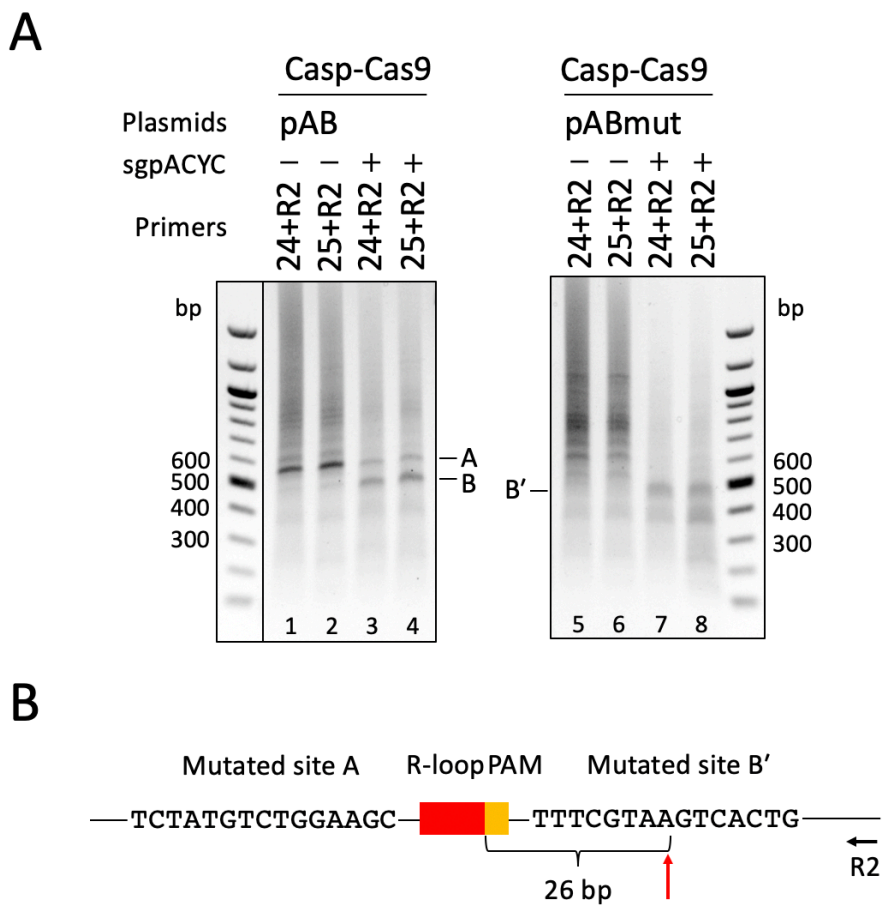


Figure 47. sgRNA-guided TK2425 integration by fusion proteins into pAB and pABmut.

(A) Casp-Cas9 with or without sgpACYC was incubated with pAB and pABmut respectively for TK2425 DNA integration. Integrated plasmid products were purified and PCR amplified to determine integration sites of TK2425. DNA bands A and B corresponded to site A and B integration products respectively. The DNA band B' corresponded to integration into mutated site B. DNA in the bands were gel extracted and sequenced (data shown in Appendix 7). (B) A schematic diagram showing sgpACYC-guided TK2425 integration site in pABmut. Red arrows indicate the insertion of TK24 or TK25 between the two bases. R2 represents the PCR primer and sequencing primer pUC19 ori R2.

The above PCR results confirmed that Casp-Cas9 and Casp-dCas9 catalysed targeted half-site integration, i.e., the integration of TK2425 at site B in one DNA strand in the presence of sgpACYC. We next examined if sgRNA-fusion protein complexes catalyse full-site integration, integration of both 3' ends of TK2425 DNA at site B in two DNA strands. For this, integration reactions were conducted using a site B hairpin DNA

substrate, which would result in different product size based on type of integration. The 154-nt long single-stranded oligo folds back on itself to form a 75 bp double stranded region from pACYC-Duet, which contains a sgpACYC target site and site B (Figure 48A). This site B hairpin DNA substrate was capable of forming a R-loop with sgpACYC by Cas9 and Casp-dCas9 (Figure 48B). Next, time course integration reactions were performed to detect full-site integration and determine the location of the first nucleophilic attack (Figure 48C). However, the results showed that Cy5 labelled TK24 was not integrated into the site B hairpin as there was no band up-shift observed. The inability of integrating DNA substrate into the site B hairpin might be explained by the double-stranded region being too short.

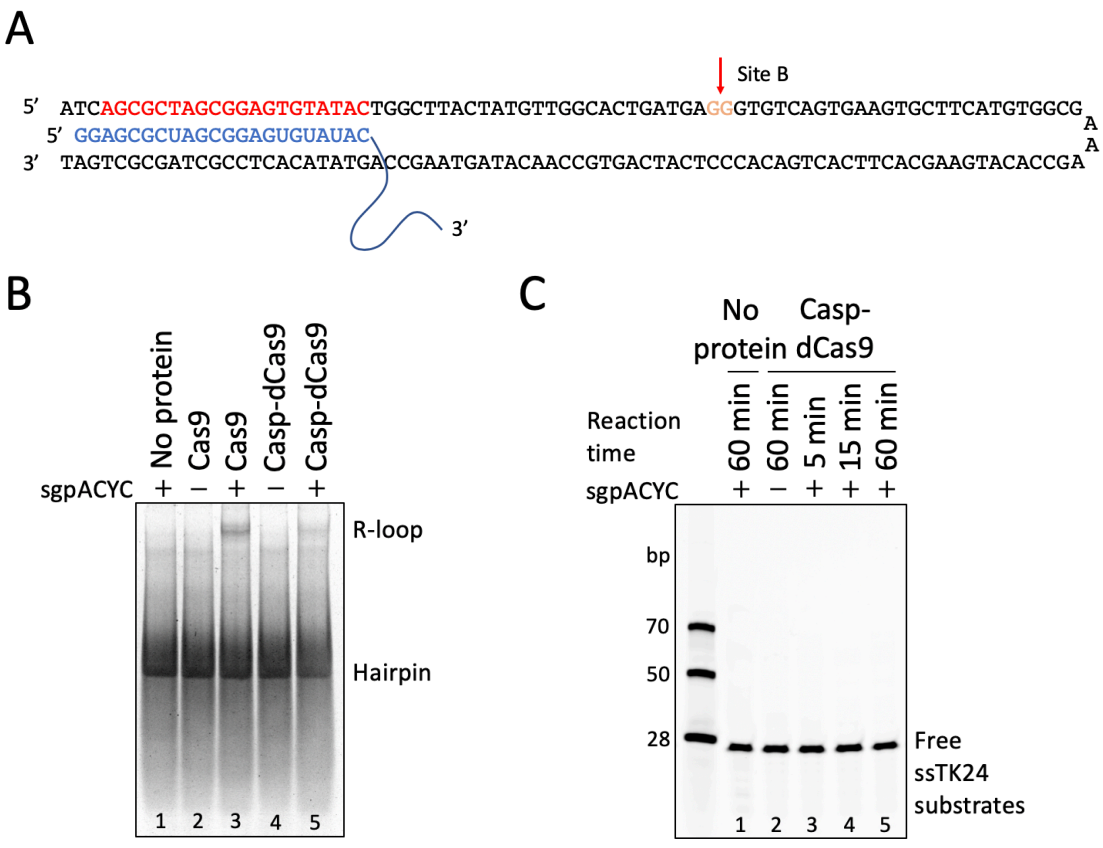


Figure 48. No integration was observed into site B hairpin by casp-hfdCas9.
(A) A schematic diagram showing a R-loop formed between a site B hairpin DNA substrate and sgpACYC RNA. A targeting sequence on the non-target strand was

labelled in red in the site B hairpin and a red arrow indicated an expected insertion of TK24 or TK25 at site B between the two guanines labelled in yellow. sgpACYC was labelled in blue and the first two guanine at the 5' end was added after in vitro transcription by T7 RNA polymerase. (B) R-loop formation between site B hairpin and sgpACYC by Cas9 and Casp-hfdCas9. Proteins at 150 nM were first incubated with or without 450 nM sgpACYC and then incubated with 450 nM site B hairpin at 37°C for 20 min. Reactions were treated with proteinase K before running on 10% polyacrylamide gels followed by visualisation using Diamond™ Nucleic Acid Dye. (C) Casp-hfdCas9 did not catalyse TK2425 integration into site B hairpin. Casp-hfdCas9 was incubated with site B hairpin as described in (B) and 30 nM Cy5 labelled TK2425 was added for incubation at 37°C. Reactions were stopped at 5, 15 and 60 min after addition of TK2425 and run on 10% polyacrylamide denaturing gel.

4.2.6. *In vitro* sgRNA-guided long DNA integration

Results above demonstrated fusion proteins are able to integrate a DNA oligo duplex TK2425 into site B, 26 bp downstream of the R-loop formation by sgRNA. We then tested if integration of a long DNA molecule can be guided to site B by sgpACYC. Mini-casposon, which was ampicillin resistance gene amp^R with the flanking of TIRs at both ends, was chosen to be the long DNA substrate because it was better integrated by casposase than those without TIRs (Table 3.1). The method was the same as for assays targeting short oligo integration except that 45 ng of mini-casposon in a 1:1 molar ratio of vector: insert was used (Figure 49A). Integrated plasmid products were ethanol precipitated and used as a template for PCR reactions to detect mini-casposon integration (Figure 49B). As expected, casposase gave the same PCR product, seen in previous assay in figure 28 lane 2, at approximately 550 bp (Figure 49B lane 2). Sequencing of this PCR product confirmed mini-casposon integrated into site A (Appendix 8). Fusion proteins with sgpACYC gave an additional band at about 480 bp which corresponded to site B integration identified by DNA sequencing (Figure 49B lanes 3-4 and Appendix 8).

The above PCR reaction results could not distinguish single-end integration products from double-end integration products as mentioned in section 3.2.4. Therefore, purified integration plasmid products were electroporated into DH5α to select for double-end integration plasmid products that conferred ampicillin resistance. Plasmids were purified from limited number of ampicillin resistant colonies and none of these showed integration at site B. The reason why a few colonies were recovered might be that integrated product plasmids were damaged during purification making them hard to transform. From purified plasmids, we observed integration at site A and elsewhere in pACYC-Duet. This off-target integration likely resulted from the casposase moiety competing with the Cas9/dCas9 moiety for binding to target plasmid because it is unlikely to find a sgRNA off-target site in pACYC-Duet with a few mismatches to the sgRNA and next to a PAM. In two integrated plasmids purified from reactions using Casp-Cas9 with sgpACYC, we observed deletions of up to 800 bp at one end of mini-casposon-plasmid junction. This might be caused by an error prone repair process during repair of the DSB generated by Cas9. Because the Cas9 moiety generates a DSB at target site resulting in unpredictable repair outcomes, further characterisation of Casp-Cas9 was discontinued.

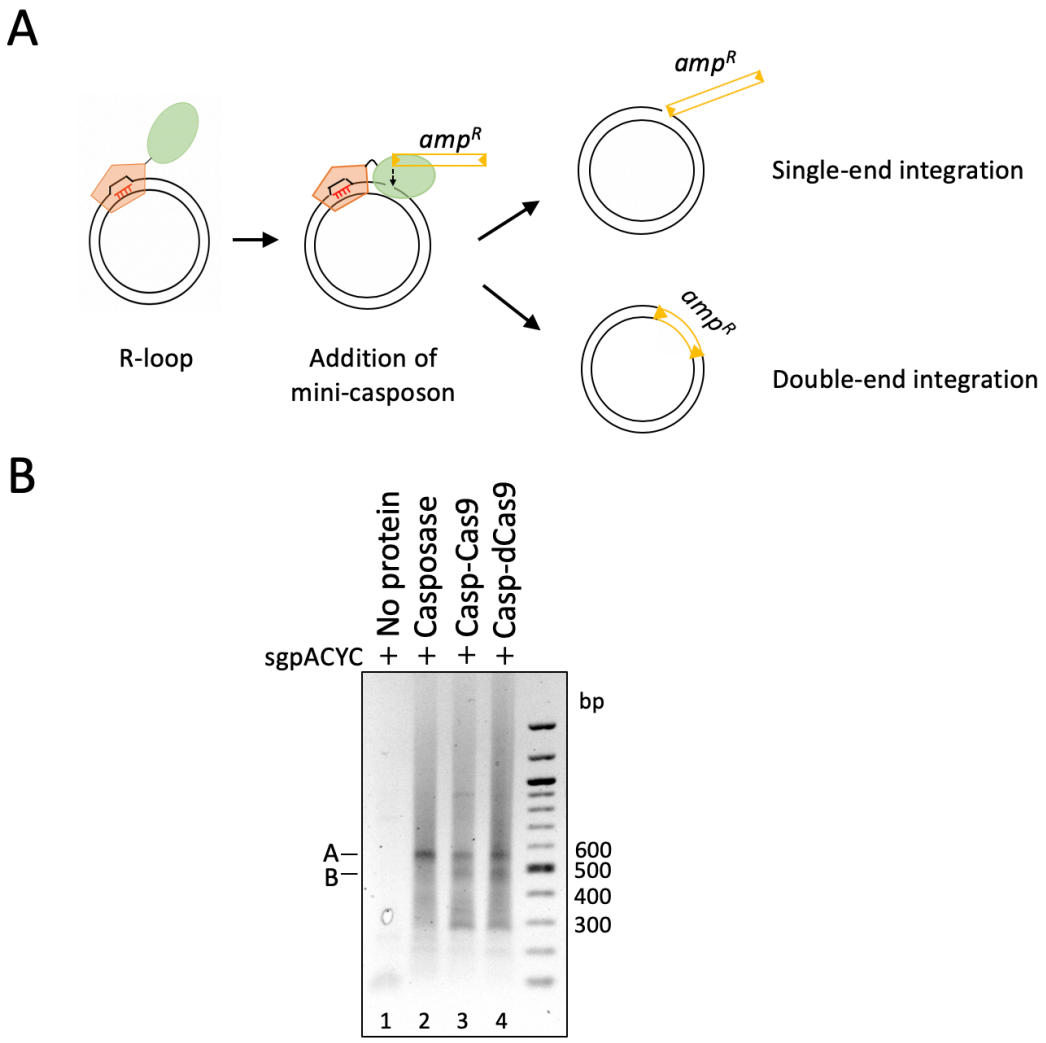


Figure 49. Fusion proteins catalysed gene insertion can be guided by sgRNA.
(A) A cartoon showing the targeted gene integration catalysed by fusion proteins and the resulting products. (B) PCR reactions showing the mini-casposon was integrated into the pACYC-Duet by casposase and fusion proteins. Integrated plasmid products were purified and PCR amplified to determine integration sites of mini-casposon. DNA bands A and B corresponded to site A and B integration products respectively.

4.2.7. Construction of high-fidelity fusion proteins and testing for sgRNA-guided DNA oligo integration

To rule out the possibility of off-target integration due to sgRNA-Cas9/dCas9, three mutations were introduced into Cas9 to increase the Cas9 specificity (Slaymaker et al., 2016). These mutations were K848A, K1003A and R1060A (Figure 50). The K848 residue in the HNH domain interacts with sgRNA while K1003 and R1060 in the RuvC

domain interact with non-target DNA strand to stabilise a R-loop. Neutralisation of these positively charged residues destabilises the R-loop and encourages rehybridization of the two DNA strands. A stable R-loop would require more stringent base pairing between the target DNA strand and sgRNA.

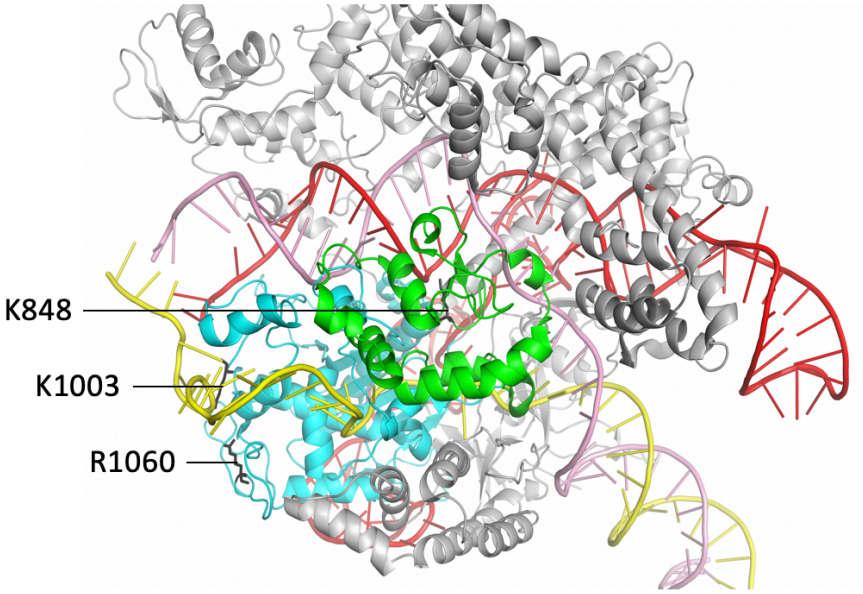


Figure 50. Crystal strucutre of DNA-sgRNA-Cas9 ternary complex
Cryo-EM structure of *S. pyogenes* Cas9 bound to sgRNA and double stranded target DNA (PDB: 5Y36). The RuvC domain was coloured cyan and the HNH domain was coloured green. A R-loop was formed among sgRNA (red), target DNA strand (pink) and non-target DNA strand (yellow). Three residues, K848, K1003 and R1060, stabilising the R-loop were labelled and shown as sticks. The image was created using PyMOL.

High-fidelity (hf) Cas9 containing the three mutations mentioned above was made and fused to casposase to generate Casp-hfCas9. From Casp-hfCas9, H840A mutation was introduced in the hfCas9 moiety to generate Casp-hfnCas9. D10A mutation was then introduced to the hfnCas9 moiety to generate Casp-hfdCas9 protein. The hf fusion proteins were purified the same as previous fusion proteins and a summary gel was shown in figure 51.

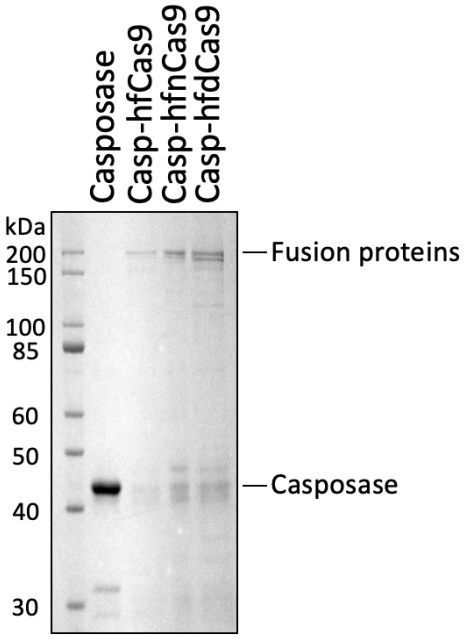


Figure 51. Summary SDS gel showing high-fidelity fusion proteins.

Purified hf fusion proteins were tested for targeted short DNA oligo integration. High-fidelity fusion proteins at 150 nM were incubated with 450 nM sgpACYC followed by incubation with 150 ng pACYC-Duet and 100 nM TK2425. Integrated plasmid products were ethanol precipitated for use in PCR reactions to detect sgRNA-guided DNA integration (Figure 52). The results showed TK2425 DNA substrate was integrated into site B by all hf fusion proteins in the presence of sgpACYC (lanes 3-5). However, integration into site A was also observed in the presence of sgpACYC suggesting the off-target integration in these assays and the assays in figures 46 and 49 are driven by autonomous activity of the casposase moiety in fusion proteins. Nevertheless, future *in vitro* assays were carried out using hf fusion proteins.

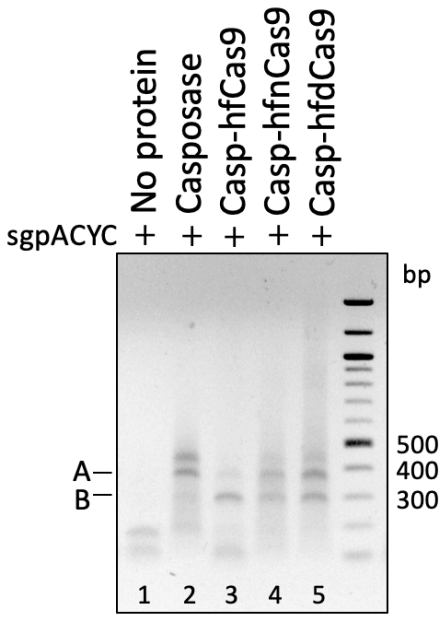


Figure 52. TK2425 was integrated into site B by high-fidelity fusion proteins in the presence of sgpACYC.

Proteins pre-incubated with sgpACYC were incubated with pACYC-Duet for TK2425 DNA integration at 37°C for one hour. Integrated plasmid products were purified and PCR amplified to determine integration sites of TK2425. Primers used in PCR reactions were TK25 and R. DNA bands A and B corresponded to site A and B integration products respectively.

4.2.8. *Casp-hfdCas9 catalysed In vitro* full-site integration was not detected

Casp-Cas9 and Casp-dCas9 fusion proteins catalysed short DNA oligo integration (Figure 45) detectable by fluorescence imaging of Cy5 labelled DNA substrate TK2425. Although PCR reactions detected sgRNA-guided integration of TK24 and TK25 at site B (Figure 46B), it was not clear whether the integration of the double-stranded TK2425 was half-site or full-site. To detect *in vitro* full-site integration, a reporter plasmid pCHL42 was constructed based on a previous study (Wright et al., 2019) (Figure 53). pCHL42 contains an extra 67 bp DNA sequence comprising sgRNA5 target sequence and site B flanking sequence from pACYC-Duet at the beginning of a chloramphenicol resistance (cm^R) gene encoding chloramphenicol acetyltransferase. This extra sequence makes the cm^R gene out of frame and pCHL42 does not confer resistance to

chloramphenicol. The open reading frame of the *cm*^{R} gene would be restored by a 65 bp (50 bp dsMW14 plus 15 bp TSD) insertion or 44 bp (29 bp CL5/6 plus 15 bp TSD) insertion resulted from full site integration of dsMW14 or CL5/6 respectively. Casp-hfdCas9 at 150 nM was first incubated with 450 nM sgRNA5 followed by incubation with 150 ng pCHL42 and 200 nM dsMW14 or CL5/6. Integrated plasmid products were purified and transformed into DH5α to select for chloramphenicol resistant colonies. The result showed no colony grew on the chloramphenicol plate suggesting the DNA substrates were not integrated or the sgRNA-guided integration was half-site integration.

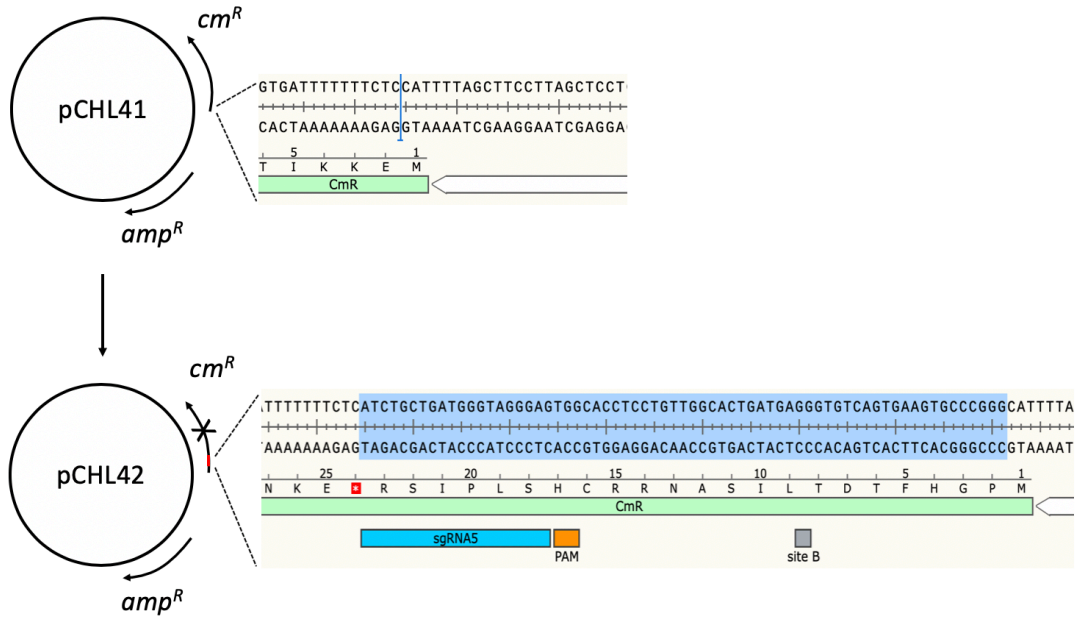


Figure 53. Schematic diagram showing chloramphenicol resistance gene structure of a reporter plasmid pCHL42.

A 67 bp DNA sequence (highlighted) containing sgRNA5 target sequence (blue box) and site B for sgRNA-guided integration was inserted into the beginning of a *cm*^{R} gene in pCHL41. This made the *cm*^{R} gene out-of-frame with an early stop codon (red star) so the resulting plasmid pCHL42 confers only ampicillin resistance. Full-site integration at site B (grey box) of dsMW14 and CL5/6 would result in 65 (50 bp dsMW14 plus 15 bp TSD) and 44 bp (29 bp CL5/6 plus 15 bp TSD) insertion into *cm*^{R} gene restoring the open reading frame.

4.2.9. Casp-hfnCas9 catalysed ssDNA integration and DNA replacement were not detected

Previous results showed casposase and Casp-dCas9 were able to integrate a ssDNA oligonucleotide into a pACYC-Duet plasmid (Figures 25 and 45B). We therefore hypothesised a genome editing strategy in which casp-hfnCas9 fusion protein can introduce a base substitution or small insertion/deletions (Figure 54A). To test this, pUC19 was used as a target plasmid and blue/white screening was used as a reporter to detect successful gene editing. Casp-hfnCas9 at 150 nM was first incubated with 450 nM sgpUC19 RNA that targets the 5' end of a *lacZ*α gene in pUC19. The RNP complex was then incubated with 150 ng pUC19 to form a R-loop and the functional RuvC domain in the nCas9 moiety of the Casp-hfnCas9 generated a single-strand break at a site three bp upstream of the PAM on the non-target strand (Figure 54A red arrow). Then a ssDNA oligonucleotide CL7 at 200 nM, which contains the same sequence as the non-target strand except for an additional adenine, was added for sgpUC19-guided integration. According to the sgpACYC-guided site B integration, CL7 would be integrated 26 bp downstream of the PAM. After integration, CL7 might displace the top strand that was cleaved at both ends and bind to the bottom strand by complementary base pairing except for insertion of one additional adenine. Successfully integrated CL7 might be used as a template to repair the mismatch in the integrated plasmid products inside a cell. This would introduce a one-bp insertion into the *lacZ*α gene causing a frameshift and early termination in translation. This would be detected by transforming products into DH5α that would grow as white colonies when grown on agar containing X-gal and IPTG, because α-complementation of the chromosomal β-galactosidase gene by pUC19 would be lost (Ullmann, 1992). The

results showed white colonies on the X-gal plate but these white colonies did not grow after re-streaking on a new plate suggesting satellite colonies (Figure 54B). One possible reason for the lack of positive outcome from this assay might be because the ssDNA oligo CL7 contains a guanine and thymine at the -3 and -2 position from the 3' end that might be disfavoured by casposase (Figure 24B). There were also several uncertainties in this experiment, for example, displacement of non-target strand by CL7 and DNA repair using CL7 as template.

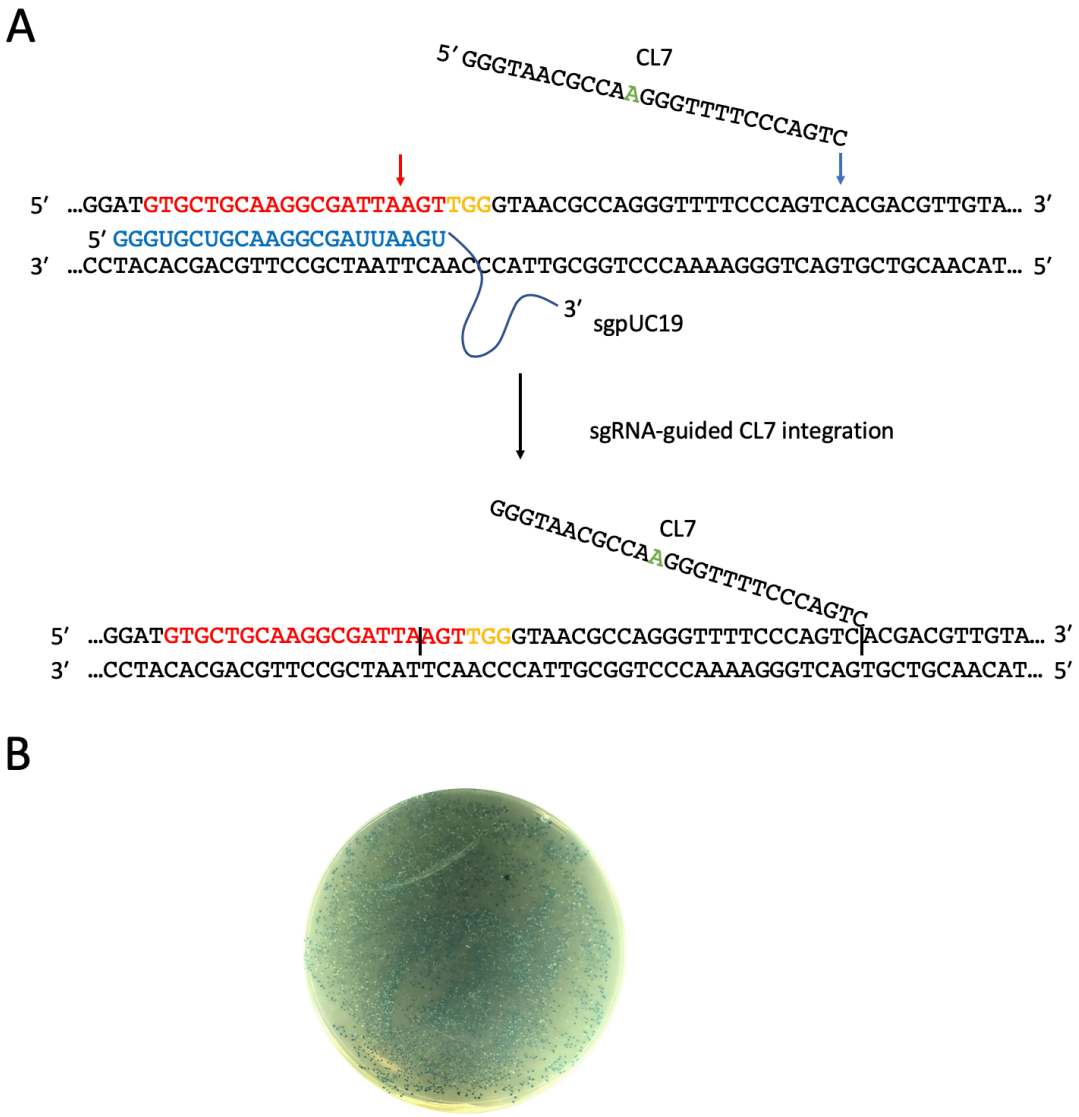


Figure 54. Casp-hfnCas9 catalysed ssDNA integration and DNA displacement were not detected.

(A) Schematic diagram showing Casp-hfnCas9 catalysed sgRNA-guided ssDNA integration and non-target strand cleavage in *lacZ*α gene in pUC19. The sgRNA targeting sequence on the non-target strand in *lacZ*α gene was labelled in red with a PAM in yellow and the sgpUC19 RNA was labelled in blue. Casp-hfnCas9 generated a single-strand break at the non-target strand (red arrow). A single-stranded oligo CL7 would be integrated into a site 26 bp from the PAM in a sgRNA-dependent manner (blue arrow). The integrated CL7 might displace a partial plasmid sequence cleaved at both ends on the top strand and bind to the complementary sequence on the bottom strand with one mismatch. (B) A blue/white screening to detect Casp-hfnCas9 catalysed CL7 integration and CL7 replacement of plasmid sequence. White colonies would correspond to positive results.

4.2.10. Casp-Cas9 catalysed DNA integration *in vivo* was not detected in *E. coli* cells

To assess if Casp-Cas9 fusions could catalyse DNA integration *in vivo* using *E. coli* cells, the protocol for expression of the fusion proteins was modified from that used for protein purification. This was because cells at OD_{600}=0.6 needed to be made electro-competent for transformation of linear DNA substrates. According to the optimised protein expression protocol for *in vivo* integration, fusion protein expression was induced at OD_{600}=0.2 and lasted until OD_{600}=0.6. Western blotting using anti-6xHis antibody confirmed that Casp-dCas9 protein was stably expressed in *E. coli* EB377 strain, which is K-12 MG1655 carrying a T7 RNA polymerase gene (Figure 55).

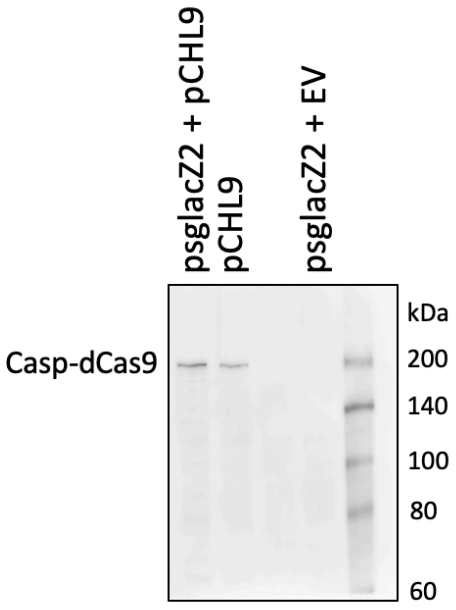


Figure 55. Casp-dCas9 expression using expression protocol for in vivo integration was confirmed by western blotting.
E. coli strain EB377 was transformed with pCHL9 for Casp-dCas9 expression and psglacZ2 for sgRNA expression. EB377 harbouring an empty vector (EV) was used as a control. Cells were induced protein expression at OD_{600}=0.2 and were harvested for SDS-PAGE at OD_{600}=0.6. Casp-dCas9 expression was confirmed by western blotting using anti-6xHis antibody.

We next tested if fusion proteins could bind to target DNA in a sgRNA-dependent manner *in vivo*. Two sgRNAs targeting to a *lacZ* gene, which was either on a plasmid or on a chromosome, were tested and β-galactosidase activity was measured by Miller assays as a read-out (Figure 56A). *E. coli* strain CL003, K-12 MG1655 with *lac* operon deletion and inducible T7 RNA polymerase production was transformed with pRC7, a very low copy number plasmid containing the *lac* operon. Then cells were respectively transformed with pCHL35, pCHL37 and pCHL50 for expressing fusion proteins and sglacZ RNA on the same plasmid. Miller assays showed co-expression of fusion proteins and sglacZ resulted in a large decrease in β-galactosidase activity (Figure 56B). The relative Miller unit dropped to nearly zero in cells expressing Casp-hfCas9-sglacZ1 because repair of the DSB generated by hfCas9 led to insertions/deletions in the *lacZ* gene. It was observed that the targeting efficiency of sglacZ2 was better than sglacZ1 consistent with previous studies showing that different sgRNAs targeting to the same gene show different editing efficiency (Doench et al., 2014; Xu et al., 2017). The difference in sgRNA targeting efficiency might be explained by the stability of R-loops formed between target DNA and sgRNA. Cells expressing Casp-hfdCas9-sglacZ2 showed nearly five-fold reduced β-galactosidase activity compared with cells expressing only Casp-hfdCas9. Similar observations were noted in Miller assays using BL21-AI strain for chromosomal *lacZ* targeting (Figure 56C). The target efficiency of Casp-dCas9-sglacZ2 was the same as dCas9-sglacZ2 that they both caused five-fold decrease in β-galactosidase activity compared with cells harbouring an empty vector. In addition, it was observed that cells expressing fusion protein and sglacZ1 on different plasmids showed lower β-galactosidase activity than those expressing fusion protein and sglacZ1 on the same plasmid. This might be explained by the amount of

sglacZ1 present in cells due to the difference in copy number of the sglacZ1 expressing plasmids. To conclude, sglacZ2 bound fusion proteins effectively target *lacZ* gene on a plasmid or on a chromosome.

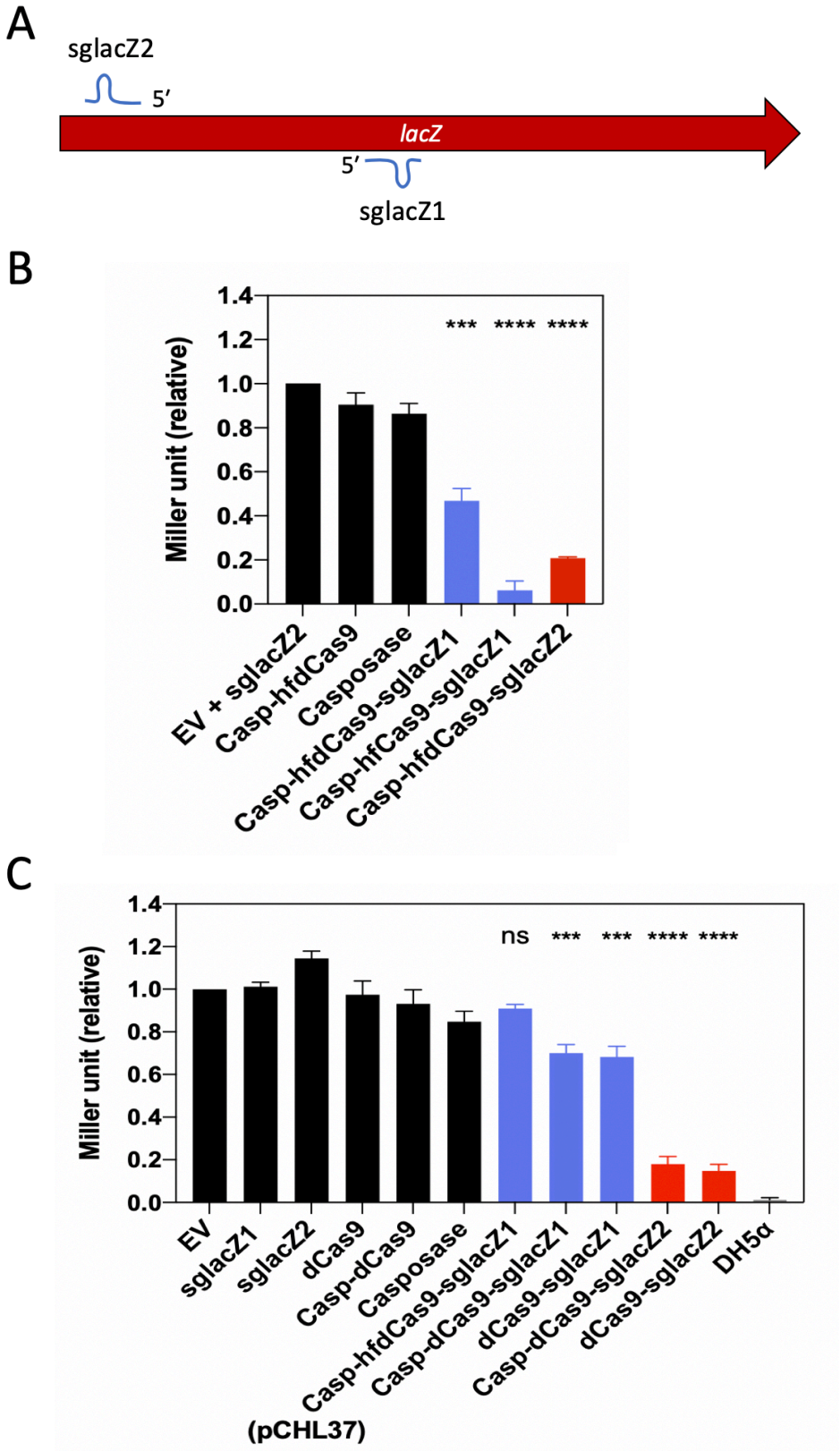


Figure 56. Fusion proteins specifically bound to lacZ gene in the presence of sglacZ RNA revealed by Miller assays

(A) Schematic representation of targeting sites of sglacZ1 and sglacZ2 RNA in *lacZ* gene. (B and C) Miller assays confirmed fusion proteins targeting of *lacZ* gene on an

extremely low copy number plasmid pRC7 (B) or on the chromosome (C). Miller units were normalised to a 'empty vector' (EV) control that lacks protein encoding genes. Columns coloured in black represent controls. Blue columns represent samples using sglacZ1 and red columns represent sglacZ2 samples. One-way ANOVA with multiple comparison to EV control was performed to determine statistical significance. ***: p-value ranges from 0.0001 to 0.001, ****: p-value < 0.0001, ns: non-significant. (B) *E. coli* CL003 strain harbouring pRC7 was transformed with a plasmid encoding protein only, sglacZ RNA only or encoding both protein and sglacZ RNA. N=2. (C) *E. coli* BL21-AI strain was transformed with plasmids for individually expressing proteins and sglacZ RNA. An exception was pCHL37 that expresses both Casp-hfdCas9 and sglacZ1. N=3. Error bars represent standard error of the mean.

We next extended this method by introducing TK2425 (3 µM) by electroporation into CL003 cells expressing a fusion protein and sglacZ1 that targets *lacZ* gene on pRC7. One microliter of cells was lysed after recovery from electroporation and was used in PCR as template to detect both half-site and full-site integration (Figure 57A). Targeted integration by sglacZ1 bound hf fusion proteins was expected to give a 1.2 kb PCR product when primers lacZ F and TK24 or TK25 were used. PCR reactions using TK24 did not show DNA amplification while PCR reactions using TK25 showed the expected DNA band at 1.2 kb. However, DNA sequencing followed by a BLAST search revealed non-specific binding of TK25 and lacZ F primers to *E. coli* chromosomal DNA (Appendix 9). The remaining electroporated cells after recovery were diluted 10^7-fold and plated on X-gal agar plate for blue/white screening (Figure 57B). Cells expressing Casp-hfCas9-sglacZ1 showed white colonies that are consistent with a *lacZ* gene that has been disrupted. Colony PCR suggested the *lacZ* gene was deleted in these cells upon repair of hfCas9 generated DSB (Figure 57C). Cells expressing casposase, Casp-hfnCas9-sglacZ1 or Casp-hfdCas9-sglacZ1 all showed blue colonies on X-gal plates representing intact *lacZ* gene that was verified by colony PCR (Figure 57B and 57C).

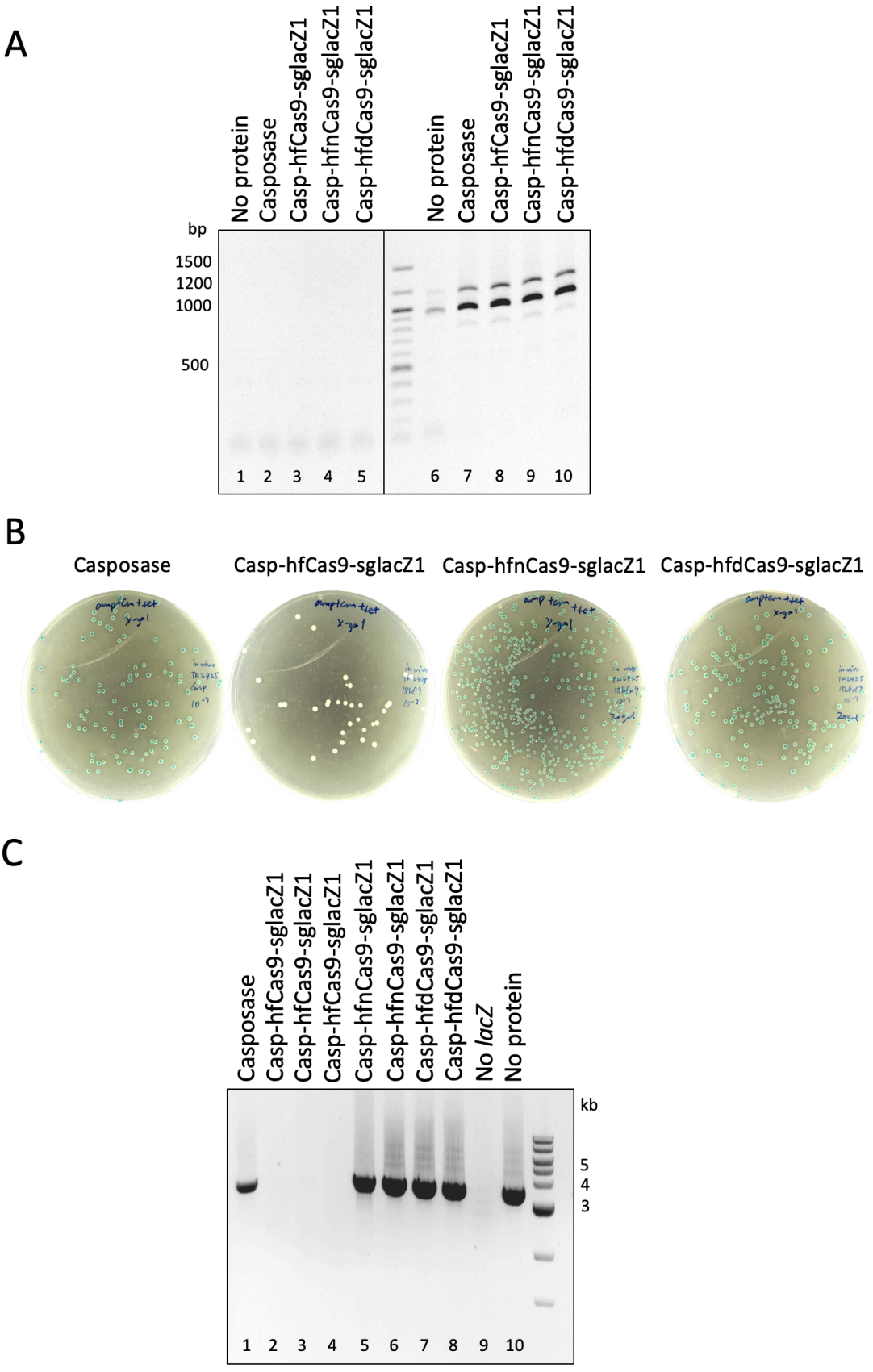


Figure 57. In vivo TK2425 integration was not detected.

(A) Liquid culture PCR reactions of cells transformed with TK2425 did not detect in vivo TK2425 integration. CL003 harbouring pRC7 and expressing casposase or fusion

proteins with sglacZ1 were electroporated with 3 µM of TK2425. After one-hour recovery in SOC, cells were diluted 100-fold in sterile distilled water and one µL of lysed cells was used in PCR as template. Primers used were TK24 and lacZ F (lanes 1-5), TK25 and lacZ F (lanes 6-10). (B) Blue/white screening of cells transformed with TK2425. Cells electroporated with TK2425 in (A) were plated on agar plates containing X-gal for blue/white screening. Blue colonies corresponded to intact *lacZ* gene and white colonies corresponded to edited *lacZ*. (C) Colony PCR showed cells expressing Casp-hfCas9-sglacZ1 had *lacZ* gene deleted. Colonies appeared on X-gal plates in (B) were picked for colony PCR to amplify *lacZ* gene using lacZ F and lacZ R primers. 'No *lacZ*' control was from a colony of CL003 and 'no protein' control was from CL003 harbouring pRC7.

We repeated the *in vivo* integration assay targeting *lacZ* on pRC7 using dsMW14 as DNA substrate (Figure 58). Cells expressing casposase or Casp-hfdCas9-sglacZ2 were electroporated with three µM dsMW14 and plated on X-gal agar plates for blue/white screening. Full-site integration of dsMW14 would result in frameshift of the *lacZ* gene giving white colonies on a X-gal plate. However, we observed no white colonies from cells expressing Casp-hfdCas9-sglacZ2 (Figure 58).

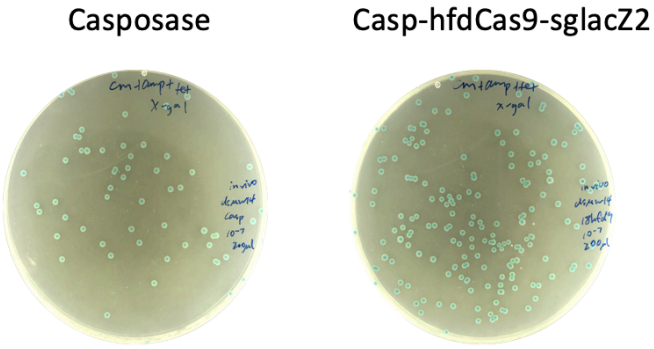


Figure 58. In vivo dsMW14 integration was not detected by blue/white screening. *E. coli* CL003 strain harbouring pRC7 and expressing casposase or Casp-hfdCas9-sglacZ2 were electroporated with 3 µM of dsMW14. Cells were plated on agar plates containing X-gal for blue/white screening. Blue colonies corresponded to intact *lacZ* gene and white colonies corresponded to edited *lacZ*.

The *in vivo* TK2425 integration experiments were next repeated to target chromosomal *lacZ* of *E. coli* EB377 strain that is K-12 MG1655 carrying T7 RNA polymerase gene. Because sglacZ2 RNA showed better targeting efficiency than sglacZ1 (Figure 56C), cells expressing sglacZ2 RNA and hf fusion proteins were electroporated with 3 μM of TK2425. One microliter of cells after recovery was lysed and diluted 100-fold for PCR amplification of the whole *lacZ* gene (Figure 59A and 59B). The results showed cells expressing Casp-hfCas9-sglacZ2 gave a product just below 500 bp indicating deletion of the *lacZ* gene (Figure 59B lane 4). Cells expressing Casp-hfnCas9-sglacZ2 and Casp-hfdCas9-sglacZ2 both gave a product at 3.3 kb corresponding to intact *lacZ* gene. DNA of the 3.3 kb bands were extracted and subjected to a second round of PCR to detect TK2425 integration (Figure 59C). However, we did not observe a DNA band at 300 bp that would correspond to sglacZ2-guided integration.

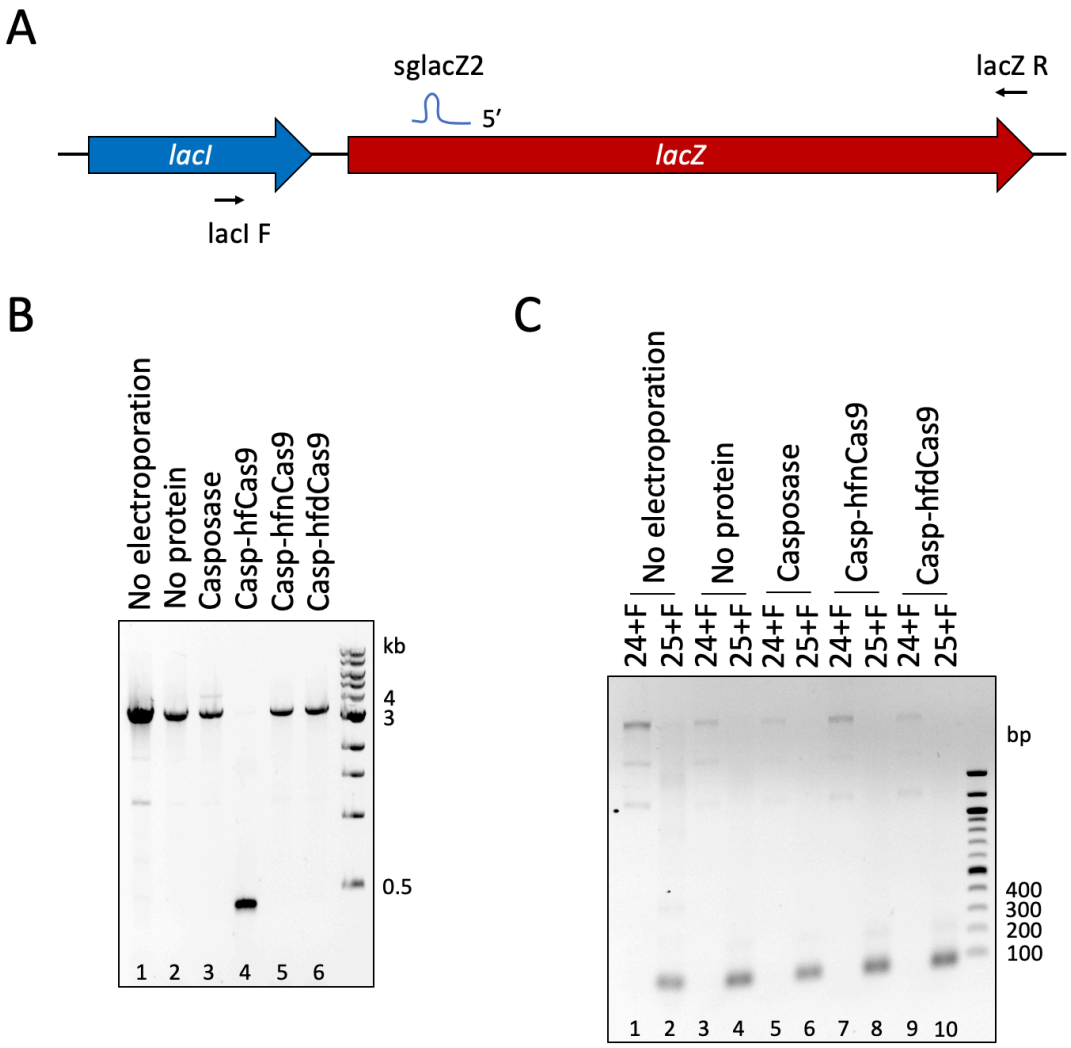


Figure 59. Liquid culture PCR did not detect in vivo Tk2425 integration.
(A) Schematic representation of sglacZ2 targeting site in *lacZ* gene and primers used in PCR reactions. (B) Amplification of *lacZ* gene after electroporating 3 µM TK2425 into EB377 expressing different proteins and sglacZ2. Primers used were lacI F and lacZ R and the expected product is at 3.3 kb. 'No electroporation' control was EB377 expressing Casp-hfdCas9 and sglacZ2 before electroporation of the TK2425 substrate. 'No protein' control was EB377 containing pACYC-Duet empty vector and sglacZ2 expressing plasmid. (C) The second round of PCR using DNA extracted from the 3.3 kb product in (B) did not detect TK24 or TK25 integration. Primers used were TK24 and lacI F or TK25 and lacI F. If targeted integration occurred, the expected product size would be 300 bp.

EB377 cells after recovery of electroporating TK2425 into cells were plated on agar plate containing X-gal (Figure 60A). The result was consistent with the above PCR results that cells expressing Casp-hfCas9-sglacZ2 gave white colonies representing disrupted *lacZ* gene. Cells expressing Casp-hfnCas9-sglacZ2 and Casp-hfdCas9-sglacZ2 gave blue colonies representing functional β-galactosidase. Colony PCR was carried out to detect TK2425 integration, in case TK2425 integration did not change the open reading frame resulting in functional β-galactosidase (Figure 60B and 60C). Two rounds of PCR were carried out to increase the PCR specificity. Results of the first round of PCR were consistent with the blue/white screening results that cells expressing Casp-hfnCas9 and Casp-hfdCas9 showed an expected product at 3.3 kb (Figure 60B lanes 3-5). The second round of PCR using DNA extracted from the 3.3 kb band did not show sglacZ2-guided TK2425 integration that should have given a product at 300 bp (Figure 60C).

We also tested if a mini-casposon (150 ng) could be integrated into the chromosomal *lacZ* of EB377 cells expressing Casp-hfnCas9-sglacZ2 or Casp-hfdCas9-sglacZ2. Cells were plated on agar plates containing ampicillin and X-gal to detect targeted integration, but no colonies were observed. Our inability to detect *in vivo* DNA integration by the casposase moiety in fusions with Cas9 might be owing to different reaction conditions compared with *in vitro* assays or rapid degradation of linear DNA substrate by host's RecBCD complex.

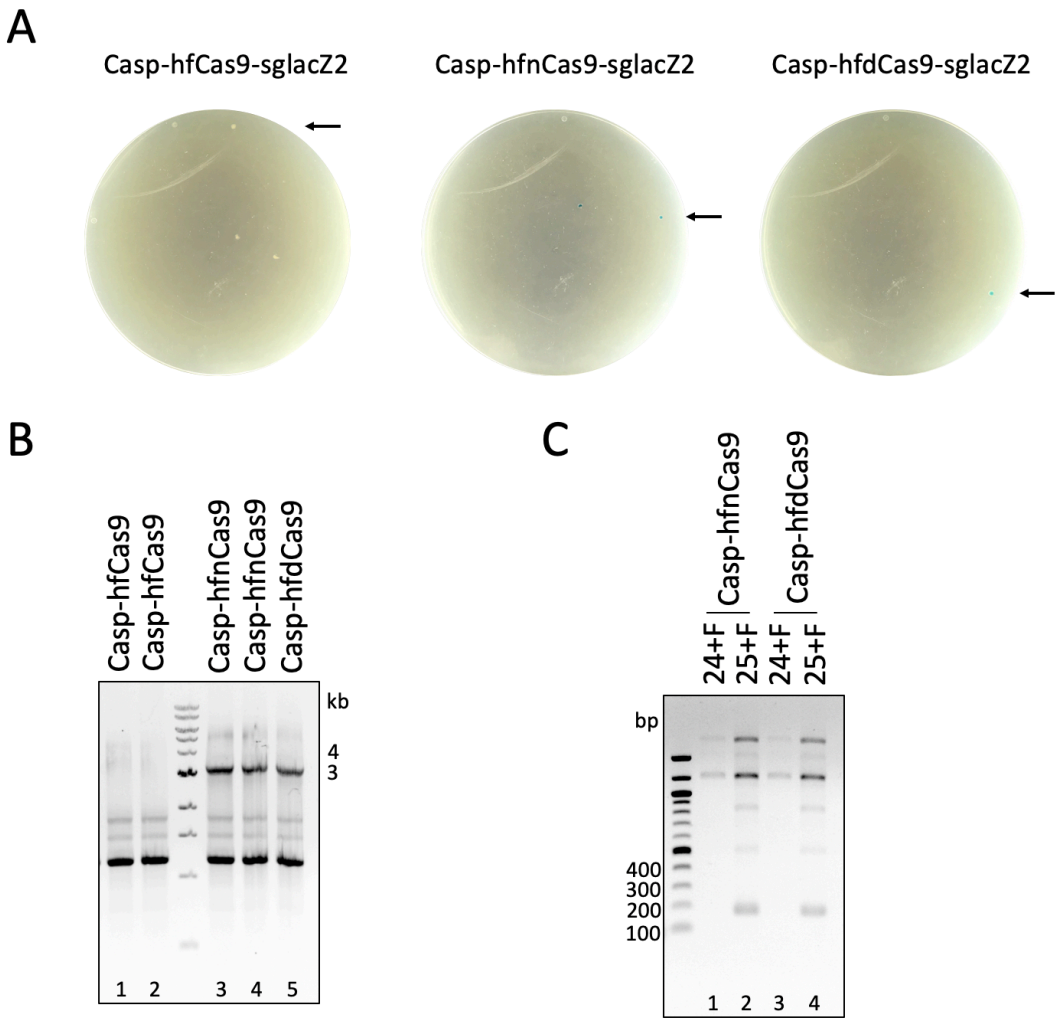


Figure 60. Colony PCR did not detect in vivo TK2425 integration.
(A) Blue/white screening of cells transformed with TK2425. EB377 expressing high fidelity fusion proteins and sglacZ2 were electroporated with 3 µM of TK2425. Cells were diluted 10^7-fold and plated on X-gal agar plate. Black arrows indicate the location of colonies. (B) Colonies seen in (A) were picked for PCR to amplify the entire lacZ gene. Primers used were lacI F and lacZ R and the expected product was 3.3 kb long. (C) DNA of the 3.3 kb band in (B) was extracted and subjected to second round of PCR to detect TK24 or TK25 integration. Primers used were TK24 and lacI F or TK25 and lacI F. sglacZ2-guided integration product was 300 bp long.

4.2.11. *In vivo* gene editing trials in a human cell line

The Casp-Cas9 high-fidelity fusion proteins were also tested for their activities in a human osteosarcoma U2OS cell line. Before assessing activities *in vivo,* we first assayed for DNA integration catalysed in cell-free extracts (CFE) of the U2OS cells, to mimic *in vivo* conditions more closely than *in vitro* assays containing only a few defined

components. Previous *in vitro* DNA integration assays were conducted in the presence of 5 mM Mn^{2+} that is scarce in living cells (Bischof et al., 2019; Varga et al., 2005). Thus, a more physiologically relevant divalent cation Mg^{2+} was used as an alternative. Casposase at 150 nM concentration was incubated with 250 ng of pACYC-Duet and 100 nM Cy5 end-labelled TK2425 in a reaction buffer containing 5 mM Mn^{2+} or 5 mM Mg^{2+} and in the presence or absence of 18 µg of U2OS CFE (Figure 61). Casposase integrated the DNA substrate less effectively in Mg^{2+} than in Mn^{2+} (Figure 61A and 61B lanes 3-4, 7-8) and this was consistent with the previous results seen in casposase-catalysed disintegration (Figure 19B). The DNA integration efficiency was decreased in the presence of U2OS CFE probably due to large amount of proteins present. However, these assays indicated that casposase is active at integrating DNA using Mg^{2+} in human CFE, albeit at lower efficiency. This suggests the casposase moiety in fusion proteins might be functional under *in vivo* conditions.

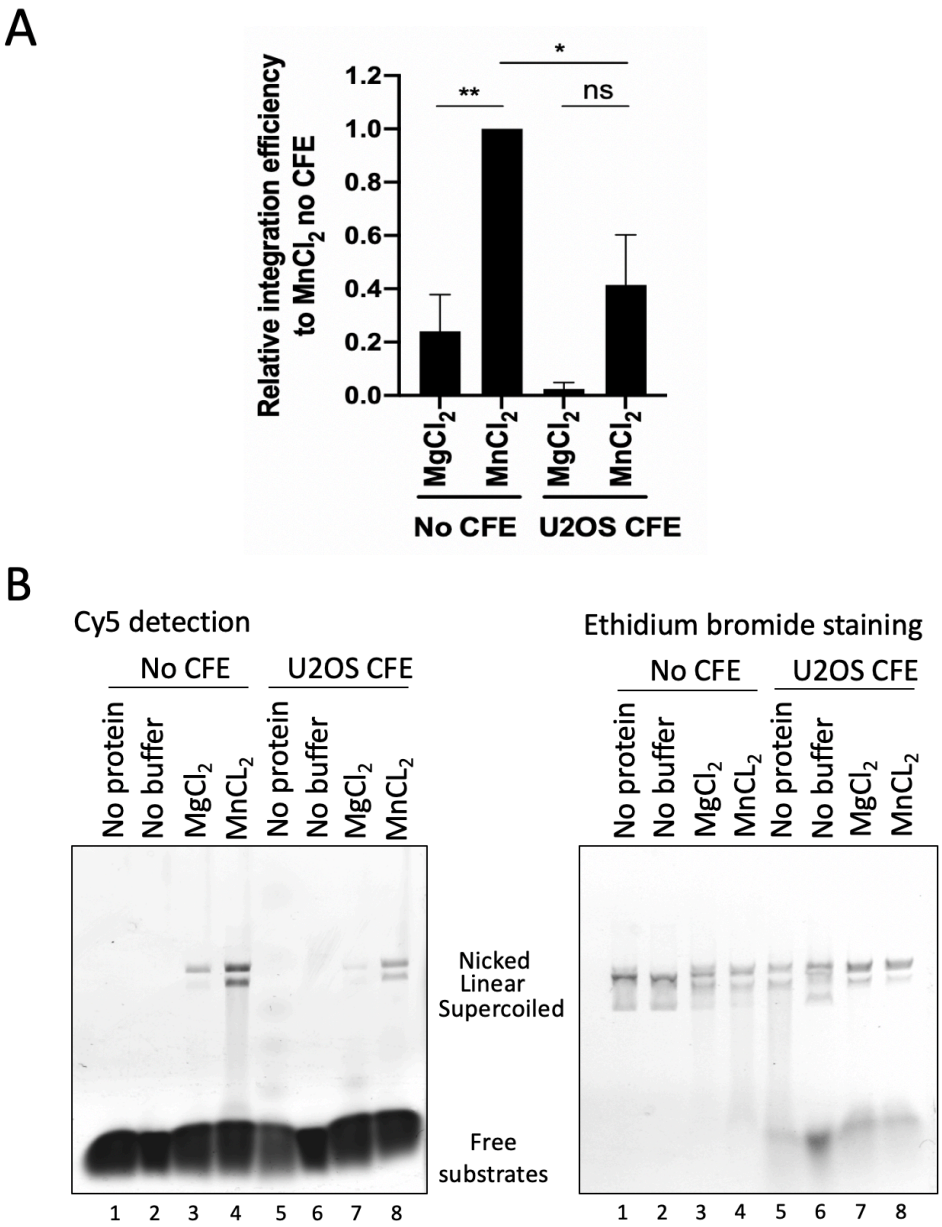


Figure 61. Casposase catalysed in vitro DNA integration in the presence of human cell-free extract.

(A) Quantification of band intensity from TK2425 DNA integration catalysed by casposase in different reaction conditions. The band intensity was normalised to reactions using $MnCl_2$ buffer and without cell-free extract to obtain relative integration efficiency. Error bars represent standard error of the mean. N=3. One-way ANOVA with multiple comparison was performed to determine statistical significance. *: p-value <0.05, **: p-value<0.01, ns: non-significant. (B) A representative gel for quantification in (A). Casposase at 150 nM was incubated with 250 ng pACYC-Duet and 100 nM Cy5 labelled TK2425 at 37°C for one hour in $MnCl_2$ containing integration buffer used in previous *in vitro* assays or in the same buffer with 5 mM $MgCl_2$ substituting for 5 mM $MnCl_2$. The reactions were also repeated in the presence of 18 µg of U2OS CFE. Integration of TK2425 into pACYC-Duet was assessed using ethidium bromide to indicate supercoiled, linear and nicked plasmids (right), and fluorescence to detect Cy5 labelled TK2425 that was integrated in to the plasmid (left).

To test Casp-Cas9 for genetic editing in eukaryotic cells, it was necessary to add a nuclear localisation signal to the protein sequence for protein import into the cell nucleus (Kosugi et al., 2009). High-fidelity fusion proteins Casp-hfnCas9 and Casp-hfdCas9, and additionally unfused Cas9, were sub-cloned into pNLS-His-StrepII plasmid so the proteins contain N-terminal 6xHis tag followed by an NLS and C-terminal NLS followed by a Strep-tag® II (Figure 62A). NLS-Cas9 was purified through Ni^{2+} column followed by Strep-Tactin® resin (Figure 62B). It was observed that NLS-hf fusion proteins did not bind to Strep-Tactin® resin (data not shown) probably because the Strep-tag® II was not exposed on the protein surface. Therefore, the NLS-hf fusion proteins were purified through Ni^{2+} column followed by heparin column as previous fusion proteins. A summary SDS polyacrylamide gel of purified NLS containing proteins was shown in figure 62C.

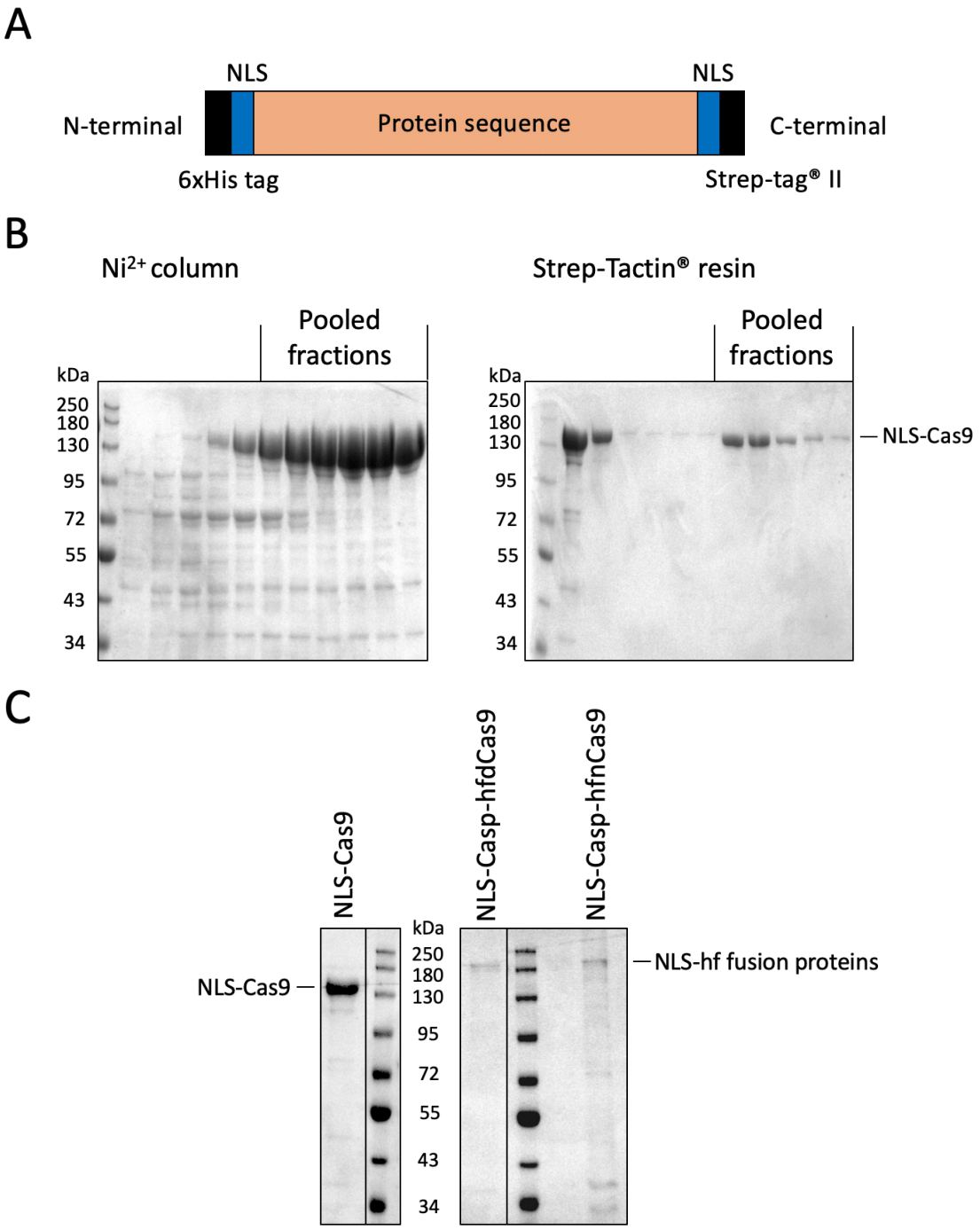


Figure 62. Purification of NLS containing proteins for in vivo gene editing in human cells.

(A) Schematic representation of NLS containing protein construct. (B) Purification of NLS-Cas9 through Ni^{2+} column followed by Strep-Tactin® resin. Elution fractions between two black lines were pooled together. (C) Summary SDS polyacrylamide gel of purified NLS containing proteins.

The *in vivo* gene editing assays in human cells utilised EGFP as a reporter. Thus, before carrying out *in vivo* gene editing assays, an EGFP expressing cell line was generated by transfecting pEGFP-c1 plasmid into an U2OS cell line. The transfected cells were cultured in the presence of 400 μg/mL of G418 antibiotic for more than two weeks to select for cells harbouring the EGFP expressing plasmid. Cells were subsequently cultured in 200 μg/mL of G418 for plasmid maintenance while minimising its effect on cells. At the end of selection, cells were imaged at four different locations under a fluorescence microscope and the average percentage of EGFP expressing cells in the population was 57% (Figure 63).

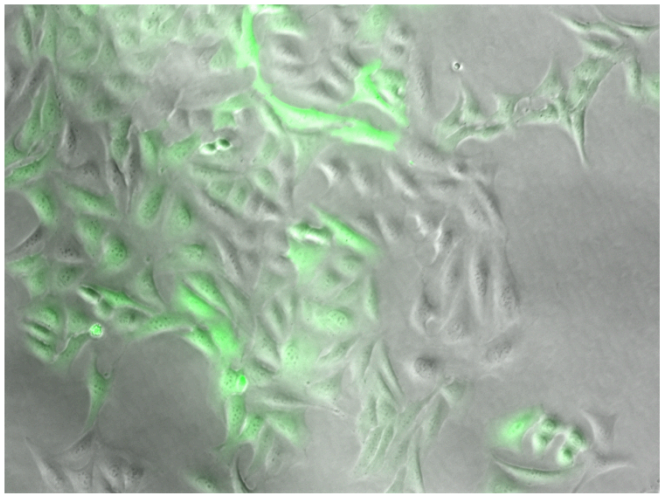


Figure 63. Image of EGFP expressing U2OS cells.
A representative image of U2OS cells transfected with pEGFP-c1 plasmid taken at four different locations. Green cells were successfully transfected cells expressing EGFP and dark cells were non-transfected cells. The EGFP gene on the pEGFP-c1 plasmid was used as a target for experiments utilising NLS-Cas9, and the NLS-Casp-Cas9 fusions.

The readout for successful DNA integration that edits the EGFP gene in cells was EGFP conversion to BFP resulting from the integrated DNA conferring two amino acid substitutions, T66S and Y67H (Glaser et al., 2016) (Figure 64A). The results could be

quantified by flow cytometry to measure the editing efficiency. Cas9 mediated HDR using a ssODN was used as a positive control (Figure 64B). Cas9 generates a DSB three bp from a PAM in the EGFP gene and the DSB would be repaired using a ssODN with 60 base homology arms at both ends flanking the cut site as a template. The ssODN introduces two base substitutions, 194C > G and 196T > C, in close proximity to the Cas9 cut site. For gene editing using casposase-Cas9 mediated recombination (CaspoR), Casp-hfnCas9 and Casp-hfdCas9 might integrate a single-stranded EGFP CaspoR oligo 26 bp in the 3' direction from the PAM (Figure 64C). The EGFP CaspoR oligo carries the same DNA sequence as the antisense strand of EGFP gene except for two different bases at the 3' end for base substitutions mentioned above.

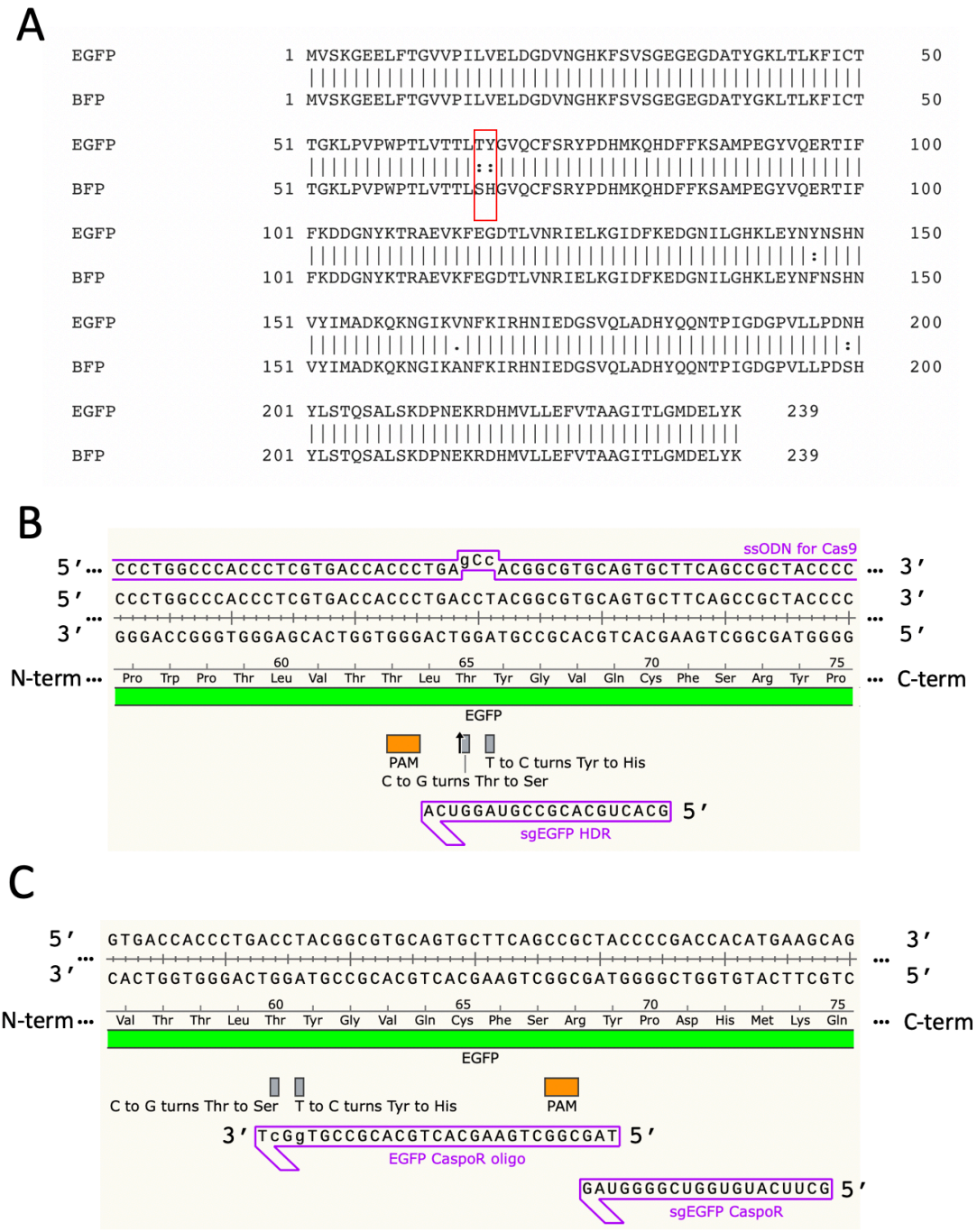


Figure 64. Conversion of EGFP to BFP by Cas9 mediated HDR using a ssODN or Casp-Cas9 fusions mediated ssDNA integration.

(A) Protein sequence alignment between EGFP and BFP by EMBOSS Needle. Two amino acids indicated by a red box correspond to the difference in the fluorescence excitation and emission spectra of the proteins. (B) CRISPR-Cas9 mediated HDR using a ssODN as a template to convert EGFP to BFP. Only the targeting sequence of sgEGFP HDR is shown. A black arrow indicates the Cas9 cleavage site. (C) Casposase-Cas9 mediated recombination (CaspoR) using a ssDNA oligo that converts EGFP to BFP. The reaction utilised casp-hfnCas9 or casp-hfdCas9 and sgEGFP CaspoR that shows only the targeting sequence. The EGFP CaspoR oligo would introduce two base substitutions which are shown in lower case if the oligo was integrated.

For both gene editing strategies, RNP complexes were first assembled *in vitro* and delivered into cells by electroporation using two methods, the Neon transfection system (Thermo Fisher), and by RNP transfection using jetCRISPR® transfection reagent. Because the Neon transfection system requires high concentration of RNP complexes containing 1 μM proteins and purified proteins and sgRNAs were not sterile, severe contamination of EGFP expressing U2OS cells was observed 24 hours after electroporation. On the other hand, the jetCRISPR® transfection method required lower concentration of RNP complex for transfection. For Cas9 mediated gene editing, NLS-Cas9-sgEGFP HDR RNP complex (30 nM) and ssODN for Cas9 (1 μg) were transfected into 1.5×10^5 EGFP expressing U2OS cells. However, in this preliminary analysis of CaspoR mediated gene editing, a final concentration of 3 nM NLS-hf fusion protein RNP complexes and EGFP CaspoR oligo (1 μg) were transfected into cells. This was because of low quantity of purified NLS-hf fusion proteins. Plasmid pEGFP-c1 was purified from transfected cells by QIAGEN miniprep kit 48 hours after transfection. A 530 bp EGFP DNA fragment flanking the two sgEGFP targeting sites was PCR amplified using primers EGFP F and EGFP R (Figure 65A). The PCR products from cells transfected with different RNP complexes were heated and slowly cooled to room temperature for re-hybridisation. These reannealed products were treated with T7 endonuclease I that catalyses DNA DSBs at DNA mismatches (Figure 65B). If the EGFP gene was successfully edited, the 530 bp product would give two fragments at 230 bp and 300 bp after the T7 endonuclease I treatment. However, no expected bands were observed for all RNP complexes indicating the EGFP gene was not edited.

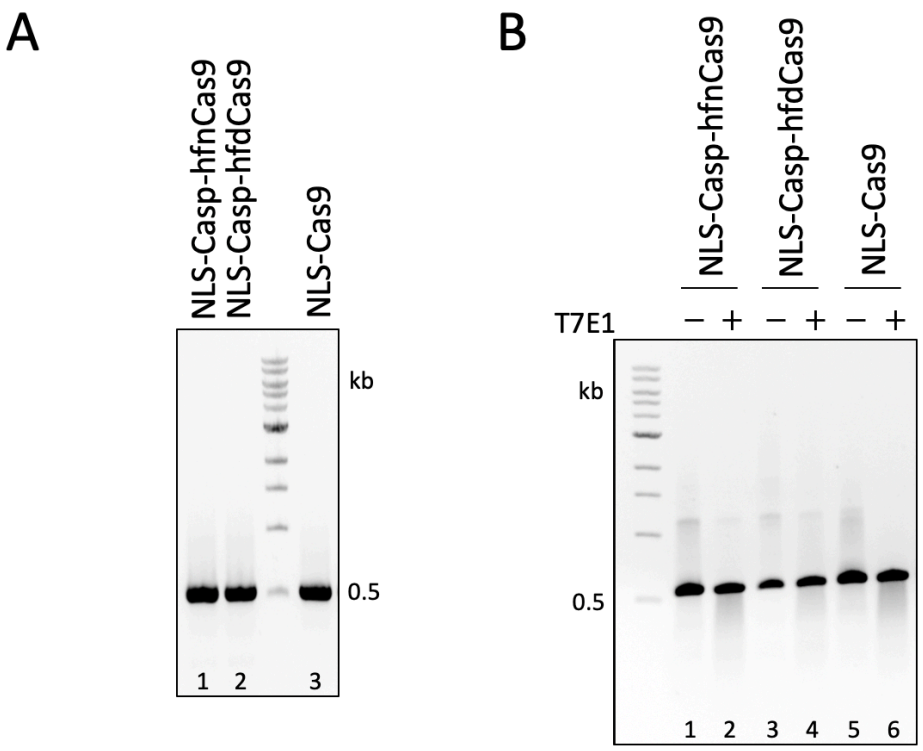


Figure 65. The EGFP gene in pEGFP-c1 plasmid was not edited by NLS-Cas9 or NLS-hf fusion proteins in vivo.

(A) PCR amplification of a 530 bp EGFP DNA fragment from pEGFP-c1 purified from cells transfected with RNP complexes comprising different NLS-proteins and sgEGFP RNA. (B) T7 endonuclease I assays detected no editing of the EGFP gene in pEGFP-c1. PCR products in (A) were reannealed and treated with T7 endonuclease I at 37°C for 30 min to detect DNA mismatches.

4.3. Discussion and conclusion

Precise programmable genome editing has long been sought after by researchers. Although targeted DSBs repaired that are repaired by HDR can introduce desired mutations, it occurs at low efficiency. In the past, targeted DSBs were achieved by ZFNs and TALENs that have limitations (Urnov, 2018). After the discovery of CRISPR-Cas systems, CRISPR-Cas9 was soon employed in genome editing as it offers several advantages over TALENs and ZFNs, such as easy target design and multiplexed editing (Gupta and Musunuru, 2014). Because target recognition in CRISPR systems relies on R-loop formation between RNA and target DNA and not protein-DNA recognition, it

makes gene editing more cost-effective and less time-consuming by just synthesising guide RNA. In addition, Cas9 utilises two nuclease domains, HNH domain and RuvC domain, that cleave the target DNA strand and non-target strand respectively. Introducing mutations of active site residues to one or both nuclease domains has generated nCas9 and dCas9 (Nishimasu et al., 2014). By fusing different effectors to Cas9 mutants that serve as programmable DNA binding domain, fusion proteins could play different roles in genome editing, for example base editors, or in epigenome editing to regulate gene transcription (Gajula, 2019; Komor et al., 2016; Thakore et al., 2016).

In chapter 3, considerable substrate flexibility of *A. boonei* casposase was observed as the enzyme integrated a variety of DNA substrates with or without TIRs. Thus, casposase was fused to Cas9 and Cas9 mutants to test if casposase-catalysed DNA integration could be targeted to specific DNA sites by sgRNAs. If this worked, this would offer potential for genome editing by inserting user-defined DNA sequences into user-defined target sites obviating the need for generating DSBs and subsequent HDR. Indeed, several studies have reported targeted insertion of transposon into specific DNA loci by fusing transposases to DNA binding proteins, such as zinc-finger proteins and dCas9 (Bhatt and Chalmers, 2019; Feng et al., 2010). Recent discoveries also revealed that naturally existing CRISPR-associated Tn7-like transposases catalyses sgRNA-guided DNA transposition (Klompe et al., 2019; Strecker et al., 2019). Here we demonstrate the casposase and Cas9 moieties of a fusion protein were both active. Fusion proteins were capable of integrating DNA substrates without TIRs and forming R-loop with target DNA in the presence of sgRNA. On top of these, it was observed that fusion protein-catalysed DNA integration could be guided by sgRNA to a site 26

bp from the Cas9 PAM. *In vitro* DNA integrations of a 33-nt TK2425 oligo and a 1.2 kb mini-casposon into this sgRNA-guided site were only detected by PCR amplification and were not observed after transforming integrated products into *E. coli*. This might be explained by low transformation efficiency of cells or because the sgRNA-guided full-site integration occurred at very low efficiency.

Random DNA integration was still observed, even though sgRNA was added to the fusion proteins. This off-target effect most likely arose from the casposase moiety due to lack of casposase native target tRNA-Pro gene (Béguin et al., 2016) rather than off-target R-loop formation by the Cas9 moiety. This was confirmed by similar results being observed between Casp-Cas9/dCas9 and Casp-hfCas9/hfdCas9. In addition, a four kb plasmid pACYC-Duet unlikely has multiple sgRNA-Cas9 off-target sites. We reasoned that some of the random DNA integration products were due to 'free' casposase present in the purified fusion protein stocks (Figures 38C and 51). The 210 kDa fusion proteins were susceptible to degradation in *E. coli* during protein synthesis (Wingfield, 2015). To improve this, MBP tag can be fused to current fusion proteins because the naturally occurring MBP tag helps fusion protein solubility and is suggested to protect fusion protein from proteolytic degradation (Peti and Page, 2007). If the yield of fusion proteins was increased, a high-resolution size exclusion chromatography column could be used in the final step of protein purification to remove the 'free' casposase. The rest of the random integration products resulted from the casposase moiety competing with the sgRNA-Cas9 moiety for DNA binding. The target specificity of fusion proteins might be improved by directed evolution that has been used in expanding CRISPR-Cas9 PAM recognition (Chen et al., 2019).

Casp-hfdCas9 and Casp-hfnCas9 might serve as prototypes for further studies because DSBs generated by active Cas9 lead to unpredictable mutations. The reasons *in vivo* integration by fusion proteins was not detected was the same as those for casposase described in chapter 3. The casposase moiety might require other proteins from the casposon for *in vivo* activity. Further fundamental biology studies of casposase and casposon could help the fusion protein design and integration strategies. As a naturally occurring protein, casposase is not very active because this is harmful to the host. Fusion proteins comprising hyperactive casposase might be useful in biotechnology and this might be achievable by directed evolution of casposase.

Chapter 5

5. Testing potential alternatives to Casp-Cas9 for HDR independent DNA integration

5.1. Introduction

The widely used *S. pyogenes* Cas9 is a relatively large protein of 1,366 amino acids and it becomes even larger when fused to an effector. This is problematic because large recombinant proteins are normally poorly expressed with low protein solubility in bacteria (Wingfield, 2015), and in mammalian cells large protein size creates a delivery challenge due to limited packaging capacity of viral gene delivery systems or reduced RNP transfection efficiency (Adli, 2018). Therefore, smaller Cas9 variants and alternative non-Cas9 interference enzymes from natural CRISPR-Cas systems are being scrutinised for applications in genetic editing. One example is Cas12, a protein from CRISPR type V systems that induces HDR after generation of DSBs (Tóth et al., 2016), and although of similar molecular weight to Cas9 (160 kDa) it has a distinct DNA cleavage mechanism that generates staggered DSBs. It was shown that five bp sticky ends generated by Cas12a facilitate precise integration of exogenous DNA into the target genomic locus, thus providing a new method for targeted gene knock-in (Li et al., 2019). Cas12 also offers different targeting capabilities because PAM sequences targeted by Cas12 proteins are predominantly T-rich compared with the G-rich Cas9 PAM (Safari et al., 2019). However, it was revealed that a Cas12a-crRNA complex bound to the complementary target DNA triggers non-specific ssDNA trans-cleavage activity and non-specific dsDNA nicking in trans (Chen et al., 2018; Murugan et al., 2020). This characteristic makes Cas12a less suitable for genome editing than Cas9 but is exploited in nucleic acid detection (Gootenberg et al., 2018).

A Deltaproteobacterial CasX protein (also known as Cas12e) illustrated a unique protein domain and distinct sgRNA folds compared with Cas9 and Cas12a (Liu et al., 2019). CasX protein is less than 1,000 amino acids, generates staggered DSBs and showed minimal trans-cleavage ssDNA activity after target binding. We therefore raised the interesting proposition of fusing casposase to catalytically dead CasX (dCasX) to test for sgRNA-guided DNA integration and to compare with Casp-dCas9.

Precise programmable genome manipulation can be also achieved by targeted recombination catalysed by non-CRISPR enzymes, called site-specific recombinases (SSEs). They catalyse DNA cleavage, strand exchange and DNA re-ligation (Karimova et al., 2016). Because site-specific recombination does not induce cell intrinsic DNA repair pathways, the outcomes of genome manipulation are highly predictable and do not contain undesired mutations. It also offers advantages over targeted nucleases that it results in not only DNA integration and excision but also inversion. However, the use of site-specific recombinase in genome editing has been limited by the presence of rare pseudo-recognition sites of the recombinase in the host genome or pre-introduction of the recognition site into the host genome by homologous recombination (Gaj et al., 2014). To expand the targeting scope of SSEs, researchers have fused hyperactive catalytic domain of serine recombinases to DNA binding protein such as zinc-finger proteins and dCas9 (Akopian et al., 2003; Chaikind et al., 2016; Standage-Beier et al., 2019). Recently, an archaeal integrase belonging to tyrosine recombinases from *Thermococcus nautili* plasmid pTN3 (Int^{pTN3}), was shown to catalyse low sequence specificity recombination akin to homologous recombination, albeit with lower efficiency than site-specific recombination (Cossu et al., 2017) (Figure 66). This inspired us to fuse Int^{pTN3} to dCas9 to test if sgRNA-guided

DNA binding of the dCas9 moiety could increase the efficiency of Int^{pTN3}-catalysed recombination between two homologous sequences.

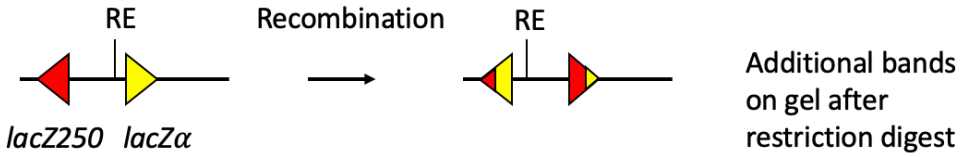


Figure 66. Int^{pTN3}-catalysed low sequence recombination described in previous work. Int^{pTN3} catalysed inversion of a linear DNA substrate between two *lacZα* segments, one full-length and one 250 bp, in the opposite orientation. Inversion of the DNA substrate was detected by appearance of new DNA bands on gel after restriction digest. RE represents a restriction enzyme cutting site.

5.2. Results

5.2.1. Casp-dCasX was successfully expressed in *E. coli* cells but was not purified well.

To construct a casposase-dCasX (Casp-dCasX) fusion, the *casposase* gene was first inserted into a pNLS-His-StrepII plasmid generated by a BBSRC DTP student (Mr Andrew Cubbon). The vector backbone pNLS-His-StrepII was chosen because the fusion protein generated by this DNA construct is flanked by NLS sequences for protein import into the eukaryotic cell nucleus and flanked by two protein tags that facilitate protein purification (Figure 67A). The *dcasX* gene was then inserted immediately downstream of the *casposase* gene and a $(GGS)_8$ linker sequence was inserted between the *casposase* and *dCasX* genes by Q5 site-directed mutagenesis PCR. The protein was successfully expressed in *E. coli* BL21-AI strain demonstrated by western blotting (Figure 67B). Casp-dCasX was purified by a BBSRC DTP rotation student (Mr Ashley Parkes) as part of his training and protein fractions eluted from a Ni^{2+} column showed most Casp-dCasX did not bind to the column (Figure 67C). However, because

the eluted Casp-dCasX was too diluted, it was not observed on an SDS gel after purification through a heparin column or Strep-tactin® resin.

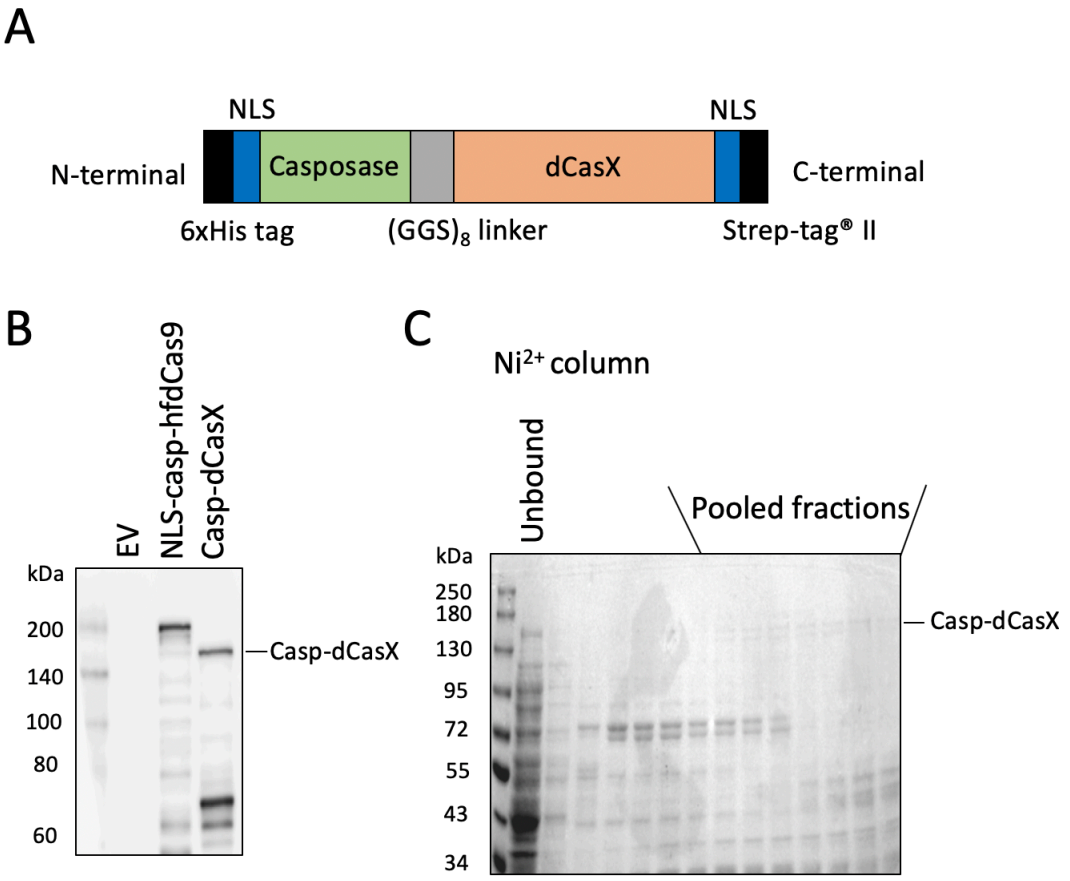


Figure 67. Difficulties in purification of a Casp-dCasX fusion protein
(A) Schematic representation of Casp-dCasX gene construct. (B) Casp-dCasX at 168 kDa was successfully expressed in *E. coli* BL21-AI strain revealed by western blotting. BL21-AI harbouring pdCaspoX plasmid was cultured at 37°C until OD_{600}=0.6. Protein expression was then induced and cultured at 18°C overnight. Cells were lysed and run on SDS gel followed by western blotting using anti-6xHis antibody. BL21-AI harbouring pNLS (empty vector, EV) was used as a negative control and harbouring pNLS-hfdCaspoR expressing the 214 kDa NLS-Casp-hfdCas9 as a positive control. (C) Purification of Casp-dCasX through a Ni^{2+} column.

5.2.2. Purification of Int^{pTN3} and Int^{pTN3}-dCas9.

A codon-optimised *Int^{pTN3}* gene (NCBI gene ID: 17125032) for expression in *E. coli* cells was synthesised by GeneArt® and inserted into pACYC-Duet to generate pCHL11. The plasmid was transformed into BL21-AI for Int^{pTN3} expression and purification. Int^{pTN3} was purified by a summer vacation undergraduate student (Miss Dora Barisic) through a Ni^{2+} column followed by a heparin column (Figure 68).

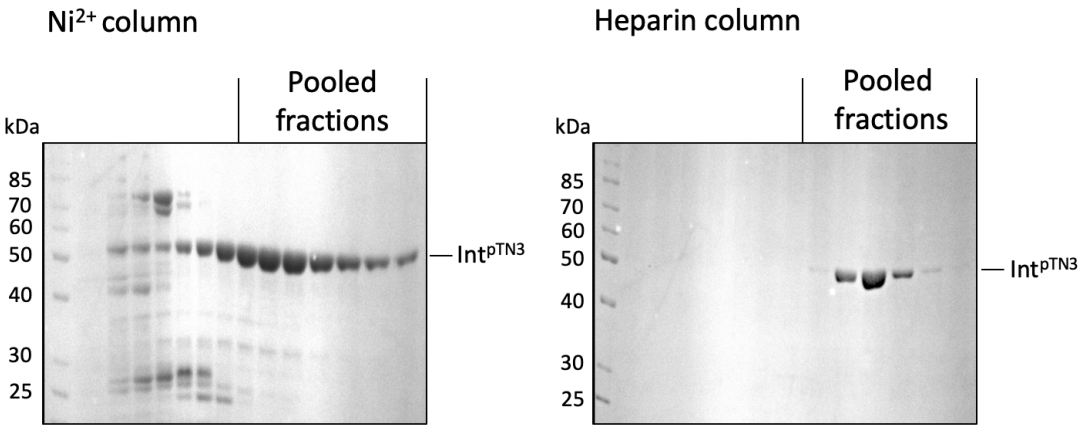


Figure 68. Purification of Int^{pTN3}.
Purification of Int^{pTN3} through a Ni^{2+} column followed by a heparin column. Elution fractions between two black lines were pooled together.

The gene encoding dCas9 was inserted into pCHL11 immediately downstream of the *casposase* gene followed by insertion of a $(GGS)_8$ linker between the two genes and a Strep-tag® II encoding sequence into the 3' end of the fusion gene construct to generate pCHL49 (Figure 69A). The fusion protein Int^{pTN3}-dCas9 endoed by pCHL49 was purified by Dora Barisic through a Ni^{2+} column followed by Strep-Tactin® resin and the identity of purified proteins were confirmed by western blotting (Figure 69B and 69C).

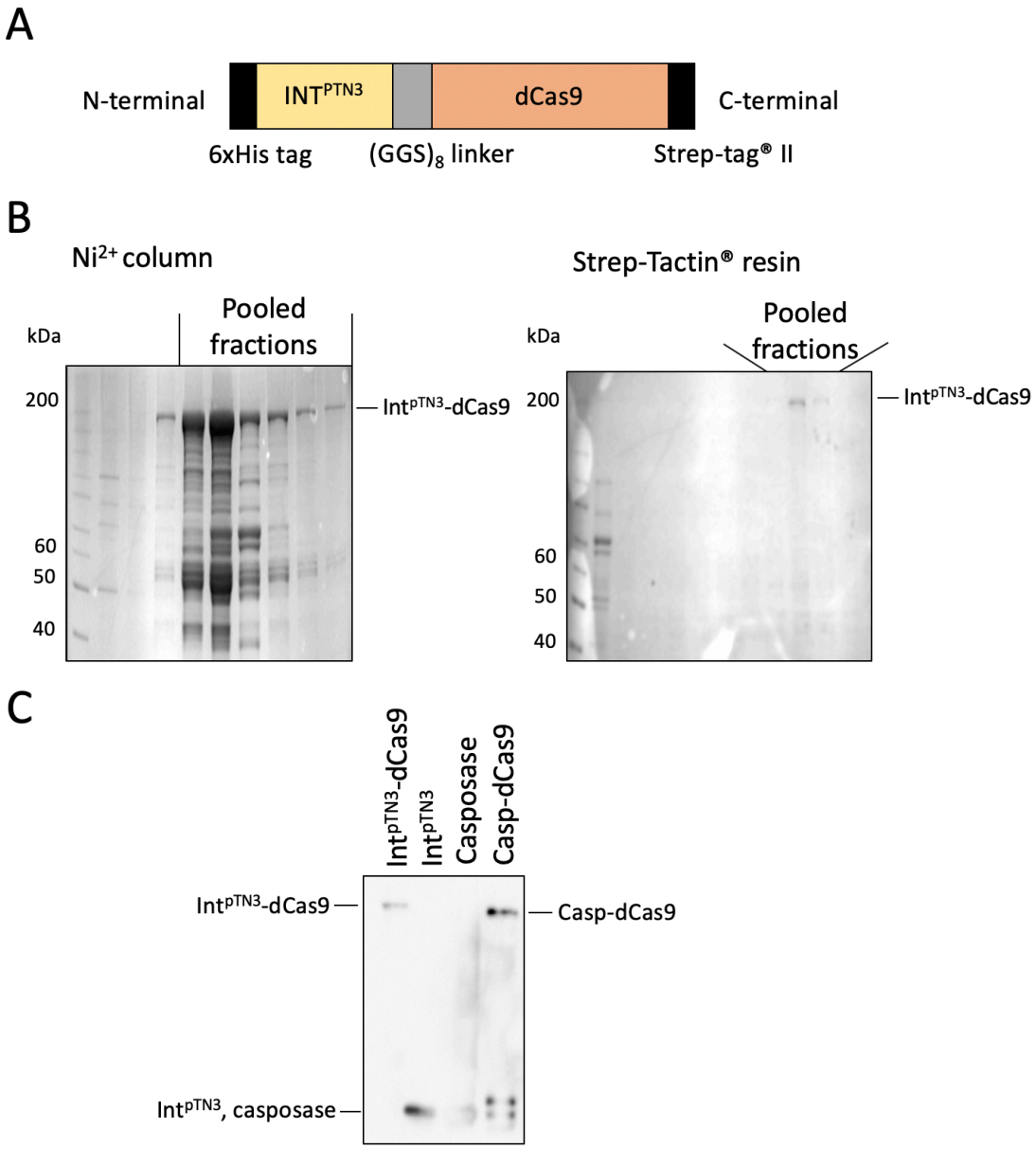


Figure 69. Purification of Int^{pTN3}-dCas9.
(A) Schematic representation of Int^{pTN3}-dCas9 protein construct. (B) Purification of Int^{pTN3}-dCas9 through a Ni^{2+} column followed by Strep-Tactin® resin. Elution fractions between two black lines were pooled together. (C) Western blot using anti-6xHis antibody confirmed the identity of purified Int^{pTN3} and Int^{pTN3}-dCas9. The 215 kDa Int^{pTN3}-dCas9 and 54 kDa Int^{pTN3} were at the right size using 210 kDa Casp-dCas9 and 50 kDa casposase as markers.

5.2.3. Int^{pTN3} activity was not detected in the Int^{pTN3}-dCas9 fusion protein.

Low sequence specificity recombination activity of Int^{pTN3} was assessed by a plasmid containing two *lacZα* segments in the same orientation (pCHL43) or in the opposite

orientation (pCHL44). pCHL43 with both *lacZ*α segments facing in one direction was used to detect deletion catalysed by Int^{pTN3} (Figure 70A). High concentrations of archaeal protein Int^{pTN3} at 500 and 1000 nM were incubated with pCHL43 at 65°C for three and a half hours. We did not observe smaller DNA products corresponding to deletion of pCHL43 through agarose gel electrophoresis (Figure 70B lanes 2-3). However, larger DNA products were observed instead in these reactions and they might be multimers of pCHL43 plasmid resulting from sequential integration. The band intensity of multimers was reduced at 1000 nM of Int^{pTN3} compared with that at 500 nM suggesting self-inhibition due to high concentration of Int^{pTN3}. This was confirmed by there being no detectable DNA multimers observed when using 2500 nM of Int^{pTN3} (Figure 70C).

For reactions using Int^{pTN3}-dCas9 fusion protein, 900 nM sgpUC19 targeting to the two *lacZ*α genes on pCHL43 was first incubated with 300 nM Int^{pTN3}-dCas9 at room temperature for 10 min. Because the Int^{pTN3}-dCas9 fusion protein contains bacterial protein moiety dCas9, which is not active at 65°C, the Int^{pTN3}-dCas9-sgpUC19 RNP complex was incubated with pCHL43 at 37°C overnight. However, analysis by gel electrophoresis of reaction products illustrated no detectable DNA products resulting from neither deletion nor sequential integration of pCHL43 (Figure 70B lane 4).

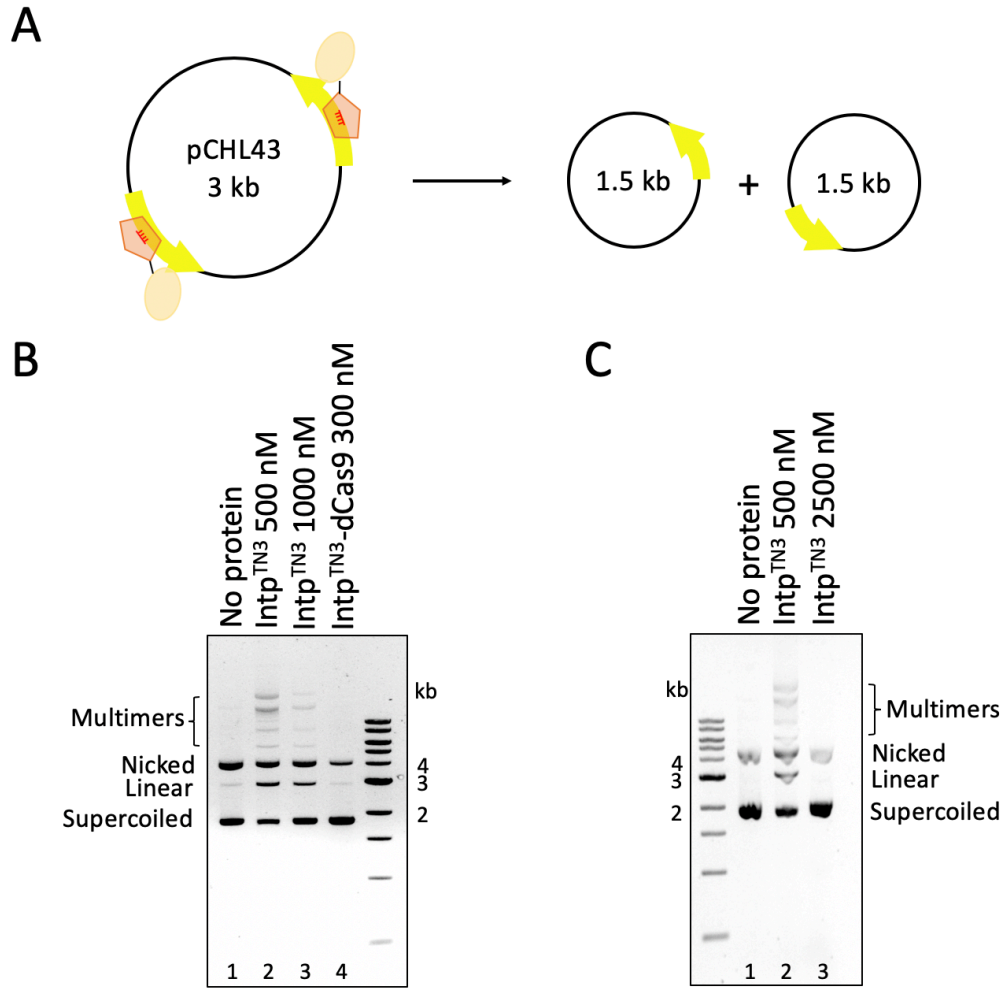


Figure 70. Int^{pTN3} did not catalyse deletion between two lacZα genes in pCHL43.
(A) Schematic diagram of deletion catalysed by Int^{pTN3}-dCas9 between two *lacZα* genes in pCHL43. Yellow arrow represents *lacZα* gene. Red pentagon represents dCas9 moiety and light orange sphere represents Int^{pTN3}. (B) Agarose gel electrophoresis of recombination reactions showed no deletion products but multimers of plasmids were seen in Int^{pTN3}. Int^{pTN3}-dCas9 at 300 nM was first incubated with 900 nM sgpUC19 before incubating with pCHL43 at 37°C overnight. For Int^{pTN3}, reactions were carried out at 65°C for 3.5 hours. (C) Self-inhibition was seen at 2500 nM of Int^{pTN3}.

The plasmid pCHL44 with both *lacZα* segments facing in the opposite direction was used to detect inversion catalysed by Int^{pTN3} similar to previous work (Cossu et al., 2017) (Figure 71A). Int^{pTN3} and Int^{pTN3}-dCas9-sgpUC19 were incubated with pCHL44 as described in above reactions using pCHL43. Reaction products were ethanol

precipitated followed by PCR amplification to detect inversion products. However, the results showed no inversion products in all samples (Figure 71B).

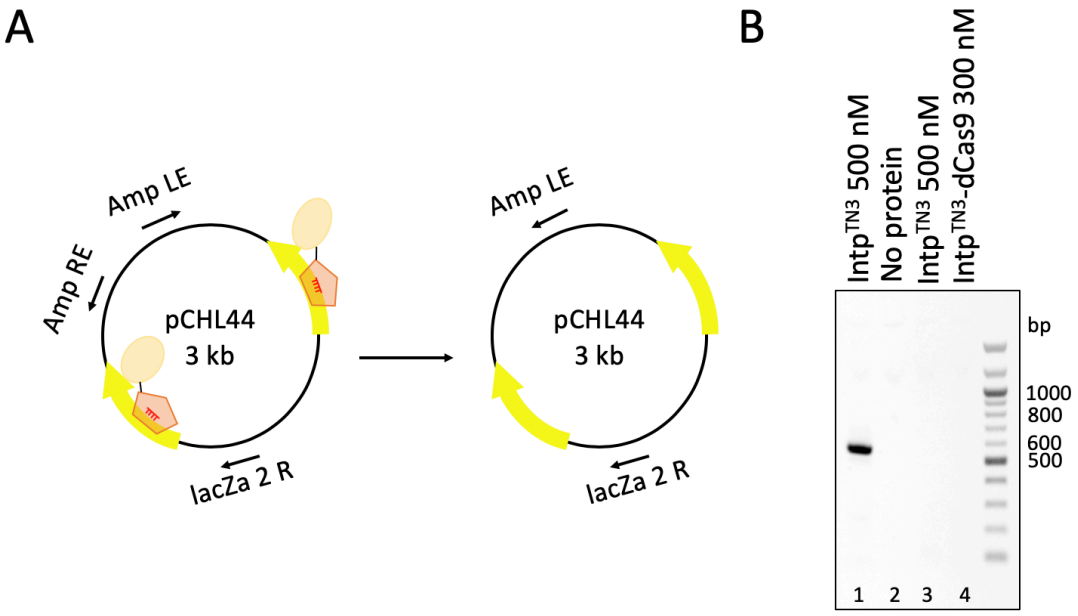


Figure 71. IntpTN3 did not catalyse inversion in pCHL44.
(A) Schematic diagram of inversion catalysed by Int^{pTN3}-dCas9 between two *lacZ*α genes in pCHL44. PCR primers used subsequent PCR amplification were shown in black arrow. Amp RE and lacZa 2 R were used in a positive control. Amp LE and lacZa 2 R were used to detect inversion. (B) Agarose gel electrophoresis of PCR reactions of recombination products showed no inversion products. Lane 1 was a positive control to show PCR reactions worked. Int^{pTN3}-dCas9 at 300 nM was first incubated with 900 nM sgpUC19 before incubating with pCHL44 at 37°C overnight. For Int^{pTN3}, reactions were carried out at 65°C for 3.5 hours.

5.3. Discussion and future work

S. pyogenes Cas9 is widely used in genome editing and biotechnology but Cas9 variants or other Cas single interference proteins may also be effective by showing different targeting scope and smaller protein molecular weight. Here, we fused casposase to a recently characterised catalytically inactive CasX (Cas12e) protein from a CRISPR type V-E system. CasX is composed of just less than 1,000 amino acids and it recognises a PAM sequence of TTCN compared with an NGG PAM for Cas9 (Liu et al., 2019). Although the fusion protein Casp-dCasX was expressed in *E. coli*, its purification

was unsuccessful. The SDS gel of elution fractions from a Ni^{2+} column revealed that most of the Casp-dCasX was found in the unbound fraction (Figure 67C). Formation of inclusion bodies in cells by aggregated Casp-dCasX might account for the inability to bind a Ni^{2+} column. To improve this, a MBP tag facilitating recombinant protein solubility (Peti and Page, 2007) can be fused to Casp-dCasX to help with protein purification.

Targeted genome editing can be also achieved by site-specific recombinases, in addition to targeted nucleases. Attempts have been reported to fuse a catalytic domain of serine recombinases to dCas9 to expand the targeting scope but the catalytic domain still requires a minimal recombinase core sequence near a sgRNA targeting site for binding and function (Chaikind et al., 2016; Standage-Beier et al., 2019). A recently characterised tyrosine recombinase IntpTN3 showed catalysing inversion between two *lacZ*α segments without the need for attachment sites mimicking homologous recombination (Cossu et al., 2017). The efficiency of this low sequence specificity recombination is much lower than site-specific recombination between native attachment sites. Thus, we tethered Int^{pTN3} to dCas9 to see if the efficiency of low sequence specificity recombination could be increased by sgRNA targeting. Our results showed the purified Int^{pTN3}-dCas9 fusion protein might be inactive because no multimers of plasmid were observed after incubating with pCHL43 compared with reactions using wild-type Int^{pTN3}. For Int^{pTN3}, no inversions or deletions were detected between two identical *lacZ*α genes on the same plasmid but multimers of plasmid were observed after running recombination products on gels. The plasmid multimers were probably produced by sequential integration by Int^{pTN3} that occurred outside of the two *lacZ*α genes. The same observation was also noted in previous work

when incubating Int^{pTN3} with another plasmid pBR322 (Cossu et al., 2017). Although Int^{pTN3} is able to catalyse low sequence specificity recombination, the protein might still have a preference binding site or sequence motif that needs further analysis. A hyperactive mutant of Int^{pTN3} might be more suitable than the wild-type in biotechnology and this could be generated by directed evolution.

Chapter 6

6. Final discussion

Recent studies suggested the CRISPR repeats and the adaptation machinery originated from ancestral casposons (Hickman and Dyda, 2015; Krupovic et al., 2014, 2017). As casposons encoded casposases are CRISPR Cas1 homologues, it was suggested that casposases are more closely related to the ancestral state of Cas1. Here, we biochemically characterised *A. boonei* casposase and tried to elucidate the evolutionary link between casposases and Cas1. Our results showed *A. boonei* casposase did not only integrate ssDNA and dsDNA oligos derived from the casposon TIR as shown in a recent study using the same casposase (Béguin et al., 2019), but it also integrated dsDNA oligos with different sequences and different lengths. Our results were partly consistent with another study that *M. mazei* casposase specifically integrated random ssDNA and dsDNA into a plasmid containing the native target site of the *M. mazei* casposase (Hickman et al., 2020). The DNA integration activity was abolished in the absence of *M. mazei* casposase target site, whereas *A. boonei* casposase can integrate DNA into random sites in the absence of the target site. Therefore, *M. mazei* casposase is more site-specific than that of *A. boonei*. Our results support the idea that ancestral casposase gave rise to CRISPR Cas1 because the promiscuous DNA substrate integration activity of the ancestral casposase has passed onto the CRISPR Cas1. However, during the evolutionary transformation, there were changes in protein-protein interactions that fixed the integrating DNA length making Cas1 specialise in short spacer integration. The recently discovered type V-C and V-D adaptation modules lack Cas2 in these systems and type V-C and V-D Cas1s serve as

an evolutionary intermediate between casposase and Cas1. A type V-C Cas1 was shown active in spacer integration as a homotetramer *in vitro* (Wright et al., 2019). The integrating spacer length of type V-C and V-D Cas1s is shorter than other CRISPR adaptation complexes that contain Cas2. As Cas1 further evolved, there were changes in protein-protein interactions that favoured Cas2 binding over tetramerization (Hickman et al., 2020). This Cas2 recruitment has increased the integrating DNA length. Casposons are a novel type of transposons and the integrase they encode, casposase, functions as transposases (Hickman and Dyda, 2015; Krupovic et al., 2014). Our results also showed *A. boonei* casposase is able to integrate long DNA substrates longer than 1 kb into a plasmid lacking a target site. These long DNA substrates can be flanked with casposon TIRs or without, although the integration efficiency is much lower for those without TIRs. This promiscuous substrate capturing is unusual for a transposase because transposable elements are selfish and TIRs flanking the transposable elements allow specific binding of transposases (Ichikawa et al., 1987; Szabó et al., 2010). Since the integration of long dsDNA without TIRs flanking was carried out *in vitro* and DNA substrates was provided in a linear form, *in vivo* characterisation is required to answer the role of this activity in physiological conditions.

Targeted DNA insertion can be achieved not only by the HDR-mediated pathway or site-specific recombinase, it can be also achieved by targeted transposition (Urnov, 2018). This involves fusing a transposase to a DNA binding protein such as zinc-finger protein or CRISPR interference effector protein (Bhatt and Chalmers, 2019; Feng et al., 2010). This project investigated the potential of using casposase to achieve targeted DNA insertion by fusing it to Cas9. Our results showed the Cas9 component in the fusion proteins worked perfectly that it is able to specifically target plasmid and

chromosomal DNA in the presence of sgRNA. Although the casposase component in the fusion proteins is able to integrate DNA substrates at a fixed distance to the sgRNA targeting site of the Cas9 component, we could not prove the targeted integration was full-site integration. Another limitation of these fusion proteins was the DNA integration into random sites even in the presence of sgRNA. Therefore, *A. boonei* casposase may not be the suitable fusing component for the sgRNA-guided DNA insertion. Recent studies revealed there are natural occurring Tn7-like transposon systems exploiting CRISPR Cas proteins for targeted transposition (Peters et al., 2017). The CRISPR crRNA-effector protein complex in these Tn7-like transposons only binds to target DNA without inducing target degradation and it directly binds TniQ that recruits TnsABC transposition protein complex. The characterised Tn7-like transposons utilise either a CRISPR type I Cascade variant complex or a CRISPR type V-K effector protein, Cas12k (Klompe et al., 2019; Strecker et al., 2019). The two independent studies showed these CRISPR-associated Tn7-like transposons can achieve around 50% insertion frequency and with more than 50% on-target specificity. These Tn7-like transposons may serve a better targeted DNA insertion tool that does not require homology-directed repair.

7. References

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

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. Nat Commun *9*, 1911.

Ahn, S., and Joyner, A.L. (2004). Dynamic Changes in the Response of Cells to Positive Hedgehog Signaling during Mouse Limb Patterning. Cell *118*, 505–516.

Aird, E.J., Lovendahl, K.N., St. Martin, A., Harris, R.S., and Gordon, W.R. (2018). Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. Commun Biol *1*, 54.

Akopian, A., He, J., Boocock, M.R., and Stark, W.M. (2003). Chimeric recombinases with designed DNA sequence recognition. Proceedings of the National Academy of Sciences *100*, 8688–8691.

Amitai, G., and Sorek, R. (2016). CRISPR–Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol *14*, 67–76.

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.

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.

Béguin, P., Charpin, N., Koonin, E.V., Forterre, P., and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res gkw821.

Béguin, P., Chekli, Y., Sezonov, G., Forterre, P., and Krupovic, M. (2019). Sequence motifs recognized by the casposon integrase of Aciduliprofundum boonei. Nucleic Acids Research *47*, 6386–6395.

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

Bhatt, S., and Chalmers, R. (2019). Targeted DNA transposition in vitro using a dCas9-transposase fusion protein. Nucleic Acids Research *47*, 8126–8135.

Bikard, D., and Marraffini, L.A. (2012). Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. Current Opinion in Immunology *24*, 15–20.

Bischof, H., Burgstaller, S., Waldeck-Weiermair, M., Rauter, T., Schinagl, M., Ramadani-Muja, J., Graier, W.F., and Malli, R. (2019). Live-Cell Imaging of Physiologically Relevant Metal Ions Using Genetically Encoded FRET-Based Probes. Cells *8*, 492.

Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proceedings of the National Academy of Sciences *95*, 10570–10575.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science *326*, 1509–1512.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol *19*, 199.

Chaikind, B., Bessen, J.L., Thompson, D.B., Hu, J.H., and Liu, D.R. (2016). A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. Nucleic Acids Res gkw707.

Chandler, M. (2017). Prokaryotic DNA Transposons: Classes and Mechanism. In ELS, John Wiley & Sons Ltd, ed. (Chichester, UK: John Wiley & Sons, Ltd), pp. 1–16.

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.

Charpentier, M., Khedher, A.H.Y., Menoret, S., Brion, A., Lamribet, K., Dardillac, E., Boix, C., Perrouault, L., Tesson, L., Geny, S., et al. (2018). CtIP fusion to Cas9 enhances transgene integration by homology-dependent repair. Nat Commun *9*, 1133.

Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. Nat Methods *12*, 326–328.

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.

Chen, W., Zhang, H., Zhang, Y., Wang, Y., Gan, J., and Ji, Q. (2019). Molecular basis for the PAM expansion and fidelity enhancement of an evolved Cas9 nuclease. PLoS Biol *17*, e3000496.

Choulika, A., Perrin, A., Dujon, B., and Nicolas, J.F. (1995). Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol. Cell. Biol. *15*, 1968–1973.

Chu, V.T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol *33*, 543–548.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Research *42*, 6091–6105.

Collins, F.S., Rossant, J., and Wurst, W. (2007). A Mouse for All Reasons. Cell *128*, 9–13.

Cong, L., Zhou, R., Kuo, Y., Cunniff, M., and Zhang, F. (2012). Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. Nat Commun *3*, 968.

Cossu, M., Badel, C., Catchpole, R., Gadelle, D., Marguet, E., Barbe, V., Forterre, P., and Oberto, J. (2017). Flipping chromosomes in deep-sea archaea. PLoS Genet *13*, e1006847.

Cubbon, A., Ivancic-Bace, I., and Bolt, E.L. (2018). CRISPR-Cas immunity, DNA repair and genome stability. Bioscience Reports *38*, BSR20180457.

Cui, T.J., and Joo, C. (2019). Facilitated diffusion of Argonaute-mediated target search. RNA Biology *16*, 1093–1107.

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

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., and Yan, N. (2012). Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. Science *335*, 720–723.

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

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat Biotechnol *32*, 1262–1267.

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

Donoho, G., Jasin, M., and Berg, P. (1998). Analysis of Gene Targeting and Intrachromosomal Homologous Recombination Stimulated by Genomic Double-Strand Breaks in Mouse Embryonic Stem Cells. Mol. Cell. Biol. *18*, 4070–4078.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H.D., 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.

Faure, G., Makarova, K.S., and Koonin, E.V. (2019). CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. Journal of Molecular Biology *431*, 3–20.

Feng, X., Bednarz, A.L., and Colloms, S.D. (2010). Precise targeted integration by a chimaeric transposase zinc-finger fusion protein. Nucleic Acids Research *38*, 1204–1216.

Ferreira, I., Amarante, T.D., and Weber, G. (2015). DNA terminal base pairs have weaker hydrogen bonds especially for AT under low salt concentration. The Journal of Chemical Physics *143*, 175101.

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.

Gaj, T., Sirk, S.J., and Barbas, C.F. (2014). Expanding the scope of site-specific recombinases for genetic and metabolic engineering: Expanding the Scope of Site-Specific Recombinases. Biotechnol. Bioeng. *111*, 1–15.

Gajula, K.S. (2019). Designing an Elusive C•G→G•C CRISPR Base Editor. Trends in Biochemical Sciences *44*, 91–94.

Garton, M., Najafabadi, H.S., Schmitges, F.W., Radovani, E., Hughes, T.R., and Kim, P.M. (2015). A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. Nucleic Acids Res *43*, 9147–9157.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471.

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

Giaever, G., and Nislow, C. (2014). The Yeast Deletion Collection: A Decade of Functional Genomics. Genetics *197*, 451–465.

Glaser, A., McColl, B., and Vadolas, J. (2016). GFP to BFP Conversion: A Versatile Assay for the Quantification of CRISPR/Cas9-mediated Genome Editing. Molecular Therapy - Nucleic Acids *5*, e334.

Globyte, V., Lee, S.H., Bae, T., Kim, J., and Joo, C. (2019). CRISPR /Cas9 searches for a protospacer adjacent motif by lateral diffusion. EMBO J *38*.

Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science *360*, 439–444.

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

Grindley, N.D.F., Whiteson, K.L., and Rice, P.A. (2006). Mechanisms of Site-Specific Recombination. Annu. Rev. Biochem. *75*, 567–605.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat Biotechnol *32*, 577–582.

Guirouilh-Barbat, J., Lambert, S., Bertrand, P., and Lopez, B.S. (2014). Is homologous recombination really an error-free process? Front. Genet. *5*.

Gupta, R.M., and Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. J. Clin. Invest. *124*, 4154–4161.

Gurumurthy, C.B., Grati, M., Ohtsuka, M., Schilit, S.L.P., Quadros, R.M., and Liu, X.Z. (2016). CRISPR: a versatile tool for both forward and reverse genetics research. Hum Genet *135*, 971–976.

Halford, S.E. (2004). How do site-specific DNA-binding proteins find their targets? Nucleic Acids Research *32*, 3040–3052.

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

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 *11*, 208.

Hickman, A.B., and Dyda, F. (2014). CRISPR-Cas immunity and mobile DNA: a new superfamily of DNA transposons encoding a Cas1 endonuclease. Mobile DNA *5*, 23.

Hickman, A.B., and Dyda, F. (2015). The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. Nucleic Acids Res *43*, 10576–10587.

Hickman, A.B., Kailasan, S., Genzor, P., Haase, A.D., and Dyda, F. (2020). Casposase structure and the mechanistic link between DNA transposition and spacer acquisition by CRISPR-Cas. ELife *9*, e50004.

Horii, T., and Hatada, I. (2016). Challenges to increasing targeting efficiency in genome engineering. Journal of Reproduction and Development *62*, 7–9.

Hu, J.H., Miller, S.M., Geurts, M.H., Tang, W., Chen, L., Sun, N., Zeina, C.M., Gao, X., Rees, H.A., Lin, Z., et al. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature *556*, 57–63.

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

Ichikawa, H., Ikeda, K., Wishart, W.L., and Ohtsubo, E. (1987). Specific binding of transposase to terminal inverted repeats of transposable element Tn3. Proceedings of the National Academy of Sciences *84*, 8220–8224.

Ivančić-Baće, I., Cass, S.D., Wearne, S.J., and Bolt, E.L. (2015). Different genome stability proteins underpin primed and naïve adaptation in *E. coli* CRISPR-Cas immunity. Nucleic Acids Res *43*, 10821–10830.

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

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

Jansen, Ruud., Embden, Jan.D.A. van, Gaastra, Wim., and Schouls, Leo.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol *43*, 1565–1575.

Jeon, Y., Choi, Y.H., Jang, Y., Yu, J., Goo, J., Lee, G., Jeong, Y.K., Lee, S.H., Kim, I.-S., Kim, J.-S., et al. (2018). Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. Nat Commun *9*, 2777.

Jiang, F., and Doudna, J.A. (2017). CRISPR–Cas9 Structures and Mechanisms. Annu. Rev. Biophys. *46*, 505–529.

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

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

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol *31*, 233–239.

Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J., and Yang, S. (2015b). Multigene Editing in the Escherichia coli Genome via the CRISPR-Cas9 System. Appl. Environ. Microbiol. *81*, 2506–2514.

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.

Jung, T.-Y., An, Y., Park, K.-H., Lee, M.-H., Oh, B.-H., and Woo, E. (2015). Crystal Structure of the Csm1 Subunit of the Csm Complex and Its Single-Stranded DNA-Specific Nuclease Activity. Structure *23*, 782–790.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat Protoc *7*, 1511–1522.

Kapitonov, V.V., Makarova, K.S., and Koonin, E.V. (2016). ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. J. Bacteriol. *198*, 797–807.

Karimova, M., Splith, V., Karpinski, J., Pisabarro, M.T., and Buchholz, F. (2016). Discovery of Nigri/nox and Panto/pox site-specific recombinase systems facilitates advanced genome engineering. Sci Rep *6*, 30130.

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.

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.

Kim, H.S., Jeong, Y.K., Hur, J.K., Kim, J.-S., and Bae, S. (2019a). Adenine base editors catalyze cytosine conversions in human cells. Nat Biotechnol *37*, 1145–1148.

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

Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. Nature *579*, 141–145.

Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proceedings of the National Academy of Sciences *93*, 1156–1160.

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

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.

Koonin, E.V. (2017). Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. Biol Direct *12*, 5.

Koonin, E.V., and Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet *16*, 184–192.

Koonin, E.V., and Makarova, K.S. (2019). Origins and evolution of CRISPR-Cas systems. Phil. Trans. R. Soc. B *374*, 20180087.

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M., and Yanagawa, H. (2009). Six Classes of Nuclear Localization Signals Specific to Different Binding Grooves of Importin α. J. Biol. Chem. *284*, 478–485.

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol *12*, 36.

Krupovic, M., Shmakov, S., Makarova, K.S., Forterre, P., and Koonin, E.V. (2016). Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. Genome Biol Evol *8*, 375–386.

Krupovic, M., Béguin, P., and Koonin, E.V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Current Opinion in Microbiology *38*, 36–43.

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

Lagace, D.C., Whitman, M.C., Noonan, M.A., Ables, J.L., DeCarolis, N.A., Arguello, A.A., Donovan, M.H., Fischer, S.J., Farnbauch, L.A., Beech, R.D., et al. (2007). Dynamic Contribution of Nestin-Expressing Stem Cells to Adult Neurogenesis. Journal of Neuroscience *27*, 12623–12629.

Lander, E.S. (2016). The Heroes of CRISPR. Cell *164*, 18–28.

Lau, C.H., Reeves, R., and Bolt, E.L. (2019). Adaptation processes that build CRISPR immunity: creative destruction, updated. Essays in Biochemistry *63*, 227–235.

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.e5.

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

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

Li, L., Wv, L.P., and Chandrasegaran, S. (1992). Functional domains in Fok I restriction endonuclease. Proc. Natl. Acad. Sci. USA 5.

Li, P., Zhang, L., Li, Z., Xu, C., Du, X., and Wu, S. (2019). Cas12a mediates efficient and precise endogenous gene tagging via MITI: microhomology-dependent targeted integrations. Cell. Mol. Life Sci.

Li, W., Kamtekar, S., Xiong, Y., Sarkis, G.J., Grindley, N.D.F., and Steitz, T.A. (2005). Structure of a Synaptic gd Resolvase Tetramer Covalently Linked to Two Cleaved DNAs. *309*, 7.

Liu, J.-J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019). CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature *566*, 218–223.

Loeff, L., Brouns, S.J.J., and Joo, C. (2018). Repetitive DNA Reeling by the Cascade-Cas3 Complex in Nucleotide Unwinding Steps. Molecular Cell *70*, 385-394.e3.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). The basic building blocks and evolution of CRISPR–Cas systems. Biochemical Society Transactions *41*, 1392–1400.

Makarova, K.S., Krupovic, M., and Koonin, E.V. (2014). Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. Front. Microbiol. *5*.

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

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

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

Mansour, S.L., Thomas, K.R., and Capecchi, M.R. (1988). Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. Nature *336*, 348–352.

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

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

Meeske, A.J., Nakandakari-Higa, S., and Marraffini, L.A. (2019). Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. Nature *570*, 241–245.

Miller, J.H. (1972). Experiments in Molecular Genetics (Cold Spring Harbor Laboratory).

Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. The EMBO Journal *4*, 1609–1614.

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.

Mojica, F.J.M., D□ez-Villase□or, C., Garc□a-Mart□nez, J., and Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J Mol Evol *60*, 174–182.

Moscou, M.J., and Bogdanove, A.J. (2009). A Simple Cipher Governs DNA Recognition by TAL Effectors. Science *326*, 1501–1501.

Munoz-Lopez, M., and Garcia-Perez, J. (2010). DNA Transposons: Nature and Applications in Genomics. CG *11*, 115–128.

Murugan, K., Seetharam, A.S., Severin, A.J., and Sashital, D.G. (2020). CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects. J. Biol. Chem. *295*, 5538–5553.

Nami, F., Basiri, M., Satarian, L., Curtiss, C., Baharvand, H., and Verfaillie, C. (2018). Strategies for In Vivo Genome Editing in Nondividing Cells. Trends in Biotechnology *36*, 770–786.

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.

Nishimasu, H., and Nureki, O. (2017). Structures and mechanisms of CRISPR RNA-guided effector nucleases. Current Opinion in Structural Biology *43*, 68–78.

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.

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

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

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

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

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

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

Olorunniji, F.J., Rosser, S.J., and Stark, W.M. (2016). Site-specific recombinases: molecular machines for the Genetic Revolution. Biochemical Journal *473*, 673–684.

Pal, A., and Levy, Y. (2019). Structure, stability and specificity of the binding of ssDNA and ssRNA with proteins. PLoS Comput Biol *15*, e1006768.

Pavletich, N., and Pabo, C. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809–817.

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

Peti, W., and Page, R. (2007). Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. Protein Expression and Purification *51*, 1–10.

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.

Ramachandran, A., Summerville, L., Learn, B.A., DeBell, L., and Bailey, S. (2020). Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. J. Biol. Chem. *295*, 3403–3414.

Ramakrishnan, M., Zhou, M., Pan, C., Hänninen, H., Yrjälä, K., Vinod, K.K., and Tang, D. (2019). Affinities of Terminal Inverted Repeats to DNA Binding Domain of Transposase Affect the Transposition Activity of Bamboo Ppmar2 Mariner-Like Element. IJMS *20*, 3692.

Rocha-Martins, M., Cavalheiro, G.R., Matos-Rodrigues, G.E., and Martins, R.A.P. (2015). From Gene Targeting to Genome Editing: Transgenic animals applications and beyond. An. Acad. Bras. Ciênc. *87*, 1323–1348.

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

Römer, P., Hahn, S., Jordan, T., Strauß, T., Bonas, U., and Lahaye, T. (2007). Plant Pathogen Recognition Mediated by Promoter Activation of the Pepper *Bs3* Resistance Gene. Science *318*, 645–648.

Rostøl, J.T., and Marraffini, L.A. (2019). Non-specific degradation of transcripts promotes plasmid clearance during type III-A CRISPR–Cas immunity. Nat Microbiol *4*, 656–662.

Rouet, P., Smih, F., and Jasin, M. (1994). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol. Cell. Biol. *14*, 8096–8106.

Safari, F., Zare, K., Negahdaripour, M., Barekati-Mowahed, M., and Ghasemi, Y. (2019). CRISPR Cpf1 proteins: structure, function and implications for genome editing. Cell Biosci *9*, 36.

San-Miguel, T., Pérez-Bermúdez, P., and Gavidia, I. (2013). Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. SpringerPlus *2*, 89.

Scherer, S., and Davis, R.W. (1979). Replacement of chromosome segments with altered DNA sequences constructed in vitro. Proceedings of the National Academy of Sciences *76*, 4951–4955.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.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.

Shapiro, J., Iancu, O., Jacobi, A.M., McNeill, M.S., Turk, R., Rettig, G.R., Amit, I., Tovin-Recht, A., Yakhini, Z., Behlke, M.A., et al. (2020). Increasing CRISPR Efficiency and Measuring Its Specificity in HSPCs Using a Clinically Relevant System. Molecular Therapy - Methods & Clinical Development *17*, 1097–1107.

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.e6.

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

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

Short, F.L., Akusobi, C., Broadhurst, W.R., and Salmond, G.P.C. (2018). The bacterial Type III toxin-antitoxin system, ToxIN, is a dynamic protein-RNA complex with stability-dependent antiviral abortive infection activity. Sci Rep *8*, 1013.

Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., and Paques, F. (2011). Meganucleases and Other Tools for Targeted Genome Engineering: Perspectives and Challenges for Gene Therapy. CGT *11*, 11–27.

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: Cas3 nuclease/helicase. The EMBO Journal *30*, 1335–1342.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Song, J., Yang, D., Xu, J., Zhu, T., Chen, Y.E., and Zhang, J. (2016). RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun *7*, 10548.

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

Standage-Beier, K., Brookhouser, N., Balachandran, P., Zhang, Q., Brafman, D.A., and Wang, X. (2019). RNA-Guided Recombinase-Cas9 Fusion Targets Genomic DNA Deletion and Integration. The CRISPR Journal *2*, 209–222.

Stella, S., Alcón, P., and Montoya, G. (2017). Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. Nature *546*, 559–563.

Stella, S., Mesa, P., Thomsen, J., Paul, B., Alcón, P., Jensen, S.B., Saligram, B., Moses, M.E., Hatzakis, N.S., and Montoya, G. (2018). Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. Cell *175*, 1856-1871.e21.

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

Studier, F.W. (2005). Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification *41*, 207–234.

Sundaresan, R., Parameshwaran, H.P., Yogesha, S.D., Keilbarth, M.W., and Rajan, R. (2018). RNA-Independent DNA Cleavage Activities of Cas9 and Cas12a. 29.

Susanne Gräslund, Pär Nordlund, Johan Weigelt, B Martin Hallberg, James Bray, Opher Gileadi, and Stefan Knapp (2008). Protein production and purification. Nat Methods *5*, 135–146.

Swarts, D.C., and Jinek, M. (2018). Mechanistic Insights into the *Cis*- and *Trans* -acting Deoxyribonuclease Activities of Cas12a (Biochemistry).

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Molecular Cell *66*, 221-233.e4.

Szabó, M., Kiss, J., and Olasz, F. (2010). Functional Organization of the Inverted Repeats of IS30. JB *192*, 3414–3423.

Thakore, P.I., Black, J.B., Hilton, I.B., and Gersbach, C.A. (2016). Editing the epigenome: technologies for programmable transcription and epigenetic modulation. Nat Methods *13*, 127–137.

Thomas, K.R., and Capecchi, M.R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell *51*, 503–512.

Thomason, L.C., Costantino, N., and Court, D.L. (2007). E. coli Genome Manipulation by P1 Transduction. In Current Protocols in Molecular Biology, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), p. 1.17.1-1.17.8.

Tóth, E., Weinhardt, N., Bencsura, P., Huszár, K., Kulcsár, P.I., Tálas, A., Fodor, E., and Welker, E. (2016). Cpf1 nucleases demonstrate robust activity to induce DNA modification by exploiting homology directed repair pathways in mammalian cells. Biol Direct *11*, 46.

Ullmann, A. (1992). Roots: Complementation in β-galactosidase: From protein structure to genetic engineering. Bioessays *14*, 201–205.

Urnov, F.D. (2018). Genome Editing B.C. (Before CRISPR): Lasting Lessons from the "Old Testament." The CRISPR Journal *1*, 34–46.

Varga, I., Szebeni, Á., Szoboszlai, N., and Kovács, B. (2005). Determination of trace elements in human liver biopsy samples by ICP–MS and TXRF: hepatic steatosis and nickel accumulation. Anal Bioanal Chem *383*, 476–482.

Wakefield, N., Rajan, R., and Sontheimer, E.J. (2015). Primary processing of CRISPR RNA by the endonuclease Cas6 in *Staphylococcus epidermidis*. FEBS Letters *589*, 3197–3204.

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.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. Structure *17*, 904–912.

Wiktor, J., van der Does, M., Büller, L., Sherratt, D.J., and Dekker, C. (2018). Direct observation of end resection by RecBCD during double-stranded DNA break repair in vivo. Nucleic Acids Research *46*, 1821–1833.

Wilkinson, M., Troman, L., Wan Nur Ismah, W.A., Chaban, Y., Avison, M.B., Dillingham, M.S., and Wigley, D.B. (2016). Structural basis for the inhibition of RecBCD by Gam and its synergistic antibacterial effect with quinolones. ELife *5*, e22963.

Wingfield, P.T. (2015). Overview of the Purification of Recombinant Proteins. Current Protocols in Protein Science *80*.

Wright, A. V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.

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

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

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.

Xu, X., Duan, D., and Chen, S.-J. (2017). CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. Sci Rep *7*, 143.

Xue, C., Zhu, Y., Zhang, X., Shin, Y.-K., and Sashital, D.G. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Reports *21*, 3717–3727.

Yamano, T., Zetsche, B., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Structural Basis for the Canonical and Non-canonical PAM Recognition by CRISPR-Cpf1. Molecular Cell *67*, 633-645.e3.

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.e12.

Zhang, X.-H., Tee, L.Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H. (2015). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Molecular Therapy - Nucleic Acids *4*, e264.

Zhang, Y., Buchholz, F., Muyrers, J.P.P., and Stewart, A.F. (1998). A new logic for DNA engineering using recombination in Escherichia coli. Nat Genet *20*, 123–128.

8. Appendix 1

Primers used in this study

Primer name	Sequence from 5’ to 3’
Casp pDuet F	GATGGATCCTATGAACCCTCTTTTAGTTAGTGG
Casp pDuet R	GCCCTGCAGCTATTTTAATTTACTCTTTACCTTCCC
Casp H242A F	TGGTTTTTTGGCAGAACTAGCTTCCTCTAAGAC
Casp H242A R	ATCGATGGGTCTAGACCTAC
Casp D254A F	TCTTGTATATGCACTTCAAGAGCTTTTTAGG
Casp D254A R	GGGGTCTTAGAGGAAGCT
Cas9 pDuet F	GGCCGGCTGCAGATGGATAAGAAATACTCAATAGGCTTAG
Cas9 pDuet R	TAAATGCGGCCGCTCAGTCACCTCCTAGCTGA
Assem Casp F	CGGCTGCAGATGAACCCTCTTTTAGTTAGTGG
Assem Casp R	GGAGCCTCCCGAGCCACCTTTTAATTTACTCTTTACCTTCCC
Assem Cas9 F	AAGGTGGCTCGGGAGGCTCCATGGATAAGAAATACTCAATAGGC
Casp-GGS6-Cas9 F	GGAGGTAGCGGTGGGTCAGGCGGTTCGATGGATAAGAAATACTCAATAG
Casp-GGS6-Cas9 R	GGAGCCTCCCGATCCGCCACTGCCACCTTTTAATTTACTCTTTACCTTCC
Casp-GGS8-Cas9 F	GGTGGGTCAGGCGGTTCGGGAGGCAGTGGTGGAAGCATGGATAAGAAATACTCAATAG
Casp-GGS8-Cas9 R	ACTACCACCACTTCCGCCAGACCCTCCAGATCCGCCTTTTAATTTACTCTTTACCTTCC
Casp-XTEN-Cas9 F	AGTGAGTCTGCCACTCCGGAATCGATGGATAAGAAATACTCAATAG
Casp-XTEN-Cas9 F	GGTCCCAGGCGTTTCTGAGCCGGATTTTAATTTACTCTTTACCTTCC
Cas9 H840A F	TGATGTCGATGCCATTGTTCCAC
Cas9 H840A R	TAATCACTTAAACGATTAATATCTAATTC
Cas9 D10A F	ATAGGCTTAGCTATCGGCACAAATAG
Cas9 D10A R	TGAGTATTTCTTATCGCTTCC

Cas9 K848A F	AAGTTTCCTTGCGGACGATTCAATAGACAATAAG
Cas9 K848A R	TGTGGAACAATGTGATCG
Cas9 K1003A F	GAAATATCCAGCGCTTGAATCGGAGTTTG
Cas9 K1003A R	TTAATCAAAGCAGTTCCAAC
Cas9 R1060A F	GATTCGCAAAGCGCCTCTAATCGAAACTAATG
Cas9 R1060A R	TCTCCATTTGCAAGTGTAATTTC
p15A ori R	CGTGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACG
SPIN seq	GCGGAGCCTATGGAAAAACG
Caspamp LE F	GGGATGTATATATATCCCCGATAAGCTTTAATGCGGTAG
Caspamp LE R	CTCTTAAGTTCCCTTTTCAGATGATAAGCTGTCAAACATG
Caspamp RE F	CTCTTAAGTTCCCTTTTTATATCAAAAAGGATCTTCACC
Caspamp RE R	GGGATATATATATATCCCCATCTCATGACCAAAATCC
Caspamp F	GGGGATATATATACATCCCCTCTTAAGTTCCCTTTTCAGATGATAAGC
Caspamp R	GGGGATATATATATATCCCCTCTTAAGTTCCCTTTTTATATCAAAAAGGATCTTCACC
AmpR F	TTCTTAGACGTCAGGTGGCAC
AmpR R	TATTAGACGTCGAGTAAACTTGGTCTGACAGTTACC
LE seq	GATAATACCGCGCCACATAGC
RE seq	ATGGTAAGCCCTCCCGTATC
T7 F	TAATACGACTCACTATAGG
sgRNA R	AAGCACCGACTCGGT
sglacZ1 F	TTGAAACCCAGTTTTAGAGCTAGAAATAGC
sglacZ1 R	TATTGGCTTCACTAGTATTATACCTAGGAC
sglacZ2 F	TGTAAAACGAGTTTTAGAGCTAGAAATAGC
sglacZ1 R	ACGTCGTGACACTAGTATTATACCTAGGAC
sgRNA F	ATTCTACTGCAGCCGAAAAGTGCCACCTGAC
sgRNA R2	TCAAAACTGCAGTCTAGACTCGAGTAAGGATC
pCHL42 cmR F	GGCACTGATGAGGGTGTCAGTGAAGTGCCCGGGCATTTTAGCTTCCTTAGC

pCHL42 cmR R	AACAGGAGGTGCCACTCCCTACCCATCAGCAGATGAGAAAAAAATCACTGGATATAC
pUC19 ori R2	GCACGAGGGAGCTTCCAGGGGGAAACGC
lacI F	CAGGGCCAGGCGGTGAAGGGCAATCAGC
lacZ R	TTTGTCGACTTATTTTTGACACCAGACCAACTGGTAATGG
CaspoR R	TAAAACGGCCGAGTCACCTCCTAGCTGACTCAA
EGFP F	GTCAGATCCGCTAGCGCTAC
EGFP R	GCGGATCTTGAAGTTCACCT
Casp SalI R	TAAAAGTCGACTTTTAATTTACTCTTTACCTTCCC
dCasX SalI F	TAAAAGTCGACGAGAAGCGTATTAACAAGATTCGTAAG
dCasX EagI R	TAAAACGGCCGCAGCATTCGGCTTCCAGACCTC
Casp-GGS8-dCasX F	GGTGGGTCAGGCGGTTCGGGAGGCAGTGGTGGAAGCGAGAAGCGTATTAACAAG
Casp-GGS8-dCasX R	ACTACCACCACTTCCGCCAGACCCTCCAGATCCGCCTTTTAATTTACTCTTTACCTTCC
IntpTN3 F	ATCGTAGGATCCAATGGTTAAAAGCGGTGG
IntpTN3 R	AAACAGCTGCAGTTACAGTTCCAG
IntpTN3-GGS8-dCas9 F	GGTGGTTCAGGTGGCAGCGGAGGGTCAGGAGGCTCTATGGATAAGAAATACTCAATAG
IntpTN3-GGS8-dCas9 R	GCTGCCGCCACTACCTCCGGACCCTCCGGATCCACCCAGTTCCAGAATACCCAG
Strep insert F	TCCACAGTTTGAGAAGTGAGCGGCCGCATAATGC
Strep insert R	TGAGACCAAGAAGAGCCGTCACCTCCTAGCTGACTCAAATC
pUC XhoI insert F	CTCGAGACAGTTACCAATGCTTAATC
pUC XhoI insert R	CAGACCAAGTTTACTCATATATAC
lacZ XhoI F	AAAATGCTCGAGATGACCATGATTACGCCAAGCTTG
lacZ XhoI R	AAAATGCTCGAGCTATGCGGCATCAGAGCAG
lacZa 2 R	CGCTCAGTGGAACGAAAACTC

9. Appendix 2

CLUSTAL multiple sequence alignment by MUSCLE of Cas1s from different CRISPR types and casposases from family 2 casposons. Residues highlighted in yellow are active site residues. A * (asterisk) indicates positions which have a single, fully conserved residue. A : (colon) indicates conservation between groups of strongly similar properties - scoring > 0.5 in the Gonnet PAM 250 matrix. A . (period) indicates conservation between groups of weakly similar properties - scoring =< 0.5 in the Gonnet PAM 250 matrix.

```
Ecoli           ----------MTWLPLNPIPL-----KDRVSMIFL-----QYGQIDVIDGAFVLIDKTG
VC_Cas1         MDQGNQTIENQTINCDQHPDFLWTWKSNKRGSRVSVWLPYFSQAKKIPRSKKWSVAYNGG
Strthe          ----------MAGWRTVVV-----NIHS-KLSY-----KNNHLIFR-------NSYK
Sulsol          ----------MISVRTLVI-----SEYGAYVYV-----KKNMLVIK-------KGDK
Metsp2          ----------MKLLLL-----NGHGINMHV-----DGAKLHIKDGRFSTTEEPQ
Metmaz          ----------MKLLLL-----NGHGINMHV-----DGAKLHIKDGRFSTTEEPQ
Metmet          ----------MKLLLL-----NGHGINMRV-----DNAKLHIKDGRFTTTEEPQ
Metbur          ----------MKLLLL-----NGHGINMRV-----DGAKLHIKDGRFSATEDPQ
Aciboo          ----------MNPLLV-----SGYGISINV-----DKRKLVIRE------KGKQ
Metarv          ----------MNPLLI-----QGFGTRISV-----EKRRLKIS-------TETN
                           .   :     .     :       .

Ecoli           IRTHIP--VGSVACIMLEPGTRVSHAAVRLAAQVGTLLVWVGEAGVRVYASGQP-GGARS
VC_Cas1         SIEFDLKETDL--IMFYGATGELPLEFLDDASKNGVMILIHRRNVLQPYVFYPSVIGDEE
Strthe          TEMIHLSEIDI--LLLETTDIVLTTMLVKRLVDENILVIFCDDKRLPTAFLTPYYARHDS
Sulsol          KVEISPSEVDE---ILITVSCSISTSALSLALTHGISVMFLNSRETPWGILLPSIVTETV
Metsp2          EYIFSPKRIDIDSIIIYGKRGNLTFEAVRWLIKHNVQVTILNWNGKLLTTMLPP-ESTNV
Metmaz          EYVFSPKRIDIDGIIIYGKSGNLTLEAIRWLIKHNVQVSILDWNGKLLTTMLPP-ESTNL
Metmet          EYVFSPKRMDIDSIIVYGQSGSLSFEAIRWLIKHNIQITILDWNGKLLTTMLPP-ESTNV
Metbur          EYVFSPKRIDIDSIVVYGRSGSLSFEAIRWLIKHNVQVTMLDWNGKFLTTMLPS-ESTNV
Aciboo          VHEFYPHQINYDSLIIEGYYGNISFEAIRWLMKHNITVSVLNWNGNLLSVFLPK-EPING
Metarv          DYEFYPHQIDHDTIVIDGFTGNISFEAMRWIMKHKINLTLLDWNGNLLGTWMPK-ETSVG
                         .    :.      :.   :         :

Ecoli           DKLLYQAKLALDEDL----RLKVVRKMFELRFGEPAP-----------------------
VC_Cas1         DILTKQIQFRTNERK----RLYIAKTLIKKRLENMGSTIPISAPLLRQLS----------
Strthe          SLQIARQIAWKENVKCEVWTAIIAQKILNQSYYLGECSFFEKSQSIMELYHGLERFDPSN
Sulsol          KTKKAQYEAIVVRKD----N-RYGEEIISSKIYNQSVHL----KYWARVTGTKNDYKELL
Metsp2          KTKFAQYHAFEDQET----RLEIAKKFIEAKFDKSKTVLDFLSQRYPQIKFDVSDELTKL
Metmaz          RTKFAQYHAFEDKEA----RLEIAKKFIEAKFYKSKAVLDFLSQRYPEINFDILDGLTKL
Metmet          KTKFAQYHAYEDQES----RIKLAKKFIEAKISKSEAVLDYLKQRYPEIEYDFSDDKAKL
Metbur          KTKFAQYHAYEDQDA----RVKLARKFIEAKFYKSEAVLDYLKQRYPEIEYDFSVDKGKL
Aciboo          KLKIRQYEIYINEKE----RLKIAEKILEEKIRKSENMLYELSEYYPEIEHIKVKKRIEK
Metarv          KLRVKQYAKYLDPTT----RYNIAYQIIKEKVKKSCNLLTELSDYYEELDKSEIEEAFIN
                     .                    ::.

Ecoli           ------------ARRSVEQLRGIEGSRVRATYALLAKQY-----GVTWNGRRYDPKDWE
VC_Cas1         ------------AAKSIDEVRAIEANTTARYWNKWYENL-----NIETTRR--------
Strthe          R----------------------EGHSARIYFNTLF--------GNDFTRE--------
Sulsol          D--------------------KDEPAAARVYWQNISQLLPK---DIGFDGR-------D
Metsp2          K------------DGKSIRELMGIEGGLAWKYWNEFSKAIPK---EYDFCSR-IDQYRRP
Metmaz          K------------DVKSTREILGVEGTLAGKYWIEFSKAVPK---EYDFCNR-IDQFRRA
Metmet          E------------KVKSIRDILGVEGGVAWKYWNEYAKAVPK---GYDFKAR-TDNYTRA
Metbur          E------------NAKSVREILGIEGGVASKYWNEYSKAIPD---EYDFRAR-TDNNARA
Aciboo          EEKLKRDMELKEENKPKLSYLLMYEGRVAQIYWKELSKIFNKLYPEFNFTSRSTKSYSWN
Metarv          E--YTGFVAYNKKEQHDINKIMVYEAKTAKAYWDRLTKVFNKLYPDFRFTGRRNKSNSWN
                                       *   .   :
```

Ecoli: *E. coli* MG1655 CRISPR type I-E Cas1

VC_Cas1: CRISPR type V-C Cas1 found in mouse gut metagenome

Strthe: *Streptococcus thermophilus* CRISPR type II-A Cas1

Sulsol: *Sulfolobus solfataricus P2* CRSIPR type I-A Cas1

Metsp2: *Methanosarcina sp. 2.H.A.1B.4* casposase

Metmaz: *Methanosarsina mazei strain 3.F.A.1A.1* casposase

Metmet: *Methanococcoides methylutens* casposase

Metbur: *Methanococcoides burtonii* casposase

Acidboo: *Aciduliprofundum boonei* casposase

Metarv: *Methanocella arvoryzae* casposase

10. Appendix 3

Raw data for constructing table 3.1. Repeat 1 was shown in figure 29B.

Repeat 2:

Long DNA substrates	Colony number on cm plate for 10 µL of cells	Total transformed cells per mL	Colony number on cm and amp plate for 1 mL of cells	Integration efficiency
mini-casposon	4110	4.11×10^{5}	3	7.30×10^{-6}
***amp*^{R} blunt**	3500	3.50×10^{5}	0	0
***amp*^{R} 3′ overhangs**	3810	3.81×10^{5}	0	0

Repeat 3:

Long DNA substrates	Colony number on cm plate for 10 µL of cells	Total transformed cells per mL	Colony number on cm and amp plate for 1 mL of cells	Integration efficiency
mini-casposon	3470	3.47×10^{5}	3	8.65×10^{-6}
***amp*^{R} blunt**	8650	8.65×10^{5}	2	2.31×10^{-6}
***amp*^{R} 3′ overhangs**	5912	5.91×10^{5}	0	0

11. Appendix 4

Ten pre-integration sites identified from ten sequenced mini-casposon integration sites in pACYC-Duet. These 10 sites were used to generate the DNA logo plot in figure 29D. Bases in red represented target site which would be duplicated after integration.

5′ TCAAATGCCTGAGGTTTCAGCAAAAAACCCCTCAAGACCCGTTTA

5′ TCAAATGCCTGAGGTTTCAGCAAAAAACCCCTCAAGACCCGTTTA

5′ TAGCTGAACAGGAGGGACAGCTGATAGAAACAGAAGCCACTGGAG

5′ TCCGGCGGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGA

5′ TTAATGTAAGTTAGCTCACTCATTAGGCACCGGGATCTCGACCGA

5′ GCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGA

5′ AAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATA

5′ GTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAAC

5′ ACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACG

5′ ATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTG

12. Appendix 5

Protein sequences of fusion proteins were shown below with *A. boonei* casposase sequence coloured in green and Cas9/dCas9 sequence coloured in brown. The fusion proteins contained a $(His)_6$ tag at the N-terminal and a $(GGS)_8$ linker.

Protein sequence of Casp-Cas9:

MGSSHHHHHHSQDPMNPLLVSGYGISINVDKRKLVIREKGKQVHEFYPHQINYDSLIIEGYYGNI
SFEAIRWLMKHNITVSVLNWNGNLLSVFLPKEPINGKLKIRQYEIYINEKERLKIAEKILEEKIRKSE
NMLYELSEYYPEIEHIKVKKRIEKEEKLKRDMELKEENKPKLSYLLMYEGRVAQIYWKELSKIFNKL
YPEFNFTSRSTKSYSWNMNASDEINALLNYSYALLESMIRKHINAVGLDPSIGFLHELASSKTPLV
YDLQELFRWVSDLSVIQLLEDKKLKKSSFIVTENYHIRLKPQTSKLLVEKFKLNMNKKYEVGKKRYT
LETIMFNTVRSLGKYILGKSNTLKFEIPYIRIEHIEPELTEKILKMTPEERKARGINKSTLWYQKKKLA
QGKSIKVYGKVKSKLKGGSGGSGGSGGSGGSGGSGGSGGSMDKKYSIGLDIGTNSVGWAVITD
EYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEM
AKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLA
LAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLE
NLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYAD
LFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAIL
RRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASA
QSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFK
TNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLT
LTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGF
ANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMG
RHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQ
NGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDE
NDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYG
DYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKG
RDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAY
SVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGR
KRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEF
SKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVL
DATLIHQSITGLYETRIDLSQLGGD

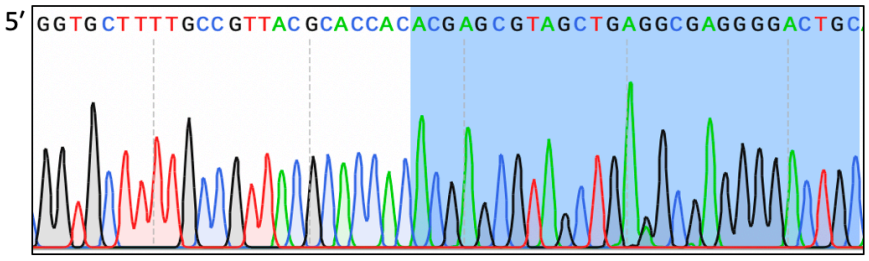
Protein sequence of Casp-dCas9:

MGSSHHHHHHSQDPMNPLLVSGYGISINVDKRKLVIREKGKQVHEFYPHQINYDSLIIEGYYGNI
SFEAIRWLMKHNITVSVLNWNGNLLSVFLPKEPINGKLKIRQYEIYINEKERLKIAEKILEEKIRKSE
NMLYELSEYYPEIEHIKVKKRIEKEEKLKRDMELKEENKPKLSYLLMYEGRVAQIYWKELSKIFNKL
YPEFNFTSRSTKSYSWNMNASDEINALLNYSYALLESMIRKHINAVGLDPSIGFLHELASSKTPLV
YDLQELFRWVSDLSVIQLLEDKKLKKSSFIVTENYHIRLKPQTSKLLVEKFKLNMNKKYEVGKKRYT
LETIMFNTVRSLGKYILGKSNTLKFEIPYIRIEHIEPELTEKILKMTPEERKARGINKSTLWYQKKKLA
QGKSIKVYGKVKSKLKGGSGGSGGSGGSGGSGGSGGSGGSMDKKYSIGLAIGTNSVGWAVITD
EYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEM
AKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLA
LAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLE
NLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYAD
LFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAIL
RRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASA
QSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFK
TNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLT
LTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGF
ANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMG
RHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQ
NGRDMYVDQELDINRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDE
NDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYG
DYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKG
RDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAY
SVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGR
KRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEF
SKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVL
DATLIHQSITGLYETRIDLSQLGGD

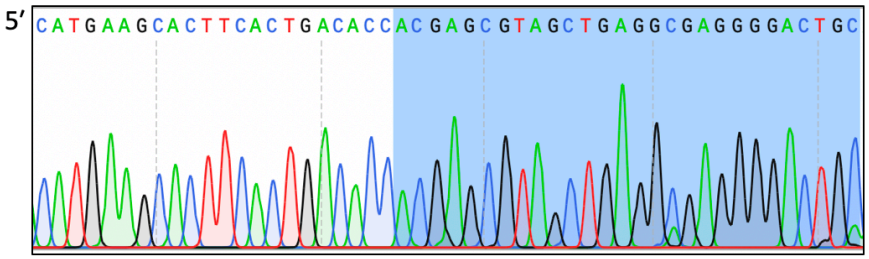
13. Appendix 6

Sequencing data of purified DNA from excised bands corresponding to site A and site B integration products in figure 46B. The highlighted regions are the complementary strand of integrated TK24 or TK25.

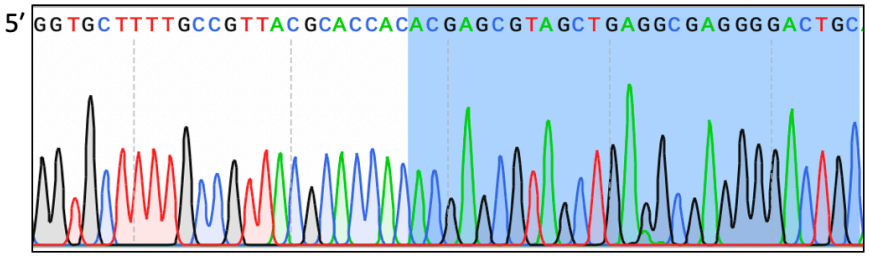
Casp-Cas9 catalysed TK24 insertion into site A in the absence of sgpACYC



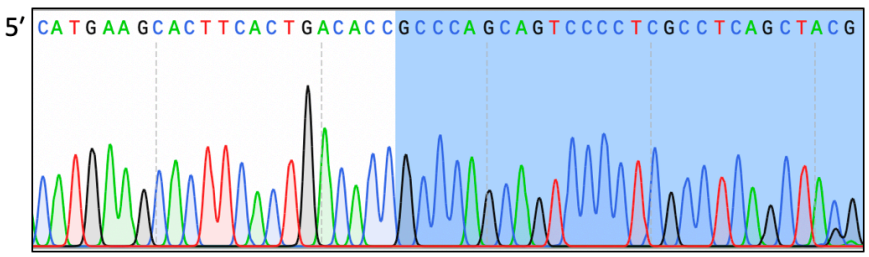
Casp-Cas9 catalysed TK24 insertion into site B in the presence of sgpACYC



Casp-dCas9 catalysed TK24 insertion into site A in the absence of sgpACYC



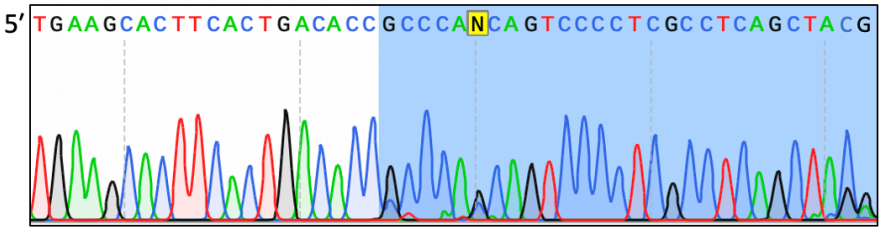
Casp-dCas9 catalysed TK25 insertion into site A in the presence of sgpACYC



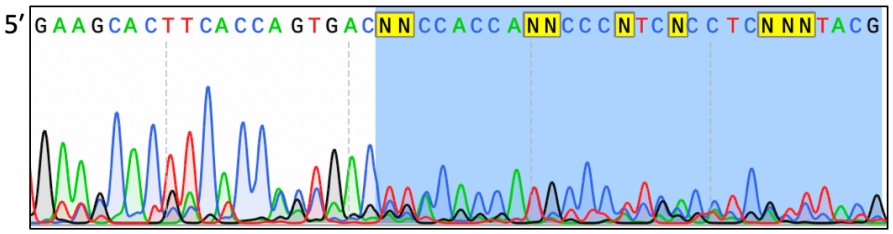
14. Appendix 7

Sequencing data of purified DNA from excised bands corresponding to site B and mutated site B integration products in figure 47A. The highlighted regions are the complementary strand of integrated TK25. The sequencing quality of mutated site B fragment was poor around the insertion site. Thus, some bases of the TK25 complementary strand were masked by other bases.

Casp-Cas9 with sgpACYC catalysed TK25 insertion into site B in pAB



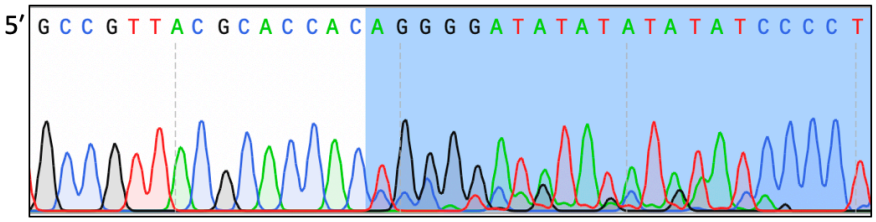
Casp-Cas9 with sgpACYC catalysed TK25 insertion into mutated site B in pABmut



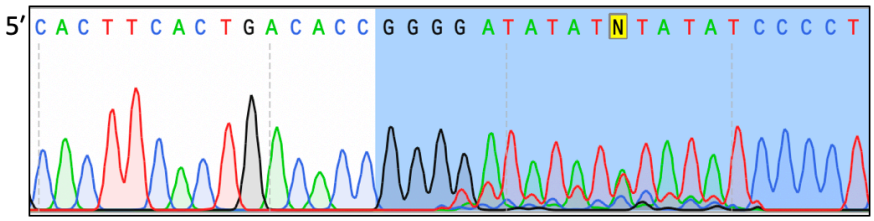
15. Appendix 8

Sequencing data of purified DNA from excised bands corresponding to site A and site B integration products in figure 49B. The highlighted regions are the LE TIR of the mini-casposon. An extra adenine at the beginning of mini-casposon was likely added by the One*Taq*® Hot Start DNA Polymerase during PCR amplification of the mini-casposon.

Casposase catalysed mini casposon insertion into site A in pACYC-Duet



Casp-Cas9 with sgpACYC catalysed mini-casposon insertion into site B in pACYC-Duet



16. Appendix 9

The 1.2 kb PCR fragment from figure 57A was sequenced at both ends using TK25 and lacZ F primers. The sequence was mapped to *E. coli* chromosomal DNA and arose from non-specific binding of both primers during PCR.