

A A T A A T T T C T A C T G
T C A T T T A A T A A G G C
T A **R I G H T** A G A A C T T
T G T T G T A G A T G C T A
T C A G A T A A **T O O L** T T
A G C T T G G C G G G C G A
A G T G C A A A A C C T T T
G T **F O R C A T H E** T A G C
G A A G G C C A A G A C G T
T A A T T T C T A C T G T T
A T **R I G H T** T T A A T A A
G T C T A A G A A C T T T A
T T G T A G A T G C T A C T
A G A T A A T T C **J O B** A G
T T G G C G G G C G A A G G

Exploring the diversity of
type V CRISPR-Cas systems

Wen Ying Wu

Propositions

1. Adaptation is the middle child of CRISPR-Cas research. (this thesis)
2. The phylogeny-based classification of type V CRISPR-Cas does not allow for prediction of mechanistic features of Cas12 variants. (this thesis)
3. Curiosity should be the main motivation in scientific research.
4. There is a need for a scientific journal in which negative results can be published.
5. Human genome editing is a double-edged sword.
6. Taking care of mental health should be a social norm, not a stigma.

Propositions belonging to the thesis entitled:

Right tool for the right job: *Exploring the diversity of type V CRISPR-Cas systems*

Wen Ying Wu
Wageningen, 12th February 2021

RIGHT TOOL FOR THE RIGHT JOB

Exploring the diversity of type V CRISPR-Cas systems

Wen Ying Wu

Thesis committee

Promotor

Prof. Dr John van der Oost
Professor of Microbial Genetics
Wageningen University & Research

Co-promotor

Dr Raymond H.J. Staals
Assistant Professor at the Laboratory of Microbiology
Wageningen University & Research

Other members

Prof. Dr Dolf Weijers, Wageningen University & Research
Prof. Dr Michiel Kleerebezem, Wageningen University & Research
Dr Gorben P. Pijlman, Wageningen University & Research
Dr Chirlmin Joo, Delft University of Technology

This research was conducted under the auspices of the Graduate School VLAG
(Advanced studies in Food Technology, Agrotechnology, Nutrition and Health Sciences)

RIGHT TOOL FOR THE RIGHT JOB

Exploring the diversity of type V CRISPR-Cas systems

THESIS

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Friday 12th February 2021

at 4 p.m. in the Aula

Right tool for the right job
Exploring the diversity of type V CRISPR-Cas systems
244 pages

PhD thesis, Wageningen University, Wageningen, NL (2021)
With references, with summary in English

ISBN: 978-94-6395-668-0
DOI: <https://doi.org/10.18174/538075>

Table of contents

Chapter 1	General introduction & thesis outline	2
<hr/>		
Chapter 2	Genome editing by natural and engineered CRISPR-associated nucleases	16
<hr/>		
Chapter 3	Adaptation in type V-A and type V-B CRISPR-Cas systems	36
<hr/>		
Chapter 4	Multiplex gene editing by CRISPR-Cas12a (Cpf1) using a single crRNA array	68
<hr/>		
Chapter 5	Cut and paste: genome editing of <i>E. coli</i> using Cas12a and T4 ligase	90
<hr/>		
Chapter 6	Characterizing a compact CRISPR-Cas12u1 enzyme	108
<hr/>		
Chapter 7	Small and mighty: MmuCas12u1 C-to-T base editors	158
<hr/>		
Chapter 8	Summary and general discussion	194
<hr/>		

ALWAYS START

----- **AND END YOUR DAY WITH A SMILE**

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
G	A	G	T	T	C	C	C
G	G	G	A	T	A	A	A
A	A	G	T	C	T	A	A

A T A A T T T C
T C H A P T E R
C G C C A G C G
C C G T T A A 1
A A C T T T G T

General introduction
& thesis outline

Science then vs now

In 1660, the Royal Society was established in London as the first national scientific institution in the world. It was a scientific community consisting of rich, Caucasian, Christian, gentlemen pursuing their curiosity-driven scientific hobby (1). Publications of single authors were published in the Proceedings of the Royal Society. The Royal Society followed the idea of acquiring knowledge through experimental investigation and had a motto: “*Nullius in verba*” (Latin), which means “Take nobody’s word for it” (2).

Fast forward to the present day. Scientific institutions have been founded all over the world, generally covering many domains of science. The scientific community has diversified, to include people of different sex, ethnicity and religious beliefs (3). Scientific findings are now submitted to a scientific journal (one of the thousands), peer-reviewed by non-biased colleagues, and eventually distributed online, making knowledge transfer quick and straightforward. A lot has changed in science, but the basics remain the same. Scientists still are curious about all aspects of life, and for that reason they still conduct “curiosity-driven” research (4).

This chapter starts with an introductory story on how curiosity-driven fundamental research can lead to extraordinary discoveries with spectacular applications, the story of CRISPR-Cas (5). Then follows an overview of the classification and mechanism of CRISPR-Cas systems. Eventually a summary is provided of one of the more recently discovered, highly diverse type V CRISPR-Cas systems.

The (short) history of CRISPR-Cas systems

In 1987, Japanese scientists were looking at the DNA sequence of an enzyme-encoding gene from the bacterium *Escherichia coli* (*E. coli*) (6) Yoshizumi Shinagawa, Hideo Makino, . Downstream the gene they found a cluster of short 30 base pair (bp) long repeated palindromic sequences. These invariable repeated sequences were interspaced by ~32 bp variable DNA sequences. At that time, the scientists were unable to come up with a physiological role for this phenomenon, and just published this information as an observation. Six years later in Spain, similar DNA repeated sequences were observed in halophilic archaea (7). Inspired by the latter authors, using the genome sequences that became available in those days, it was a group in the Netherlands that coined a name for this repeated region: **C**lustered **R**egularly **I**nterspace **S**hort **P**alindromic **R**epeats, **CRISPR** (now known as CRISPR-array) (8).

Genes that are located adjacent to CRISPRs were named **CRISPR associated (Cas)** genes (8), which led to the name of CRISPR-Cas (8). A few years later, three groups independently reported that the variable 32 bp spacer sequences of the CRISPR arrays does correspond to phage DNA (viruses that attack bacteria or archaea), which generated the idea that CRISPR-Cas is an anti-viral defense system in bacteria and archaea (9-11). Based on bioinformatic analysis of cas genes domains and CRISPR transcripts, CRISPR-Cas was predicted to function similarly to eukaryotic RNA-guided RNA interference (RNAi) systems (12). A key experimental breakthrough of CRISPR-Cas occurred in 2007, while scientists from a dairy company were searching for a phage resistant lactic acid bacterium, *Streptococcus Thermophilus*, used for yoghurt fermentation (13). A milestone in CRISPR research was the discovery that adaptation of the CRISPR array of the bacteria occurred through acquisition of new spacers from the phage genome, and that this resulted in phage resistance. However, precise spacer and target (protospacer) match was required, as phage escapers contained point mutations within the protospacer region. In addition, cas genes, such as cas9 (previously known as cas5) were required for phage immunity (13).

A year later, in 2008, first mechanistic insights were gained on CRISPR-Cas when studying the CRISPR-Cas system in *E. coli*. The *E. coli* CRISPR-Cas system (class 1, see below for details on classification) differs to the aforementioned *S. thermophilus* Cas9 system (class 2, see below), as it required a **CRISPR-associated complex** for **antiviral defense** (Cascade) consisting of five different Cas proteinases as well as a Cas3 nuclease (14). After transcription of the CRISPR-array, the generated precursor CRISPR-RNA (pre-crRNA) is processed into mature crRNA by a subunit of the Cascade complex (14). Mature crRNAs are guides that result in successful targeting (i.e. phage protection) in both the sense and the anti-sense orientation suggesting dsDNA (14), instead of the previously hypothesized RNA (9). After identifying the mature crRNAs, the first artificial CRISPR was created to alter the crRNA guide sequences, showcasing the programmability of DNA targeting by CRISPR-Cas (14).

The initial curiosity-driven search for the functional role of the unique repetitive sequences, cascaded into the discovery of the role of CRISPR-Cas as a unique adaptive immunity systems. A series of seminal fundamental studies on the structure and function of the key components of these systems in turn paved the way for establishing innovative applications of CRISPR-Cas, such as genome editing.

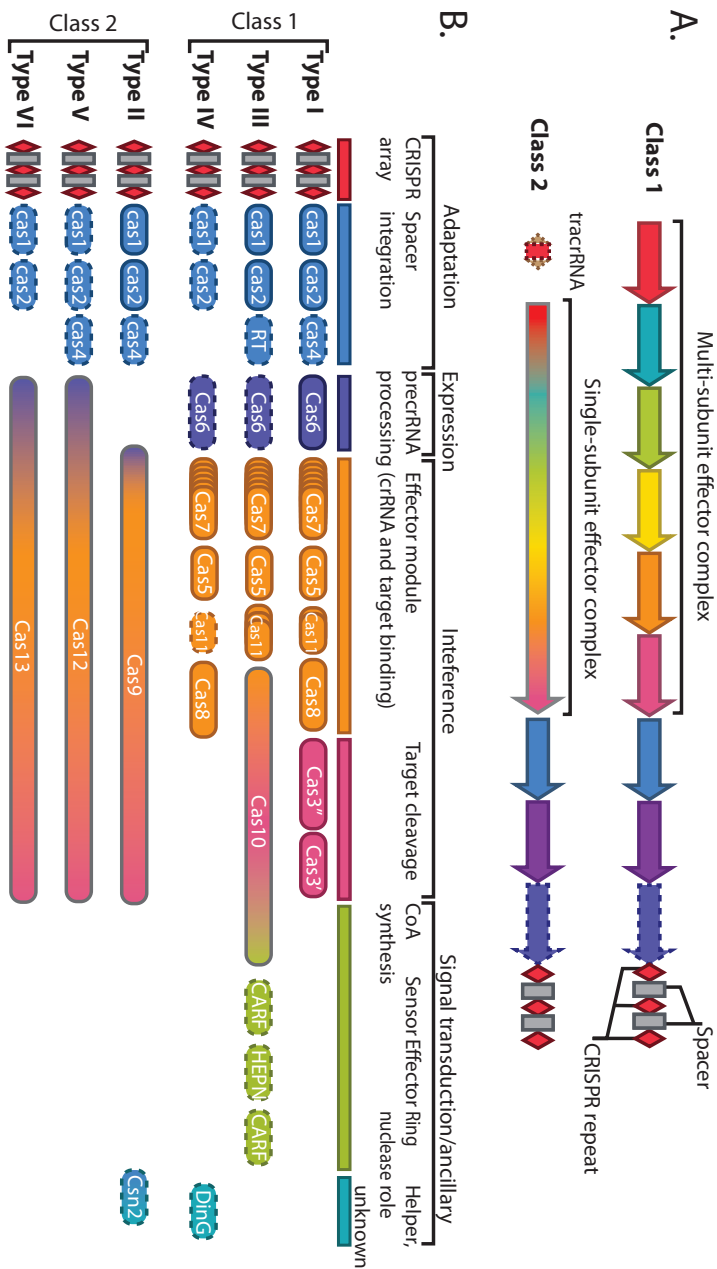


Figure 1 | Schematic classifications of CRISPR-Cas loci. Genes with dashed outlines are dispensable or missing in some CRISPR-Cas systems of that class, type or variant. **(A)** class 1 and class 2 CRISPR-Cas loci. Both classes contain a CRISPR-array (consisting of CRISPRs and spacers) and an adaptation module (genes in blue, purple and navy), but differ in their effector modules. Class 1 encompasses a multi-subunit effector complex whereas, class 2 encompasses a single-subunit effector complex. In addition, in some class 2 systems, a tracrRNA is also required. **(B)** Gene functionality and organization of six types of CRISPR-Cas systems. LS is large subunit; SS is small subunit and RT is reverse transcriptase. Within the two classes are type I, III and IV belonging to class 1 and type II, V and VI belonging to class 2. cas genes are separated and ordered based on functionality within the CRISPR-Cas mechanism, those being: adaptation, expression, interference, and signal transduction/ancillary. Figure adapted from (18).

CRISPR-Cas classification

Since its first discovery, the number of different CRISPR-Cas systems has steadily increased over the years. CRISPR-Cas systems are divided into two major categories, class 1 and class 2 (Fig. 1A) (15). This first classification is based on the effector module of the CRISPR-Cas system i.e., the protein (complex) involved in guide-based targeting. Class 1 CRISPR-Cas systems requires a multiprotein subunit complex, such as the aforementioned Cascade and variants thereof (14). In class 2 CRISPR-Cas systems, the effector modules consist of a single multidomain protein such as Cas9 (16). Each of the two classes are sub-divided into three types of CRISPR-Cas systems, which depends on the domain architecture of effectors protein(s) (Fig. 1B) (17). Class 1 consists of type I, III and IV, containing multi-subunit complexes with similar architecture: Cascade, Csm/Cmr and an “unknown” complex, respectively. Class 2 consists of type II, V and VI, containing Cas9, Cas12 and Cas13, respectively. Each type is then further divided into subtypes based on CRISPR loci organization and *cas* gene repertoires aside from the effector module (18). At least until recently, many CRISPR-Cas systems have been discovered and characterized each year. New CRISPR loci that do not meet the criteria to be included in the previously identified subtypes are assigned to new subtypes. An example of that, is type V which currently contain 11 subtypes. The most recent CRISPR-Cas classification includes 2 classes, 6 types and 34 subtypes (18).

Molecular mechanism CRISPR-Cas system

After uncovering the mystery behind the repeats, CRISPR-Cas was discovered to be an adaptive immune system in bacteria and archaea against phages or mobile genetic elements (MGE) (14, 19). CRISPR-Cas mediated adaptive immunity consists of three steps: adaptation, expression and interference (Fig. 2) (20, 21). Adaptation is the first step towards obtaining immunity and occurs when a short dsDNA (pre-spacer) of an MGE is acquired by Cas proteins and inserted into the CRISPR-array (22). During expression, the CRISPR-array is transcribed into long pre-crRNA and processed into mature crRNA. Mature crRNAs are then bound to effector protein(s) to form a ribonucleoprotein (RNP) complex (14). Then during interference, the ribonucleoprotein searches for its corresponding protospacer. Apart from matching the sequence of the spacer, the protospacer must also contain a protospacer adjacent motif (PAM) (9). This allows CRISPR-Cas systems to distinguish between self (spacer sequence in CRISPR-array) and non-self (the MGE). Once a PAM has

been found and the spacer matches the protospacer, the Cas protein cleaves the invader's DNA, eliminating the MGE from the cell (23).

What was just described, is a quick glance on the molecular mechanism of CRISPR-Cas systems. CRISPR-Cas systems are diverse containing many sophisticated distinct features between the systems during each step. Those details are further elaborated on below.

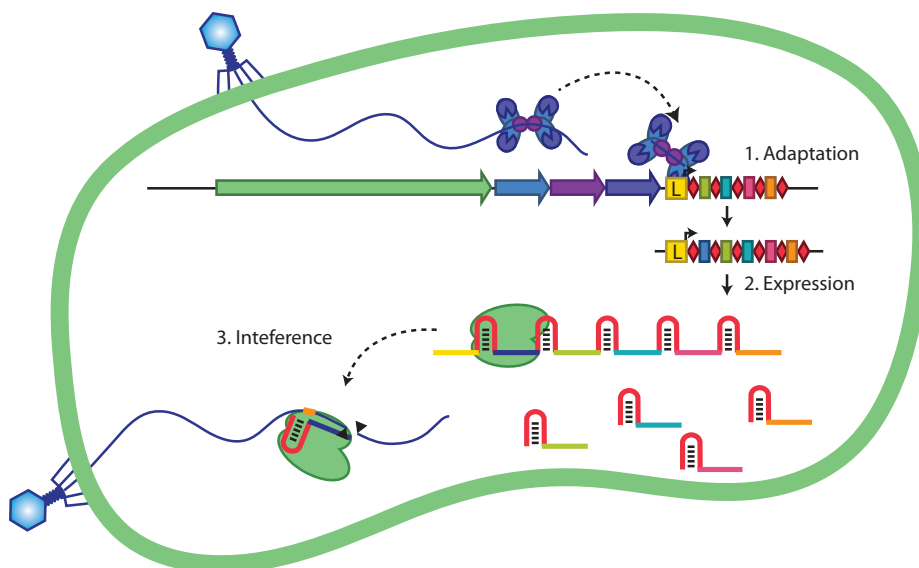


Figure 2 | Overview of the mechanism of CRISPR-Cas system. During a phage infection, DNA is injected into the cell. In the first step, **adaptation**, the adaptation complex captures a short dsDNA fragment from the phage and inserts it into the CRISPR-array at the leader end (yellow) between the duplicated halves of the first repeat (red diamond). The CRISPR-array is expanded to contain a new spacer (dark blue) against the invading phage. During **expression**, the CRISPR-array is transcribed into a long pre-crRNA which is then processed into mature crRNAs. Mature crRNAs bind to an effector complex to form an RNP complex. Lastly, in **interference**, the RNP searches for a protospacer adjacent motif (PAM) (orange) and once found the spacer attempts to base pair with one of the DNA strands of the targeted region. In case of a matching target (protospacer), then the RNP-associated nuclease cleaves the DNA of the phage, resulting in neutralization of the viral attack.

Adaptation

During the acquisition of a new spacer, the spacer is inserted into the CRISPR-array at the 'leader'-end, between the duplicated copies of the first repeat. The adaptation module often comprises of Cas1 and Cas2. Cas1 and Cas2 form a protein complex consisting of two Cas1 homodimers connected by a Cas2 homodimer (Cas1₂-Cas2-Cas1₂; Fig 2) (24). Cas1 is crucial for adaptation, as mutation in the active site

of Cas1 abolishes adaptation, whereas disruption of the active site of Cas2 does not (22, 25). In some CRISPR-Cas systems, other Cas proteins also aid in adaptation (26). Like in type I systems, Cas4 has been demonstrated to select the correct PAM containing spacers, trim the pre-spacer to its correct length and adjust the pre-spacer orientation (26-28). In Type II-A, Cas9 and Csn2 are required for adaptation to occur (29, 30). Cas9 selects for correct PAM-containing pre-spacers and Csn2 is hypothesized to stabilize the adaptation complex for capturing new spacers (31, 32).

There are two distinct ways adaptation can occur, naïve adaptation and primed adaptation (33). When a cell is exposed to a MGE for the first time, the acquisition of spacers is called naïve adaptation. Primed acquisition occurs during re-infection of a cell (that already acquired a spacer for this MGE previously) by the same MGE or a mutant MGE (34). The rate of spacer acquisition during primed adaptation is much faster than during naïve adaptation, since immunity has been previously obtained against the attacking MGE (35). In case of type I-E, cleavage by Cas3 generates short DNA degradation products, which are then used for acquisition of new spacers (36). Recently, primed adaptation was also found in class 2 systems, type II-A, where cleavage by Cas9 led to increasing adaptation rates (37). ‘

Expression

During the expression stage, transcribed pre-crRNA is processed into mature crRNA. The way pre-crRNA is processed differs between systems. For example, in type I and III systems, the Cas6 ribonuclease processes the pre-crRNA on the 3' of the repeat sequence resulting in an 8 nt repeat-product that forms the 5' handle of the mature crRNA (Fig. 3A) (38). However, in type II systems, a transactivating RNA (tracrRNA) is required that base pairs with the repeat sequence of the crRNA (39). Cas9 binds to the tracrRNA:pre-crRNA duplex to form a ribonucleoprotein. Host RNase III recognizes and cleaves the RNA duplex, leaving a 3' 2 nt overhang (Fig. 3B) (39, 40). In type II-A of *Francisella Novicida*, a small CRISPR-Cas associated RNA (scaRNA) can also base pair with the tracrRNA. Cas9 containing a tracrRNA:scaRNA duplex can target and regulate transcription to aid in the virulence of *F. novicida* (41, 42). In some type V (e.g. V-A) and all type VI systems, pre-crRNA processing is simpler and requires no tracrRNA or external protein. In these systems, comparable to the type I Cascade complex), the effector protein itself processes the pre-crRNA (43, 44). Cas12a and Cas13 recognize the hairpin structure in the palindromic repeats of the pre-crRNA. Cas12a cleaves the repeat just upstream the hairpin/pseudoknot structure, after which the RNP holds on to the mature crRNA guide (Fig. 3C) (44-46). After initial processing by either Cas6, RNase III or Cas12a, secondary processing of pre-crRNA occurs via non-Cas RNases, which trims either the 3' end (type I and type V-A) or the 5' end (type II) of the crRNA (15).

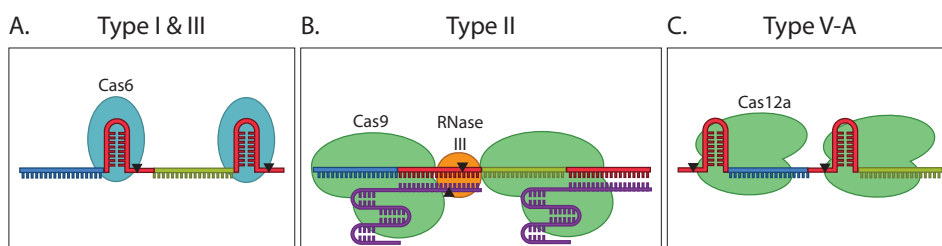


Figure 3 | Pre-crRNA processing in type I, II, III and V-A. The pre-crRNA consists of two spacers (green and blue) and two repeats (red). **(A)** In type I and type III systems processing is catalyzed by Cas6, which recognizes a hairpin formed in the repeat sequence and cleaves the 5' end of the repeat in the pre-crRNA. **(B)** In type II system, a tracrRNA, partly complementary to the repeat sequence, hybridizes to the repeat. This RNA duplex is recognized and cleaved by RNase III. **(C)** In type V-A system, Cas12a can process its own pre-crRNA by recognition of the hairpin form by the repeat. Cas12a cleaves at the 3' end of the repeat, to generate mature crRNAs.

Interference

The first step of interference is surveillance for the correct PAM, i.e. PAM scanning (47, 48). Different Cas effectors recognize different PAM sequences (49). Type III and type VI systems target RNA instead of DNA and recognize a 5' RNA PAM (rPAM) and a 3' protospacer flanking sequence (PFS), respectively (Fig. 4) (50, 51). After finding a correct PAM, base pairing occurs between the first 5–10 nucleotides of the guide, the seed region, and the target strand (52, 53). If the spacer matches or mismatches with the seed, further base pairing follows, leading to a complete unwinding of the target DNA (so-called R-loop structure), that is required for target cleavage. However, if a mismatch is present, base pairing is aborted and the ribonucleoprotein dissociates and continues to search for its target.

DNA targeting

Polynucleotide targeting differs between types. In type I, II and most type V CRISPR-Cas systems use their guide to specifically target dsDNA. In Type I systems, Cascade binds to dsDNA and recruits the Cas3 nuclease-helicase. Cas3 then nicks the non-target strand and, using its helicase activity, continues to degrade the non-target in the 3'-to-5' direction using a 'reeling' mechanism (Fig. 4) (54, 55). In type II, Cas9 binds to dsDNA and generates blunt-ended double stranded break at the PAM proximal end (56). In most type V systems, Cas12 generates staggered ended double stranded break at the PAM distal end (43, 57–59).

RNA targeting

As mentioned before, in type III and type VI systems, RNA is targeted instead of dsDNA. Type III systems are unique, as they are transcription dependent RNA and DNA nucleases. Csm/Cmr complexes bind to the targeted mRNA and cleave the mRNA in chunks of six nucleotides (60). Then Cas10 cleaves non-specific adjacent ssDNA, which is the coding strand during transcription of the targeted mRNA (Fig. 4). During ssDNA degradation, Cas10 converts ATP to cyclic oligoadenylate (c(OA)). cOA then activates Csm6/Csx1 to degrade non-specific collateral RNA in *trans* (61, 62). Collateral RNA degradation also occurs in type VI systems, where Cas13 binds and cleaves only ssRNA (44) (Fig. 4). The cleavage sites within the targeted RNA depends on the target sequence and RNA structure. Cleavage of the targeted RNA then activates Cas13 to degrade non-specific collateral RNA in *trans* (44). This collateral RNA degradation induces cell death or cell dormancy upon a severe phage infection to prevent outbreaks (63).

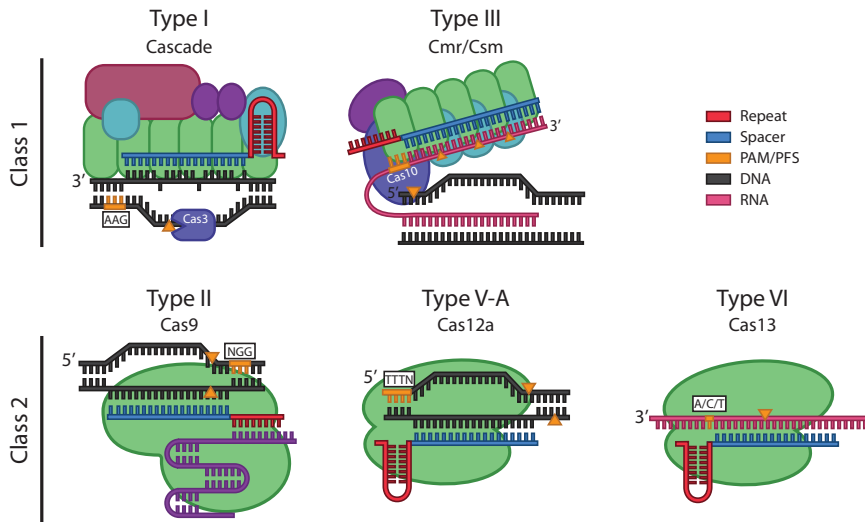


Figure 4 | Target cleavage by different types of CRISPR-Cas nucleases. Class 1 systems, type I and type III contain effector protein complexes, Cascade and Cmr, respectively. Cascade recognizes 5' CTT-protospacer adjacent motif (PAM) and binds dsDNA. Cascade then recruits Cas3, which nicks the non-target strand and continues on to cleave to the non-target strand in small pieces towards the 3' end. In type III, Cmr recognizes an rPAM and binds ssRNA, including transcripts. Once bound, Cmr cleaves the RNA in 5nt RNA fragments and Cas10 nicks the dsDNA, where transcription occurs with the target RNA. In Class 2, type II, type V-A and type VI contain single effector proteins, Cas9, Cas12a and Cas13. In type II, Cas9 recognizes a NGG- PAM, binds dsDNA and generates blunt-ended double stranded breaks (DSB) at the PAM proximal end. In type V-A, Cas12a recognizes a 5' TTTN- PAM, binds dsDNA and generates staggered-ended double stranded breaks at the PAM proximal end. In type VI, Cas13 recognizes a non-G (A/C/T) PFS, binds and cleaves ssRNA targets.

Type V CRISPR-Cas systems

Type V is currently the most diverse type of all CRISPR-Cas systems, consisting of characterized subtypes A to K and uncharacterized subtypes U1-U4 (18). By bioinformatic analysis, type V systems are proposed to have been evolved several times independently from transposon-encoded TnpB nucleases, which eventually yielded many separate subtypes containing distinct features (64). This analysis was based on the RuvC domain found in Cas12 nucleases, which is involved in cleavage of dsDNA (65). Type V-A, containing Cas12a (Cpf1), was the first type V CRISPR-Cas system to be characterized. In addition, Cas12a is also the first Cas effector protein found to process its own crRNA (43, 46). As for type V-B, Cas12b (previously called C2c1) shares many features with Cas12a, except for the requirement of a tracrRNA (58); this resembles the RNase III-dependent pre-crRNA/tracrRNA processing of Cas9. Other type V effector proteins that also require a tracrRNA are Cas12c, Cas12e (CasX), Cas12f (Cas14), Cas12g and Cas12k (57, 66-69). Cas12d (also called CasY) from type V-D was recently discovered to require short-complementarity untranslated RNA (scoutRNA) instead of a tracrRNA (70). Apart from crRNA processing, type V nucleases also differ regarding the nature of their polynucleotide target and generated cleavage products. However, one common feature shared amongst all Cas12 nucleases, is the recognition of 5' T-rich PAM (Table 1). The majority of Cas12 nucleases (Cas12a-e and Cas12h-j) target and cleave dsDNA to generate staggered ends. However, not all Cas12 nucleases generate double stranded breaks, as other Cas12 nucleases vary in their RuvC specificity (single stranded versus double stranded targets, and DNA vs RNA). For example Cas12i also targets dsDNA, but predominantly nicks dsDNA (57). Another exception is Cas12f1, which was first reported to exclusively target and cleave ssDNA (67). However, another study claims that Cas12f1 is also capable of targeting and cleaving dsDNA (71). Furthermore, Cas12g only targets and cleaves ssRNA instead of dsDNA (57). Lastly, Cas12k from type V-K (previously V-U5) contains an inactive RuvC domain and targets dsDNA. Cas12k does not cleave dsDNA, but instead recruits transposon proteins and initiates RNA guided transposition (69). After target cleavage, most of the type V nucleases (including Cas12a) were also found to cleave collateral ssDNA or ssRNA in *trans* (Table 1).

Aside from the aforementioned characterized type V CRISPR-Cas systems, several discovered type V sub-types remain to be characterized, such as type V-U1, -U2, -U3 and U-4. This thesis will focus on elucidating unknown features of type V CRISPR-Cas systems. More specifically, the characterization of fundamental features of type V-A and type V-U1 CRISPR-Cas systems and their subsequent repurposing towards genome editing applications.

Table 1 | Summary of type V nucleases characteristics

Type	Effector protein	size (aa)	PAM (5')	pre-crRNA processing	tracrRNA	Target	cleavage	collateral
V-A	Cas12a/Cpf1	~1300	TTTV	yes	no	dsDNA	4,5 nt staggered ends	ssDNA
V-B	Cas12b/C2c1	~1130	DTTD	no	yes	dsDNA	6-9 nt staggered ends	ssDNA
V-C	Cas12c	1209-1330	TN	no	yes	dsDNA	double stranded break	ssDNA/ssRNA
V-D	Cas12d/CasY	~1200	TA	no	no - scoutRNA	dsDNA	double stranded break	-
V-E	Cas12e/CasX	986	TTCN	no	yes	dsDNA	8-13 nt staggered ends	-
V-F	Cas12f/Cas14	400-700	TTN	no	yes	ssDNA/dsDNA	2-4 staggered ends	ssDNA
V-G	Cas12g	720-830	-	no	yes	ssRNA	-	ssDNA/ssRNA
V-H	Cas12h	870-924	RTR	yes	no	dsDNA	double stranded break (in vivo)	ssDNA
V-I	Cas12i	1033-1093	TTN	yes	no	dsDNA	preferentially nicks	ssDNA
V-J	Cas12j/Cas12ϕ	700-800	TBN	yes	no	dsDNA	8-12nt staggered ends	-
V-K (VU5)	Cas12k	~650	GTN	no	yes	dsDNA	guided transposition	-

Thesis outline

Chapter 1 provides a brief history of the discovery of CRISPR-Cas systems and how it started from being incredibly curious about short repeated DNA sequences in a bacterial genome. This curiosity led to the discovery of a sophisticated adaptive immune system in prokaryotes and archaea. Research in the uncovering of CRISPR-Cas systems laid down the steppingstones for the creation of a groundbreaking genome editing tool, able to modify DNA in all forms of life. The discovery of the first few CRISPR-Cas systems cascaded to the exploration of many more diverse systems, the characterization of which reveals new insights on the diversity and functionality of CRISPR-Cas systems, but also new innovative ideas towards the application Cas proteins.

Chapter 2 | *Genome editing by natural and engineered CRISPR-associated nucleases*

The second chapter reviews the different class 2 CRISPR-Cas proteins that are applied in genome editing, such as Cas9, Cas12a and Cas12b. Despite distant similarities, these proteins have unique structural and functional features that are compared for both natural and engineered CRISPR-Cas variants. In addition, other aspects of CRISPR-Cas genome editing are discussed to optimize genome editing efficiency and precision, including nuclease regulation, nuclease delivery systems, and relevant features of host repair.

Chapter 3 | *Adaptation in type V-A and type V-B CRISPR-Cas systems*

The third chapter studies the adaptation of CRISPR-Cas systems, the first step of the CRISPR-Cas mechanism, during which immunity is acquired. More specifically, this chapter uncovers the adaptation mechanism of two class 2, type V CRISPR-Cas systems, namely V-A and V-B. This chapter focusses on determining the cas genes required and their mechanistic role within adaptation. By overexpressing the CRISPR-Cas locus in *E. coli* and analyzing the spacers acquired, it was realized that adaptation in type V differs to some extent from the guide acquisition process in other class 2 systems. In type V-A, only Cas1 and Cas2 are required for adaptation, whereas in type V-B, Cas4/1 and Cas2 are required for adaptation. Although Cas4 is not required in type V-A, it increases the efficiency of adaptation by PAM-scanning for PAM containing pre-spacers.

Chapter 4 | *Multiplex gene editing by CRISPR-Cas12a (Cpf1) using a single crRNA array*

The fourth chapter elucidates the crRNA maturation mechanism of type V-A CRISPR-Cas system that differs from other class 2 CRISPR-Cas systems, such as type II. In type II systems, Cas9 requires both a crRNA and a tracrRNA, which then gets processed by an endogenous ribonuclease, RNaseIII. In type V-A systems, Cas12a requires only a single pre-crRNA and Cas12a is solely responsible for processing pre-crRNAs into mature crRNAs. This characteristic of Cas12a is highly advantageous for genome editing, as it allows for easy simultaneous editing of multiple targets (multiplexing). Multiplexing is demonstrated in this chapter by using a single

customized CRISPR array to simultaneously edit up to four genes in mammalian cells, and three genes in the mouse brain.

Chapter 5 | *Cut and paste: genome editing of *E. coli* using Cas12a and T4 ligase*

The fifth chapter demonstrates the proof of concept of a novel genome editing method in *E. coli* using Cas12a nuclease and T4 ligase, termed “cut & paste”. Cas12a targets and cleaves dsDNA while generating 5nt staggered ends. These staggered ends can be utilized and designed to be compatible, so that after cleavage of Cas12a at two different target locations, generated compatible sticky ends can be ligated and repaired by T4 ligase. Although low editing efficiency was observed, cut & paste is demonstrated to generate a genomic deletion in *E. coli*. Further improvements of the system are required to make it a more suitable tool for genome editing in prokaryotes.

Chapter 6 | *Characterizing a compact CRISPR-Cas12u1 enzyme*

The sixth chapter focuses on the characterization of a small Type V-U1 effector protein, MmuC2c4 from *Mycolicibacterium mucogenicum*. Type V-U1 lacks an adaptation module and, just like the other Cas12 variants, is thought to have evolved from the transposon-encoded TnpB. Like Cas12a, MmuCas12u1 catalyzes the maturation of its single crRNA guide, it recognizes a 5'-TTN PAM and binds double-stranded DNA. Unexpectedly, MmuCas12u1 does not cleave dsDNA, but instead enhances transcriptional silencing in *E. coli*. Using this unique feature, MmuCas12u1 has been applied as a silencing tool in *E. coli* for single and multiplex targeting. Current experiments suggest that MmuCas12u1 has an unprecedented mechanism of dsDNA-dependent mRNA transcript cleavage.

Chapter 7 | *Small and mighty: MmuCas12u1 C-to-T base editors*

The seventh chapter applies the knowledge gained on MmuCas12u1 from chapter 6, to engineer a C to T base editors (~2.8 kbp) using the small MmuCas12u1. MmuCas12u1 base editors (MmuBE) enable highly efficient C to T base editing in *E. coli* within a wide editing window. The base editing window of MmuBEs consist of two regions, a PAM-proximal (2-5) and a PAM-distal (13-19) region. In addition, preliminary results suggest that MmuBE is also active in baker's yeast. The MmuBEs presented in this chapter are excellent additions to the current base editing toolbox for prokaryotic base editing and show great promises for eukaryotic base editing.

Chapter 8 | *Summary and general discussion*

The final chapter summarizes the work described in this thesis. Moreover, some remaining questions and future perspectives on CRISPR-Cas are discussed, both from a fundamental and an application-oriented perspective.

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
G	T	T	T	T	A	G	A
T	A	G	C	A	A	G	T
A	G	G	C	T	A	G	T

A T A A T T T C
T C H A P T E R
G C T A G A A A
T A A A A T A 2
C C G T T A T C

Genome editing by natural and engineered CRISPR- associated nucleases

Wen Y. Wu, Joyce H.G. Lebbink, Roland Kanaar, Niels Geijsen, John van der Oost†

†To whom correspondence should be addressed:
J.V.D.O. (john.vanderoost@wur.nl)

This chapter has been published as:

Wu, W. Y., Lebbink, J. H., Kanaar, R., Geijsen, N., & Van Der Oost, J. (2018). Genome editing by natural and engineered CRISPR-associated nucleases. *Nature chemical biology*, 14(7), 642-651.

Abstract

Over the last decade, research on distinct types of CRISPR systems has revealed many structural and functional variations. Recently, several novel types of single-polypeptide CRISPR-associated systems have been discovered including Cas12a/Cpf1 and Cas13a/C2c2. Despite distant similarities to Cas9, these additional systems have unique structural and functional features, providing new opportunities for genome editing applications. Here, relevant fundamental features of natural and engineered CRISPR-Cas variants are compared. Moreover, practical matters are discussed that are essential for dedicated genome editing applications, including nuclease regulation and delivery, target specificity, as well as host repair diversity.

Ever since the discovery of DNA as the carrier of genetic information, researchers have been looking for ways to modify genes and genomes, either for functional analysis or for specific applications. Most directed genetic engineering approaches are based on DNA-targeting enzymes, i.e. deoxyribonucleases that generate double stranded breaks (DSBs) in a sequence specific manner. In the context of a living cell, DSBs are repaired either by non-homologous end joining (NHEJ), by homology directed recombination (HDR), or by variants thereof (see below), potentially leading to the introduction of genome modifications.

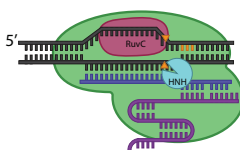
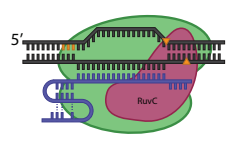
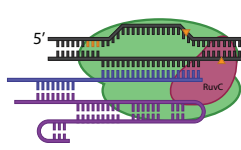
Recently, a class of RNA programmable nucleases with potential for genome editing has been discovered. These nucleases are key players of a system consisting of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins. CRISPR-Cas is an adaptive immune system in prokaryotes that protects against invasions by mobile genetic elements (15). Some Cas nucleases turned out to hold great potential for genome editing, such as Cas9 (72). This single-polypeptide nuclease is guided by two partly complementary RNA molecules, CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA); for practical reasons the two RNAs are generally fused as a single guide RNA (sgRNA) (40). The protein and the two RNAs form a ribonucleoprotein (RNP) complex that is able to recognize a DNA sequence that is complementary to the spacer sequence of the crRNA. After base pairing of the crRNA guide and the target strand, a conformational change of the multi-domain Cas9 protein results in a cleavage-competent state of the nuclease that generates a defined DSB with blunt ends (40, 56). In order to target a specific DNA sequence, an appropriate crRNA guide needs to be generated; adjusting the crRNA guide is relatively easy and inexpensive (72). Apart from Cas9, several distinct types of single-polypeptide CRISPR-associated nucleases have recently been discovered (15). This review will focus on comparing structural and functional features of these natural Cas nucleases and derived variants, as well as on the state of the art with respect to their application in genome editing.

Mechanism of Class 2 CRISPR nucleases

CRISPR-Cas systems are divided in Class 1 and Class 2, each of which is further divided into types and sub-types. This classification is based on major differences between the proteins involved in guide binding and target cleavage. Class 1 systems use crRNA binding Cascade complexes composed of multiple subunits that associate with a nuclease (Cas3 or Cas10). On the other hand, if a single, multi-domain protein is responsible for both guide binding and target cleavage, it belongs to Class 2 (17). The focus of this review is on these single-protein Class 2 systems that have successfully been repurposed for genome engineering. An important practical advantage of DNA targeting Class 2 nucleases generate a DSB through cleavage of both the target strand and the non-target strand. This is in contrast with Class 1 nucleases, which first nick the displaced non-target strand (14, 73).

Class 2 CRISPR-Cas systems currently consist of Types II, V and VI (74). The type II nuclease Cas9 was the first Class 2 effector to be discovered and characterized (19, 40, 75). Consequently, it was the first CRISPR-Cas system to be used for genome editing in mammalian cells (23, 76, 77) and in bacteria (78). Type V includes DNA targeting nucleases Cas12a and Cas12b (previously called Cpf1 and C2c1, respectively) (58, 79), whereas Type VI contains Cas13a and Cas13b (previously called C2c2), which are RNA-guided RNA cleaving nucleases (50, 74). Type V nucleases differ from Cas9 with respect to crRNA guide processing, target recognition and/or target cleavage (Table 1). In contrast to Cas9 and Cas12b, Cas12a can process its own crRNA without requiring a trans-activating crRNA (tracrRNA) (Table 1). Cas12a and Cas12b both recognize a T-rich 5' protospacer adjacent motif (PAM), whereas Cas9 recognizes a G-rich 3' PAM. Cas12a and Cas12b generate 5' staggered ends 17-18nt distal from the PAM, whereas Cas9 generates 5' blunt ends 3nt 5' from the PAM. The molecular basis for these functional differences has been revealed by in-depth biochemical and structural analyses. Integrated molecular analyses revealed that Cas9 possesses two nuclease domains, HNH and RuvC, responsible for cleaving the target and the non-target strand, respectively (40, 56, 80). Cas12a and Cas12b only contain a single nuclease domain (RuvC) that was recently proposed to cleave both DNA strands of the duplex (81-84). The prediction that Cas13 nucleases target RNA rather than DNA, based on the presence of HEPN domains (58), has indeed been confirmed experimentally (50, 74). It is now clear that the different types/subtypes of Class 2 nucleases and their crRNA guide share some general features, but they have distinctive characteristics as well. This implies that each subtype has unique mechanistic features, with potential pros and cons for application in genome editing.

Table 1 | Protein characteristics of class 2 CRISPR nucleases. One variant of each class 2 CRISPR nuclease type was chosen to represent its protein characteristics. *Streptococcus pyogenes* Cas9, SpCas9; *Acidaminococcus* sp. Cas12a, AsCas12a; *Alicyclobacillus acidoterrestris* Cas12b, AaCas12b. Each nuclease contains a crRNA (purple). In addition, some nucleases contain a tracrRNA (violet), which binds to its complementary DNA adjacent to a PAM (orange). N/A, not available.

CRISPR-Class 2			
	Type II-A	Type V-A	Type V-B
			
	SpCas9	AsCas12a	AaCas12b
Type	II-ABC	V-A	V-B
Protein size (aa)	1368	1307	1129
Target	dsDNA	dsDNA	dsDNA
tracrRNA	yes	no	yes
PAM	NGG-3'	5'-TTTN	5'-TTN
Seed (bp)	5	5	N/A
DSB	Blunt (in seed)	4-5 nt 5'overhang	7 nt 5'overhang
Spacer length (bp)	20	23	20

Engineering precision of Class 2 CRISPR nucleases

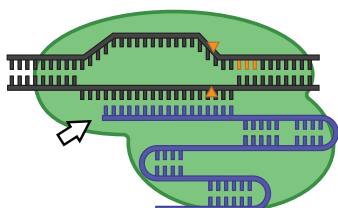
Although mechanistic insights were first obtained for Class 1 Cascade and Cas3 (15), the subsequent unravelling of the molecular details of DNA interference by Cas9 and Cas12a/b (46, 79, 84) as well as their successful functional expression in a wide range of host cells, has allowed for the development of an unprecedented general toolbox for genome editing. For many applications, including gene therapy of human cells, a nuclease should only cleave in case of perfect complementarity between the RNA guide and the DNA target sequence. In case of Cas9 and Cas12a, however, perfect matching is restricted to the seed region of the guide, whereas several mismatches are tolerated in regions more distant from the PAM (85, 86). Seed regions of Class 2 crRNA guides are generally 5 nt long at the PAM proximal end (78, 87, 88).

Off-target cleavage by Cas9 has been reduced by modifying either the sgRNA guide or the Cas9 protein. First, Cas9 with a shortened sgRNA with the variable base-pairing region (spacer) of 17-18 nt appeared to be more specific than a full length (20 nt) sgRNA (Fig. 1A) (89, 90). Second, alternative target sites with more optimal spacer sequences can be used. It has been reported that a U-rich seed sequence shows lower sgRNA expression, meaning lower concentration of active RNP in the cell and thereby leading to a higher specificity (87). Apart from specificity, sgRNA sequences also may influence cleavage efficiency, with preference of a G at position 20, i.e. at the PAM proximal end of the spacer (91). Despite some recently developed algorithms (91, 92), further optimization appears to be required for design of specific guides to reliably target genes of interest. Third, modified SpCas9 nucleases have been engineered to allow reliable genome editing: enhanced specificity Cas9 (eSpCas9), high fidelity Cas9 (Cas9-HF1) and hyper-accurate Cas9 (HypaCas9) (93-95). Using structural insights in protein-DNA interactions, major specificity improvements have been achieved by directed amino acid substitutions resulting in reduced binding affinity towards either the non-target strand (eSpCas9) or the target strand (Cas9-HF1). Just like the effect of a shorter spacer, the rationale behind the designed amino acids substitution was that, in case of attacking off-target sites, a slightly reduced binding affinity results in a subtle shift of the equilibrium from the locally unwound state towards re-hybridization of the two DNA strands, and as such to abortion of the undesired off-targeting. Hence, cleavage only occurs if there is a strong base pairing between an RNA guide and a perfectly matching DNA target (Fig. 1B) (93, 94). HypaCas9 was constructed by introducing mutations in its REC3 domain, which is involved in RNA/DNA duplex recognition that triggers repositioning of the HNH nuclease domain in its cleavage compatible state. In case of guide/target mismatches the HNH domain remains locked in its inactive state and no cleavage will occur (95). By comparing HypaCas9 with the previous eSpCas9 and Cas9-HF, it was found that all three have comparable specificity.

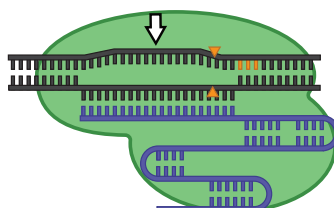
In addition, Cas9 nickases (Cas9n) were developed in which the active site of either the RuvC or the HNH domain is inactivated (Fig. 1C). Both types of Cas9n have been demonstrated to only nick one DNA strand, instead of generating a DSB. DSBs are generated only in case of a Cas9n pair with two guides that target opposite strands in close proximity (85, 96). Targeting a certain DNA sequence by using two guides, implies double selection for specificity; indeed, off-target cleavage was significantly reduced. A similar strategy has been applied to a catalytically inactive, dead Cas9 (dCas9), fused to the catalytic domain of the FokI restriction enzyme, also known as RNA-guided FokI nucleases (RNF). A DSB only occurs upon dimerization of the FokI domains, when a pair of RNFs target their complementary strands (Figure 1D) (97). RNFs have shown to have slightly higher specificity when compared to nCas9s. Alternatively to the methods mentioned above, more specific CRISPR nucleases, either natural Cas9 variants or type V nucleases, can potentially be used to reduce off-target cleavage. Cas12a for instance has been reported to be more specific than Cas9 by having the first 18nt adjacent to the PAM being highly mismatch intolerable (77, 98). Gene targeting by AsCas12a and LbCas12a did not result in off-target indel

formation for more than half of the crRNAs tested, and very few cases of undesired cleavage for the remaining guides, suggesting that these Cas12a nucleases are more specific than SpCas9 (77). Remarkably, Cas12b has been reported to not tolerate any mismatches *in vitro* (83), although *in vivo* genome wide target/non-target analysis is required to validate the high specificity of this nuclease.

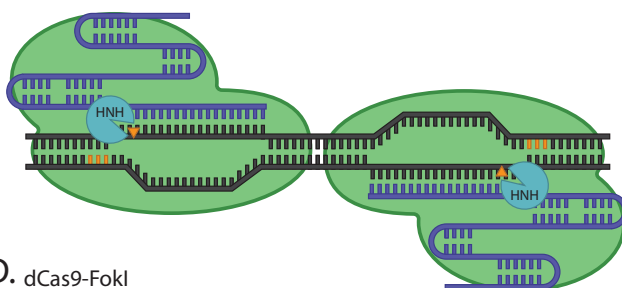
A. Truncated sgRNA-Cas9



B. eSpCas9



C. Dual nCas9



D. dCas9-FokI

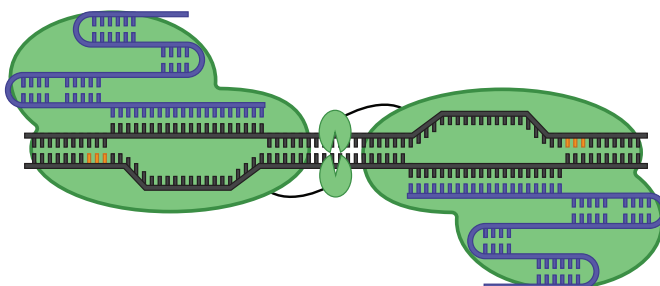


Figure 1 | Modified Cas9 or sgRNA, used to reduce off-targets. Orange arrow indicates cleavage site and * indicates a mutation in a domain resulting in no activity of that domain. **(A)** sgRNA is truncated from 20 nt to 17-18 nt. **(B)** Enhanced specificity Cas9 (eSpCas9). Charged residues are swapped with neutral residues to reduce binding affinity to the non-target strand. **(C)** Dual nickase Cas9 (nCas9) approach in which a DSB is created by two nCas9-sgRNA complexes nicking only the target strand. **(D)** Dead Cas9 fused to FokI domain (dCas9-FokI), when FokI forms a dimer and a DSB is generated.

Cas9 regulation *in vivo*

Another way of reducing off-target cleavage is to control the amount of active ribonucleoprotein (Cas9-sgRNA complex) within the cell (85, 86). Therefore, different strategies have been developed to regulate RNPs. For example, the amount of active sgRNA can be tuned using small molecules (e.g. Theophylline or Guanine). In this system, an sgRNA is bound to a ligand-inducible self-cleaving RNA (aptazyme guide RNA, agRNA), that base-pairs with the sgRNA, thus preventing target DNA binding. Upon binding to a ligand, the aptazyme will cleave itself and separate from the sgRNA, resulting in an active, DNA-targeting sgRNA. (Fig. 2a) (99). The sgRNA guide can also be used to induce dimerization of two halves of a Cas9 protein (split Cas9), resulting in an active RNP (Fig. 2b) (99). In addition, split Cas9 dimerization can also be induced with rapamycin, by fusing the C-terminal Cas9 lobe to a FK506 binding protein (FKBP) and the N-terminal Cas9 lobe to a FKBP rapamycin binding (FRB) domain (Fig. 2c), implying that the presence of rapamycin triggers assembly of the two Cas9 lobes (100). In addition, methods have been established that are based on activity induction of the intact Cas9 complex. An example is intein-Cas9, in which a fusion of Cas9 to a ligand-dependent self-splicing protein domain (intein). Splicing occurs in the presence of the ligand (4-hydroxytamoxifen, 4-HT), restoring Cas9 activity (Fig. 2d) (101). In addition, an inactive Cas9 has been constructed by substituting an essential lysine residue by a caged lysine (pyrrolysine). The pyrrolysine is converted back to lysine upon UV exposure, which restores Cas9 activity (Fig. 2e) (102).

Off-target activity can also be limited by inactivating Cas9 as soon as possible after target site cleavage has occurred. One way of removing active Cas9 proteins from the cell is the use of the Self-Limiting Circuit for Enhanced Safety and Specificity (SLICES) approach. SLICES works by co-expressing Cas9 with two guides, one targeting the gene of interest and another one auto-targeting the cas9 gene (103). A limitation of the SLICES approach is that it can only be used when Cas9 is delivered as DNA.

An approach to directly disrupt Cas9 nuclease activity at protein level would be the (appropriately timed) delivery of anti-CRISPR proteins (104). In addition, alternative methods are available that lead to reversible (in)activation of Cas9. An allosterically regulated Cas9 (arCas9) has been constructed by inserting the ligand-binding domain of the estrogen receptor- α into Cas9, rendering Cas9 inactive. Upon addition of 4-hydroxytamoxifen (4-HT), this ligand binds to the receptor domain causing a conformational change that results in activation of arCas9 (Fig. 3a) (105). Inactivation is achieved by transferring cells to 4HT-free medium. Another example is iCas, which is also based on a fusion of Cas9 and an estrogen receptor (ERT2) that is activated by 4-HT. In this case, the presence of 4-HT leads to nuclear localization of iCas, which otherwise remains in the cytoplasm (106).

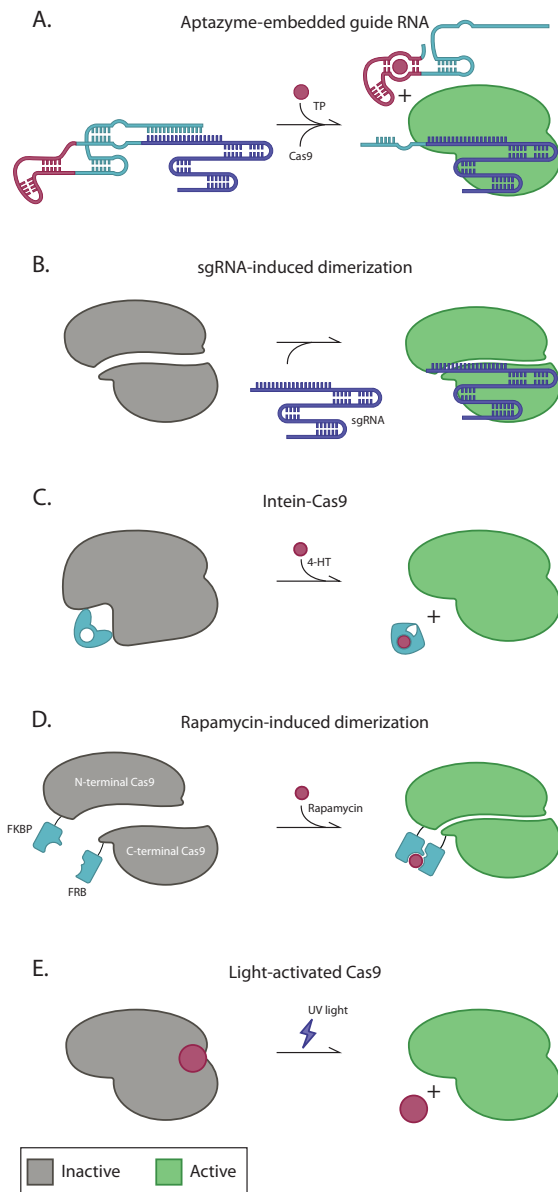


Figure 2 | Modified Cas9 or sgRNA for regulation of active RNP (irreversible). Gray Cas9 = inactive and green Cas9 = active. Arrows indicate if reactions are reversible or not. **(A)** Aptazyme embedded to an sgRNA (purple) that binds and covers the sgRNA. Theophylline (TP) (red) binds to the aptazyme, cleaves it and leaves the sgRNA. **(B)** Split Cas9s (grey) are dimerized by the addition of an sgRNA (purple). **(C)** N-terminal and C-terminal split Cas9s are fused to FKBP and FRB domains (blue) respectively. Rapamycin (red) binds to both domains, leading to split Cas9 dimerization. **(D)** A modified intein (blue) is fused to Cas9 and is spliced out when bound with 4-HT (red). **(E)** Cas9 contains a caged lysine amino acid (red), rendering it inactive. This is removed when exposed to UV light.

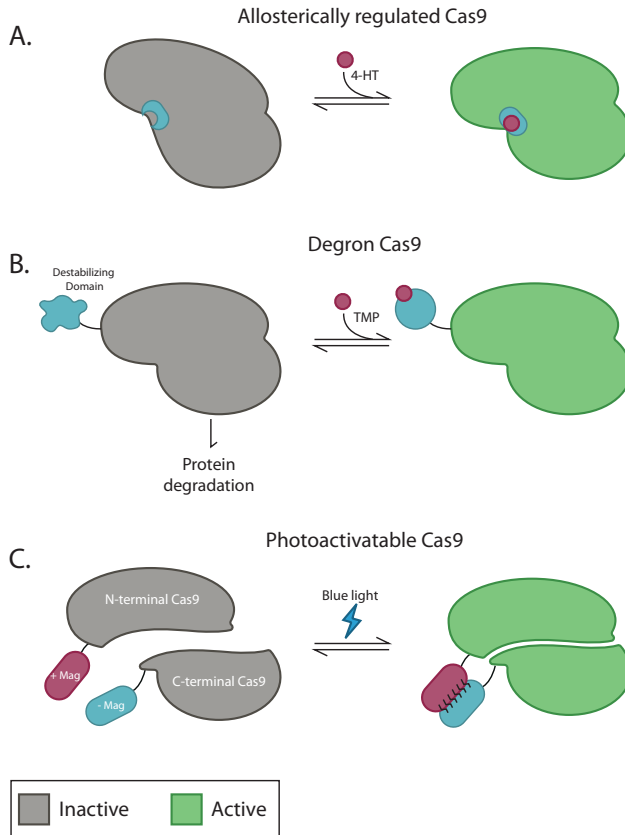


Figure 3 | Modified Cas9 or sgRNA for regulation of active RNP (reversible). Gray Cas9 = inactive and green Cas9 = active. Arrows indicate if reactions are reversible or not. **(A)** The protein conformation of Cas9 bound to an estrogen binding domain (blue) is inactive. Upon binding to 4-HT (red), Cas9 undergoes a conformational change and becomes active. **(B)** Cas9 is bound to a destabilizing domain (blue), which constantly undergoes protein degradation. Trimethoprim (TMP) (red) binds to stabilize the destabilizing domain, inhibiting protein degradation. **(C)** N-terminal and C-terminal split Cas9s are fused to positive (red) and negative (blue) magnet domains respectively. Blue light exposure makes the magnets bind to each other, leading to split Cas9 dimerization.

Yet another variant is degron Cas9, which is based on a fusion of Cas9 to a destabilizing dihydrofolate reductase (DHFR) domain (Fig. 3b). If not stabilized, degron Cas9 is rapidly degraded by proteases. Degron Cas9 activity can thus be modulated by the presence or absence of trimethoprim (TMP) that can bind to DHFR and stabilize degron Cas9 (107). Moreover, a photo-activatable Cas9 (paCas9) has been developed, in which blue light induces dimerization of split Cas9. The N-terminal Cas9 and C-terminal Cas9 are fused photo-inducible dimerization domains ('magnets') (Fig. 3c) (108). Although this design resembles the rapamycin-induced Cas9 (Fig. 2c), the reversibility of the light-dependent dimerization can be

controlled much easier. Recently, the real-time activation/deactivation dynamics of some of these tunable synthetic Cas9 variants was reported (106, 109). This was done using a droplet digital PCR assay for double-strand breaks (DSB-ddPCR), which measures DSBs and repair *in vivo* (109).

Multiplex genome editing

Another challenge for genome editing concerns the co-expression of a CRISPR-associated nuclease with different guide RNAs to perform multiplexing, i.e. to target multiple genes simultaneously. The first Cas9-based multiplex approaches in bacteria (78) and in mammalian cells (23, 76), were based on the simultaneous assembly of Cas9 complexes with different sgRNAs (crRNA fused to tracrRNA), each of which was transcribed as an individual transcription unit (promoter-guide gene-terminator) (23, 76). Alternatively, two methods have been reported to generate multiple mature sgRNA guides from a single precursor crRNA. One method relies on a cleavage site that can be recognized and cleaved by Csy4, a Class 1/Type I-F CRISPR-associated ribonuclease that should be co-expressed (97). In another method, a DNA construct has been designed in which a tRNA gene is positioned in between two sgRNA genes, resulting in processing of the transcript by endogenous RNase P and RNase Z, and release of functional sgRNAs in plants (110). Whereas Cas12b also relies on a tracrRNA and processing by RNaseIII, Cas12a systems do not possess tracrRNA (79)(Table 1). Cas12a is unique in that it possesses a domain that auto-catalyzes specific cleavage of its precursor crRNA to yield mature crRNA guides (45, 46). Multiplexing was shown for Cas12a for up to 4 genes in human cells (HEK 293T; all 4 genes targeted in 6.4% of the transformed cells) and 3 genes in the brains of living mice (all 3 genes targeted in 16.9% of the transfected cells) (46). Cas12a-based multiplex genomic recombination has recently been observed in yeast, in which knockouts of 4 genes were obtained simultaneously with 100% efficiency (111).

Delivery of gene editing systems

Improved efficacy and specificity of CRISPR systems, bring clinical application within reach, but delivery of CRISPR effectors remains a hurdle. Ideally, efficient cell targeting is combined with minimal cytotoxicity and rapid clearing of the CRISPR system after successful gene modification. However, none of the currently available delivery methods fulfils all above criteria.


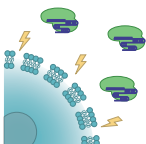
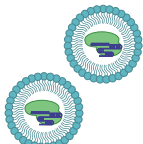
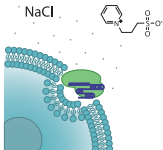
Adeno-Associated Viruses (AAV) combine low immunogenicity, low oncogenic risk and serotype-related target-cell specificity (112-115), and the use of cell-specific promoters can further restrict Cas9 expression (Table 2). However, the maximum cargo size that AAV can deliver (approximately 4.5 kb) impedes the packaging of the most commonly used CRISPR-associated nuclease genes into a single vector. A smaller version of Cas9 from *Staphylococcus aureus* (SaCas9) (116) is as effective as SpCas9 and can be appropriately packaged into an AAV vector (Table 2). The small SaCas9 has been used to restore Dystrophin expression in the skeletal muscle of a Duchenne mouse model (117, 118), but efficient gene editing was only achieved when the AAV particles were injected directly into the target muscle, an approach that undercuts the main advantage of AAV as systemic delivery tool. SpCas9 can also be split into two lobes and reconstituted intracellularly to yield a fully functional protein (100, 108, 119-123), but this approach may reduce overall efficacy.

An important drawback of viral delivery systems is the long-term presence (for months or even years (124)), which could lead to accumulated off-target cleavage. Delivery of preformed ribonucleoprotein complexes (RNPs) may provide better temporal control of CRISPR activity. Their rapid degradation (within several hours) assures a short window of activity, yet with higher editing efficiency than RNA or DNA-based delivery method (125-127). Fusion of the Cas9 protein with a cell-penetrating peptide (CPP) enhances intracellular delivery (128), but the efficiency varies between cell types (128-130). Electroporation is more widely used for RNP delivery (Table 2) (127, 131) and is a clinically accepted method for introducing large molecules into cells, and has successfully introduced CRISPR/Cas9 plasmid DNA into the skeletal muscle of a murine model of Duchenne Muscular Dystrophy (Table 2) (132, 133)133</style>. However, tissue damage caused by the electroporation process can be an obstacle for broad *in vivo* application (133).

Polymer (134, 135), lipid (136, 137) or DNA-based nanoparticles (138) are an alternative means for the intracellular delivery of RNA, DNA or RNP complex gene editing systems (Table 2). Lipid nanoparticles home effectively in the liver and allowed for repairing a murine model of hereditary tyrosinemia (139) (Table 2). Nanoparticle delivery of Cas9 RNA combined together with AAV to deliver the sgRNA and repair template DNA resulted in impressive homologous-recombination-mediated gene repair in more than 6% of the hepatocytes upon systemic application (139). Repaired cells have a competitive survival and proliferation advantage, implying that this efficiency may be therapeutically relevant. Similarly, lipid particle-mediated RNP delivery was shown to allow for transduction of a variety of cell types *in vitro* and achieving gene editing of inner ear cells *in vivo* (126). However, efficiency of lipid-based transfection reagents is tissue type dependent (140), and lipid nanoparticles have been reported to be immunogenic (141, 142). A new method, induced transduction by osmocytosis and propanebetaine (iTOP), allows efficient delivery of CRISPR/Cas9 into a wide variety of primary cell types (Table 2) (125). The iTOP approach allows for virus-free transduction of native proteins and does not rely on additional peptide tags, which may interfere with protein function or editing efficiency and is particularly effective

for transduction of cell types that are refractory to other delivery methods (125, 143). Finally, a new delivery method based on a gold nanoparticle/ DNA scaffold was reported (144). This CRISPR-Gold system simultaneously delivers CRISPR/Cas9 RNP and repair template DNA into skeletal muscle, allowing homology-directed repair of a dystrophin point mutation, albeit at low efficiency (5.4%) (144). However, the complexity and cost of the CRISPR-Gold particle may hamper wide scale adoption in research.

Table 2 | Methods for intracellular delivery of CRISPR editing system

Delivery method	Demonstrated applications	Form	Temporal presence	Main advantages (+) and limitations (-)
Adeno-associated virus (AAV) 	In vitro, ex-vivo and in vivo	Single-stranded DNA	Long-term (years in non-dividing cells)	+ Broad tissue tropism dependent on viral serotype + Low immunogenicity - Restricted packaging capacity (± 5 kb) hampers use with most CRISPR system - Possibility of genomic integrations long term presence of CRISPR system inside target and long-term effects of AAV require further analyses
Electroporation 	In vitro, ex-vivo and in vivo	DNA, RNA or RNP	Hours for RNP, days-weeks for RNA/DNA	+ Clinical-grade electroporation protocols available + High efficiency gene editing in <i>in vitro</i> and <i>ex-vivo</i> applications - Cytotoxicity and limited <i>in vivo</i> applicability
Nanoparticles 	In vitro and in vivo	DNA, RNA or RNP	Hours for RNP, days-weeks for RNA/DNA	+ Efficient <i>in vitro</i> delivery in commonly used cell lines + reported <i>in vivo</i> delivery sgRNA and template DNA into hepatocytes and delivery on RNP into inner ear cells + Simultaneous <i>in vivo</i> delivery of CRISPR/Cas9 RNP and repair template DNA in skeletal muscle - Immunogenicity and toxicity have been reported
iTOP 	In vitro and ex vivo	RNP	Hours	+ Developed for RNP transduction + Efficient transduction of primary (stem) cells that are refractory to other delivery methods - Does not allow delivery of plasmid DNA

Repair of CRISPR-induced DNA breaks

CRISPR-induced DSBs will be repaired by one of the cellular DSB repair pathways. Understanding the mechanistic details of these distinct pathways are important to guide the optimal design of targeting constructs to efficiently obtain the intended genome modification. Repair pathway choice depends on the presence of a donor repair template and the form in which this is delivered, on the kind of break introduced into the target DNA, but also on parameters such as genomic locus, cell cycle phase and cell type.

DSBs can be repaired efficiently through canonical non-homologous end joining (C-NHEJ) (Fig. 4a left) (145), which does not require sequence homology and is active throughout the G1, S and G2 phases of the cell cycle. Repair through C-NHEJ can occur in an error-free manner (146), however, a restored original sequence can be re-cleaved by the CRISPR nuclease. During error-prone repair, small insertions, or deletions (indels) often result in frame shift mutations, and (in case of Cas9) in destruction of the nuclease target site. Mutations can also be created after limited (enzyme-mediated) editing of the DSB, and error-prone repair may occur when sequence micro-homologies are used by the alternative end-joining pathway (Alt-EJ) (Fig. 4a middle). The inclusion of non-homologous donor DNA increases editing efficiencies (Fig. 4a right). Because these repair events are error-prone it is hard to control the identity of the eventual genomic mutation. Nevertheless, clever donor template design allowed EJ-mediated creation of in-frame fusion genes with techniques such as CRISPR/Cas9-mediated Precise Integration into Target Chromosome (CRIS-PITCh) (147) and homology-independent targeted integration (HITI) (148). Notably, HITI has been used successfully for efficient transgene insertion in non-dividing cells, both *in vitro* and *in vivo* (148).

When aiming for gene variants with single point mutations or for integrating complete genes, more precise surgery is required. Besides accurate targeting by Cas nucleases, this requires the engagement of error-free DNA repair pathways through homology-directed repair (HDR). Traditionally, knock-out and knock-in mutations were made via homologous recombination (HR) through the DSB repair (DSBR) sub-pathway by introducing a double-stranded DNA repair template with long homology arms (Fig. 4b left). The efficacy of this procedure has been significantly improved by employing CRISPR-Cas to generate specific DSBs. Small insertions and point mutations can also be introduced using single stranded DNA oligonucleotides (ssODN) via Synthesis-Dependent Strand Annealing (SDSA) (149) (Fig. 4b middle) or the Single Strand Annealing (SSA)-like pathway (150) (Fig. 4b right). The concomitant introduction of blocking mutations that destroy the seed region and/or the PAM motif in the genomic DNA prevents recurrent targeting by the CRISPR nuclease and reduces the frequency of undesired indels (151). Removal of the blocking mutation, resulting in scar-less editing, can be achieved via subsequent rounds ('re-guide' or 're-Cas' approaches) (151). Targeting efficiency can be increased through design

of ssODN donors complementary to the 3'-end of the non-target strand, which is asymmetrically released by Cas9 prior to complete dissociation (152). The use of exonuclease-resistant phosphorothioate-modified oligonucleotides allows for incorporation of larger insertions up to 100 bp in length (150). Upon the introduction of a DSB, repair proceeds mainly via SDSA (149). If the initiating lesion is a single strand nick (for example created by nCas9), repair occurs via SDSA or SSA depending on whether a double stranded DNA donor or a ssODN complementary to the target or the non-target strand is provided (149, 153). Note that in the SSA-like pathway, the ssODN becomes physically incorporated into the genome, while during SDSA, the ssODN is only used as template to direct nascent DNA synthesis (149) (Fig. 4b). Because both SDSA and SSA pathways involve short gene conversion tracts, it is critical that knock-in mutations are placed within the effective conversion zones, which are different for both pathways (149). In fact, this phenomenon can be exploited via distance-dependent suboptimal mutation incorporation to create monoallelic variants (151).

In mammalian cells, homologous integration of a donor construct is rare because random integration is orders of a magnitude more efficient. A potential reason is that EJ-based pathways are more efficient than HDR pathways and can operate throughout the cell cycle, while HR is normally limited to S and G2 phases. Indeed, targeting efficiency through HDR can be increased by controlled timing of CRISPR-Cas9 RNP delivery to synchronized cells (154), by synchronization of Cas9 expression with cell cycle progression through fusion of Cas9 with the N-terminus of geminin (present only in S, G2 and M phase cells) (155), or by activating HR in G1 cells through restoring DNA-end resection and an S-phase specific protein repair complex (156). Simultaneous inactivation of C-NHEJ and an Alt-EJ pathway mediated by DNA polymerase theta completely eliminates all random integrations, without affecting homologous integrations (157). Thus (pharmacological) suppression of EJ-based pathways may provide an additional means to reduce off-targets effects, which will be of utmost importance for clinical applications of CRISPR-Cas technologies.

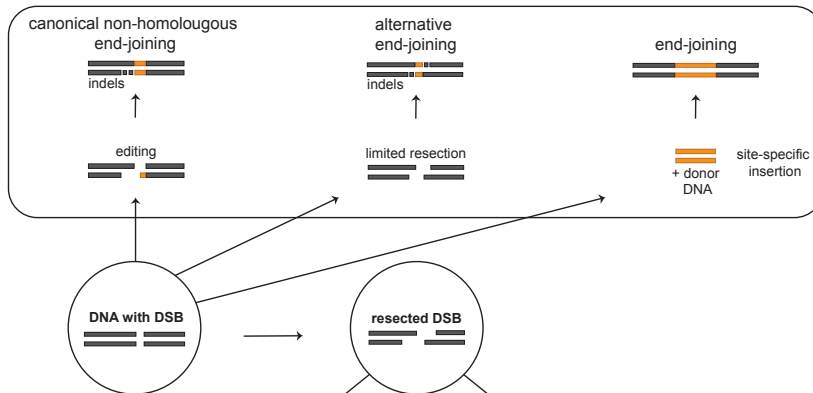
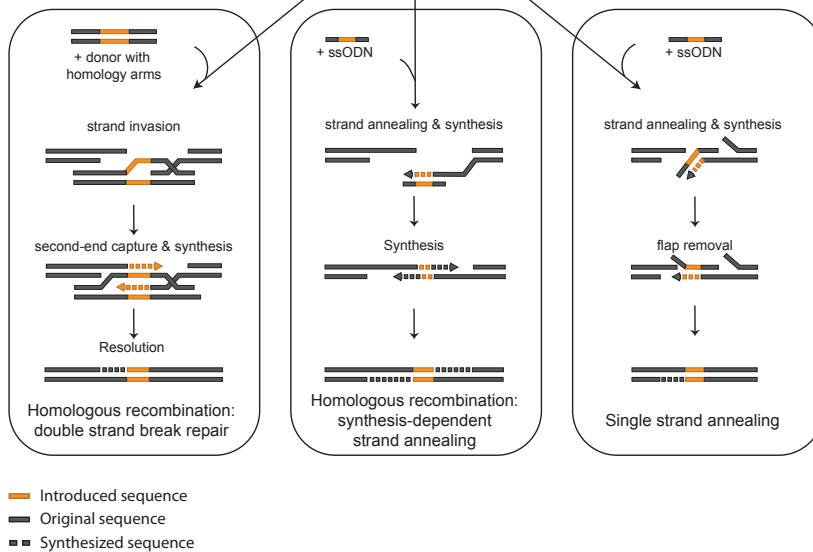
A. Non-Homologous End-Joining**B. Homology Directed Repair**

Figure 4 / Engineering the genome from a CRISPR-Cas9 induced targeted DNA double-strand break using distinct DSB repair pathways. Top panel **(A)**: the end-joining (EJ) pathways, that are used for repair of a CRISPR-Cas9 induced, which leads to targeted but unspecified mutations. Left: C-NHEJ occurs in the absence of DSB resection, with the signature of the joint either having no or very little (<5 nt) homology. Middle: Alt-EJ uses micro-homologies (up to ~ 25 nt) formed upon limited editing of the DSB. Right: Targeting efficiency via end-joining can be increased upon introduction of extra-chromosomal DNA, although the nucleotide sequence at the borders cannot be precisely controlled. Lower panels: homology-directed repair (HDR) pathways that are harnessed after more extensive resection of the DSB and require (partially) homologous donor DNA for repair. **(B)** Left panel: Introducing donor DNA with homology arms allows the DSB-like HR pathway to generate precise insertions, deletions or point mutations. During this process, both DNA ends at the break engage the template DNA and after DNA synthesis Holliday junctions are resolved into products. Middle panel: ssODNs can be used as templates for mutations introduced through SDSA. During SDSA one-ended invasion of the broken DNA is followed by DNA synthesis. The newly synthesized DNA is complementary to the other end of the DSB, which can now be engaged by annealing for further synthesis. Right panel: The SSA-like pathway also uses ssODNs that become incorporated in the genome by annealing to homologous resected regions around the DSB.

Single Base editing

In previous sections, CRISPR Class 2 nucleases have been discussed with respect to their ability to generate DSBs or nicks, which steer the mutagenic outcome via end-joining or homology directed repair pathways. Recently, precise gene editing tools have been developed to modify specific DNA bases at target sites, circumventing DSB generation and a donor repair template. These base editing tools are based on fusions of Cas9 variants and specific nucleotide-converting enzymes (Fig. 5) (158, 159). Cytidine deaminase (AID) catalyzes the irreversible deamination of cytosine (C) to uracil (U). When fused to Cas9 and an appropriate guide RNA, the desired conversion occurs within a 5-nucleotide window at the non-target strand of the selected genetic locus. Single base editing tools such as Base Editor 3 (BE3) (159) and Target-AID (160) were developed by specifically fusing domains of both AID and a uracil DNA glycosylase inhibitor (UGI) to a Cas9 nickase. Because of repair pathway management, this protein combination resulted in highly efficient base-editing with the desired C→T (non-target strand) and G→A (target strand) substitutions (Fig. 5a). UGI blocks the uracil DNA glycosylase repair pathway, which otherwise would remove the uracil and restore the original C-G base pair (159, 160). The nCas9 nicks the (non-edited) target strand which contains G opposite the uridine. The DNA mismatch repair pathway, activated by the G-U mismatch, removes the nicked DNA fragment containing the original G and replicates with an A opposite U, effectively fixing the edited base change as a stable substitution without requiring the cells having to cycle through S-phase for DNA replication (159, 160). BE3 and Target-AID are able to deaminate C bases at position 4-8 in the target site (at the distal end of the PAM, position 21-23) and 2-5 respectively. Single Base editing can be used to introduce early STOP codons to create gene knockout (CRISPR-STOP) or to incorporate single amino acid changes (161). Recently, the BE3 toolbox has been expanded to target different PAMs and optimized towards a narrower base editing window of 1-2 nucleotides (162). Moreover, a DNA and RNA adenine deamination tool has recently been developed, known as Adenine Base Editor (ABE) and RNA Editing for Programmable A to I Replacement (REPAIR), respectively. Both systems utilize nCas9 fused to an adenine deaminase to convert adenine (A) to inosine (I), which in DNA is further recognized as G by the transcription and replication machineries, resulting in A→G (non-target strand) and T→C (target strand) substitution (Fig. 5b) (163, 164).

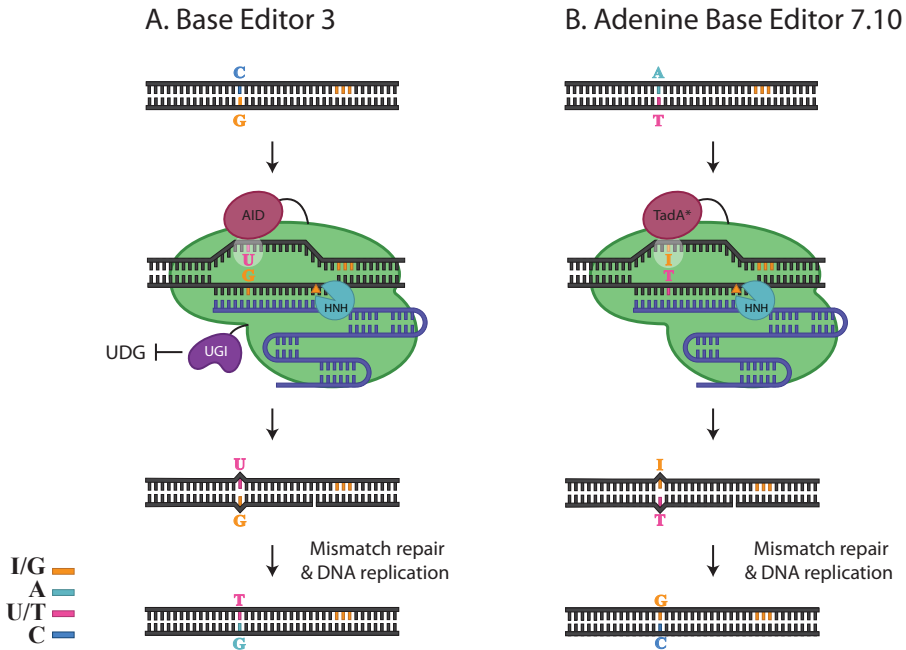


Figure 5 | Guided base editing using dead Cas9. (A) A cytosine (C) is deaminized to uracil (U) by Base Editor 3, which consists of a nCas9 (green) nicking the target strand, an AID (red) deaminizing a C within a given window (transparent white box) and an Uracil Glycosylase Inhibitor (UGI) (purple) inhibiting base excision repair. Mismatch repair will convert U-G to U-A; then a new cycle replication will produce daughters with U-A and T-A. **(B)** An adenine (A) is deaminized by Base Editor 7.10, which consists of a nCas9 (green) nicking the target strand, a mutated (*) TadA (red) deaminizing an A within a given window (transparent white box). Mismatch repair will convert I-T to I-C; then S-phase and replication will produce daughters with I-C and G-C.

Future prospects & concluding remarks

Insights into the molecular mechanism of Class 2 CRISPR-Cas nucleases have led to their repurposing into state-of-the-art genome editing tools. Due to their successful heterologous expression in cells from a wide range of organisms and the fast and cost-effective adjustment of their specificity, the CRISPR-associated nucleases have rapidly reached a status of generic applicability. Apart from the initially developed Cas9 system, alternative natural CRISPR-associated nuclease variants have recently been characterized and utilized for genetic engineering. Realization of the great promise that these genome editing tools hold for gene therapy applications, however, still requires improvement of their precision, efficacy, and delivery. Targeting precision is currently being improved by using different natural types of

Cas nucleases, by engineering variant nucleases, and by regulating RNP activity in the cell. The elucidation of general rules for selection of high-efficiency guide/target sequence pairs will benefit from cleavage efficiency studies using different nuclease types. Furthermore, the efficiency of obtaining the desired DNA sequence modification can be optimized through DNA repair pathway management. This can be achieved by selecting optimal donor template DNAs or even in the absence of any donor DNA. The latter approach employs synthetic chimeric CRISPR nucleases with innovative functionalities, such as guided base editing. In addition, delivery still is a major bottle neck for CRISPR-based gene therapy. Currently the CRISPR toolbox is being expanded very fast, not in the least because of a series of smart synthetic chimeras that has resulted in CRISPR nucleases with a wide range of innovative functionalities. All in all, the CRISPR revolution continues and will enable many spectacular applications in the near future.

Acknowledgements

J.v.d.O is supported by Netherlands Organization for Scientific Research (NWO) through a TOP grant (714.015.001). The work of J.L and R.K is part of the Onco Institute which is partly financed by the Dutch Cancer Society and was funded by the gravitation program CancerGenomiCs.nl from the Netherlands Organisation for Scientific Research (NWO). The work of N.G. is supported in part by Stichting Singelzwim Utrecht, Stichting FSHD and TKI/ Health Holland.

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
A	T	G	G	A	A	A	A
t	G	A	A	T	A	A	G
T	A	A	A	A	G	T	C

A T A A T T T C
T C H A P T E R
T T A T A T T T
G C C A C G T 3
T A A A A C T T

Adaptation in type V-A and type V-B CRISPR-Cas systems

Wen Y. Wu, Simon A. Jackson, Cristóbal Almendros, Suzan Yilmaz, Rob Joosten, Stan J.J. Brouns,
John van der Oost, Raymond H.J. Staals[†]

Manuscript in preparation

[†]To whom correspondence should be addressed: R.H.J.S. (raymond.staals@wur.nl)

Abstract

Adaptation (the acquisition of a new spacer in the CRISPR array) is an important step in the CRISPR-Cas system, as it determines towards which mobile genetic element(s) immunity is aimed. Adaptation has been well studied in class 1 systems, but not in class 2 systems. In this study, we explore the adaptation mechanisms in two type V systems: type V-A from *Francisella tularensis* subsp. *novicida* U112 and type V-B from *Alicyclobacillus acidoterrestris* ATCC 49025. Their respective CRISPR-Cas loci were heterologously expressed in *E. coli*, after which newly acquired spacers were obtained by PCR amplification and analyzed by deep sequencing. The results indicated that although adaptation occurred, spacers were acquired with non-canonical PAMs. Closer inspection of the native genes encoding Cas4 (involved in adaptation) revealed that they were truncated at the N-terminus, providing a possible explanation for the aberrant PAM selection. We confirmed this hypothesis, by removing the truncations and repeating the experiment on a smaller scale. Furthermore, we found that the adaptation mechanism in type V-A and V-B distinct to that of type II-A. In type V-A, only Cas1 and Cas2 are required for adaptation, whereas in type V-B, Cas4/1 and Cas2 are required for adaptation, but Cas4 activity is not. Spacers acquired without a functional Cas4 target protospacers containing mostly non-canonical PAMs. Thus, Cas4 activity ensures for PAM selection and acquisition of correct spacers in both type V-A and V-B.

Introduction

Bacteria and Archaea are constantly being challenged by mobile genetic elements (MGE). To combat these MGEs, these organisms have developed innate and adaptive immune systems. An example of the latter is the CRISPR-Cas system. CRISPR-Cas consists of a CRISPR array (Clustered Regularly Interspaced Short Palindromic Repeats) and its corresponding *cas* (CRISPR associated) genes. Using these two components, CRISPR-Cas protects the cell by cleaving invader double stranded DNA (dsDNA) using an RNA guide. To achieve this, the CRISPR-Cas system must first go through several steps: adaptation, expression and interference (15).

Firstly, during adaptation, immunity is acquired and occurs when a short piece of dsDNA, also known as pre-spacer, is integrated into the CRISPR array. After integrating into the CRISPR array, the pre-spacer is then termed spacer (22). Adaptation can be split into two categories, naïve adaptation and primed adaptation. Naïve adaptation occurs when no pre-existing spacer exists against a specific MGE, whereas primed adaptation occurs when a matching or partially matching spacer already exists in the CRISPR array (35, 36, 165). The rate of spacer acquisition during primed adaptation is much faster than that during naïve adaptation, since immunity has been previously obtained against the attacking MGE (35). In case of type I-E, cleavage by Cas3 generates short DNA degradation products, which are then used for acquisition of new spacers (36). The spacers acquired from primed adaptation are often found nearby the protospacer of the previously obtained spacer (166). Secondly, during expression, the *cas* genes are expressed and translated and the CRISPR array is transcribed into a long pre-CRISPR RNA transcript (pre-crRNA) and then processed into individual mature crRNAs (14, 167, 168). Lastly, during interference, mature crRNAs are bound to Cas effector proteins to form a ribonucleoprotein complex. The complex first searches for a protospacer adjacent motif (PAM). A PAM allows for distinguishing between self and non-self, as a PAM is only present on the DNA target (protospacer) and not on the CRISPR array (169). Once a PAM is found and the protospacer matches, the effector protein cleaves targeted protospacer to generate a double stranded break and eliminates the MGE (20).

Throughout different CRISPR-Cas systems, *cas1* and *cas2* were found to be the most conserved genes in all CRISPR-Cas systems and are strictly required for adaptation (22). Cas1 and Cas2 forms a complex (two Cas1 homodimers bridged by one Cas2 homodimer) to take up pre-spacer dsDNA and integrating it at the leader proximal end of the CRISPR array (24). Apart from Cas1 and Cas2, in type I and II systems, others proteins were also found to be involved in adaptation, such as Cas4, Csn2 and Cas9 (29, 30, 170). Though not required, Cas4 was found to aid adaptation in enhancing adaptation, PAM determination, spacer trimming and spacer orientation (26-28). The functionality of Cas4 can differ between systems, meaning it is hard to accurately predict the role of Cas4 without experimental testing it. On the contrary, both Csn2 and Cas9 were found to be required for adaptation

in type II-A systems. Csn2 was hypothesized to stabilize the adaptation complex and Cas9 was found to be required for PAM recognition (29, 30, 171). To date, most adaptation research has been done in class 1 systems whereas very little is known about class 2 systems apart from type II-A, II-C and V-C (172-174). In this work we elucidate the adaptation mechanism of two class 2 CRISPR-Cas systems, type V-A and type V-B from *Francisella tularensis* subsp. *novicida* U112 and *Alicyclobacillus acidoterrestris* ATCC 49025, respectively. The V-A CRISPR locus contains *cas12a*, *cas4*, *cas1* and *cas2* whereas V-B contains *cas12b*, *cas4/1* (one gene consisting of *cas4* and *cas1* domain) fusion and *cas2* (Fig. 1) (64). Previous studies showed that Cas12a and Cas12b recognized a 5'-NTTN PAM and cleave dsDNA at the PAM distal end to generate staggered ends (43). However, Cas12a generates 4-5nt staggered ends, whereas Cas12b generates 7 nt staggered ends (58). In addition, Cas12a can process its own crRNA whereas Cas12b requires a tracrRNA like that of Cas9 (Fig. 1). To study adaptation, both systems type V-A and V-B CRISPR-Cas systems were overexpressed in *E. coli* to determine functionality of the individual cas genes involved for naïve adaptation. Although primed adaptation was first thought to occur in class 1 systems exclusively, a primed adaptation setup was also performed for both systems due to recent studies indicating that this can also occur in class 2 type II-A systems (37, 166).

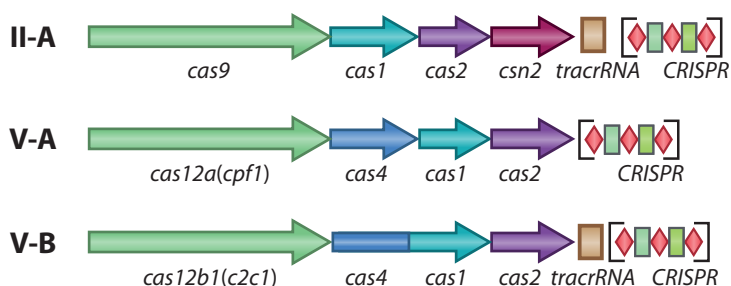


Figure 1 | Schematic CRISPR-Cas loci of type II-A, V-A and V-B systems. Effector genes are indicated in green. *cas1*, *cas4*, *cas2* and *csn2* are indicated in light blue, dark blue, purple and bordeaux, respectively. II-A and V-B also contain a *tracrRNA* (brown). The CRISPR array consists of repeats (red diamond) and spacers (green).

Results

CRISPR-Cas loci were overexpressed in *Escherichia coli* using a three-plasmid setup (Fig. 2A): pCas_adaptation, pCas_effector and pTarget. pCas_adaptation expressed the genes predicted to be involved in adaptation such as *cas4*, *cas1* and *cas2* for V-A and *cas4/1* and *cas2* for V-B. pCas_effector expressed the effector protein Cas12a and Cas12b for V-A and V-B, respectively. Lastly, pTarget was used to mimic the MGE, as a source for spacers by the adaptation machinery and was not selected against during the growth experiment. Variations of pCas_adaptation and pCas_effector plasmids were constructed to test *cas* genes functionality individually within naïve adaptation. To study whether *cas4* was involved in adaptation, *cas4* was either knocked out or made catalytically inactive (V-A: K70A, V-B: K81A) (175). Additionally, the *cas4* domain of type V-B was swapped with the closely-related Cas4 domain from type I-U to test whether adaptation still occurs, since they are found to be closely related, (176). To investigate the role of the different catalytic domains in the effector nuclease Cas12, *cas12a/b* were either mutated in the RuvC domain (V-A: D917A & E1006A, V-B: E848A) or the PAM interaction (PI) domain (V-A: K613A & K671A, V-B: R122A & G143P) (53, 177). Lastly, primed adaptation was also studied for both type V systems, by adding a protospacer or a mismatched protospacer on pTarget. As a negative control, a Cas2 knockout ($\Delta 2$) was used (Fig. 2A). Cells containing all three plasmids were grown for 48 hours in medium containing L-arabinose and IPTG to induce *cas* genes expression and antibiotics selecting for pCas_adaptation and pCas_effector, but not pTarget (Fig. 2B).

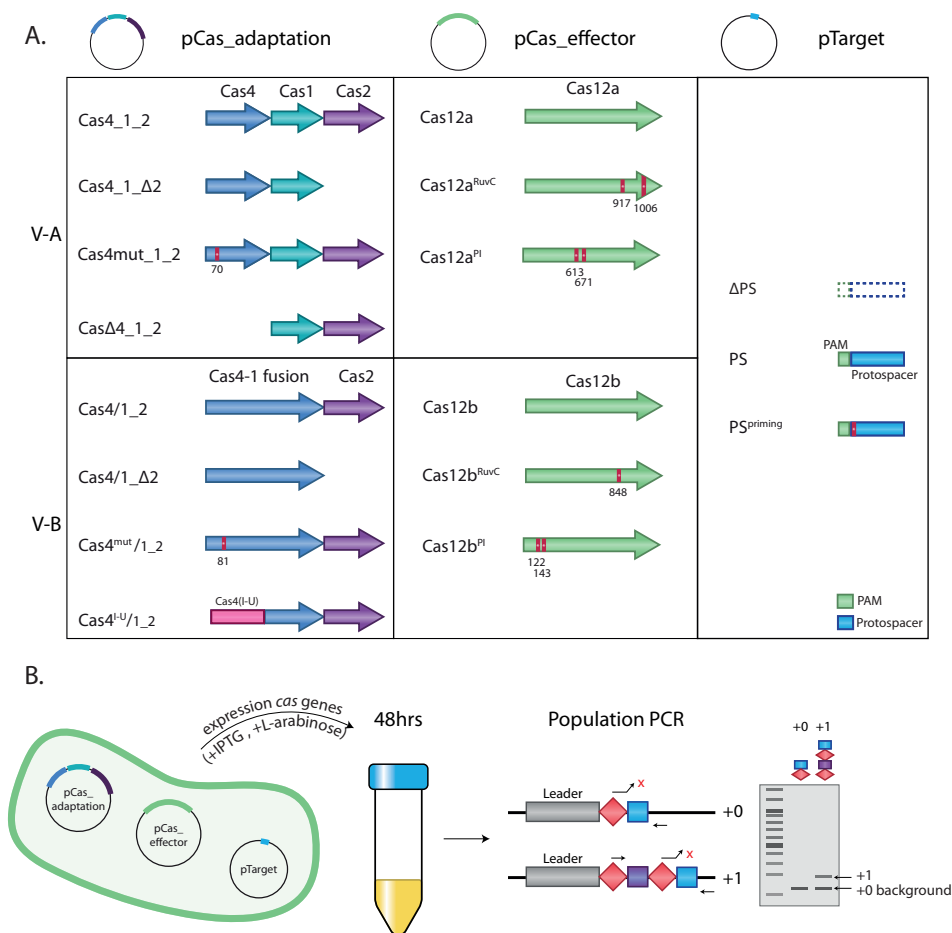


Figure 2 | Adaptation of type V-A and type V-B system in *E. coli*. (A) Variation of the three plasmids used in this study. Red square indicates position of the mutation. (B) Workflow schematic of the adaptation assay conducted using the three-plasmid system in *E. coli*. The three plasmids are, pCas_adaptation, pCas_effector and pTarget. L-arabinose and IPTG are added to induce expression of cas genes. Cells are grown for 48 hours in selective medium (except for pTarget) and subsequently used in a population PCR using degenerated primers. Amplified CRISPR arrays are visualized on an agarose gel electrophoresis.

cas genes required for adaptation

After 48 hours, cells were harvested and used for a population PCR to amplify expanded CRISPR arrays using degenerated primers and expanded arrays were visualized by agarose gel electrophoresis (Fig. 2B) (178). Adaptation was detected in all conditions except for the negative controls which were devoid of Cas2 (Fig. 3A

and S1, $\Delta 2$). These results indicated that only Cas1 and Cas2 are strictly required for adaptation for type V-A, since a mutation or knockout of either Cas4 or Cas12a did not affect the efficiency of adaptation. For type V-B, although Cas4 activity was dispensable, it was accompanied by a lower adaptation rate compared to the wildtype (Fig. 3B, 4/1_2 and 4^{mut}). In addition, swapping the Cas4 domain of Cas4/1 with that of I-U did not impact the adaptation rate, indicating that adaptation can occur with a swapped Cas4 domain or that the Cas4 domain is dispensable (Fig. 3B, 4^{IU}).

Adaptation rates using either the priming protospacer (PPS) with a single mismatch or without mismatches (PS) were not enhanced, indicating that priming does not occur in both Type V-A and V-B (Fig. 3, -, PS and PPS). However, spacer mapping should be also analyzed for primed conditions, since one of the characteristics of primed adaptation, is the acquisition of spacers near the targeted protospacer (166).

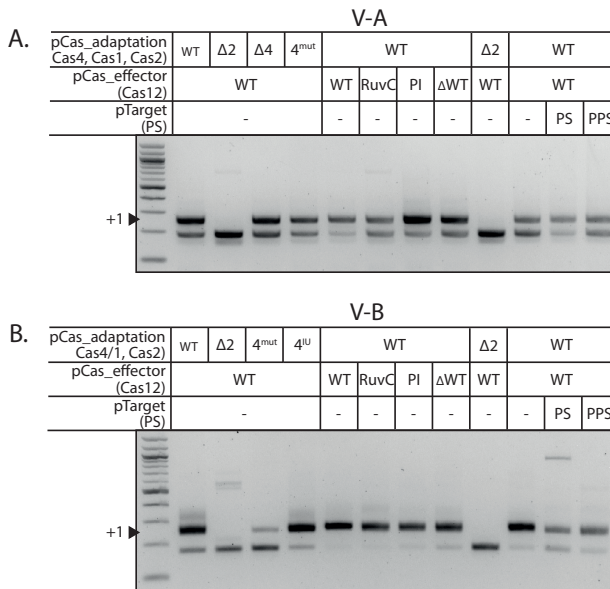


Figure 3 | Population PCR of cells expressing type V-A or V-B Cas genes (and variation hereof). CRISPR arrays were amplified after 48 hours of cas genes expression and visualized by agarose gel electrophoresis. Plasmids variants are indicated on top of the gel. WT pCas_adaptation= Cas4, Cas1 and Cas2 (V-A) or Cas4/1 and Cas2 (V-B), $\Delta 2$ = Δ Cas2, $\Delta 4$ = Δ Cas4, 4^{mut} = catalytically inactive Cas4, Δ WT pCas_effector = Δ Cas12a/b, RuvC = catalytically inactive Cas12a/b, PICas12a/b containing mutation in the PI domain, - = Δ protospacer, PS= with protospacer, PPS = priming protospacer containing a mismatch in the 1st nt of the protospacer. +1 spacer expanded is indicated by a black arrow.

PAM preference of newly acquired spacers

Next, we asked whether spacer adaptation in type V is biased for spacers containing a 5' PAM. It is known that for proper interference, FnCas12a and aaCas12b recognizes a T-rich 5'-PAM. More specifically 5'-NTTV and 5'-NTTN, respectively (43, 58). As such, amplicons of the expanded CRISPR arrays (Fig. 2) were sent for high-throughput sequencing (Table S1 and S2). Acquired spacers were extracted and mapped to the chromosome and the plasmids used. PAMs frequencies were scored by analyzing the DNA sequence upstream of the protospacers (5' PAM). PAM analysis was first done for the most wild type conditions, so conditions with and without Cas12a/b. These conditions would be the most likely to acquire functional spacers targeting a canonical T-rich PAM. The top 20 most frequently-occurring 5'-PAMs are listed in Fig. 3. However, no enrichment for a particular 5'-PAM could be observed in either condition (Fig. 4A and B). Also, nucleotide occurrence in position -4, -3, -2 and -1 of the 5'-PAM were also analyzed individually for all PAMs and found that all nucleotides showed an equal distribution, indicating that under these experimental conditions spacer integration was not PAM-dependent and not selected based on a T-rich 5'-PAMs (Fig. 4. C and D).

Influence of Cas4 on spacer length

In class 1 CRISPR-Cas systems, Cas4 trims and thereby determines the spacer length prior to integrating the spacer into the CRISPR array (26, 28, 175). We therefore analyzed spacer lengths conditions containing a wild type Cas4 (Fig. 5A (V-A), 5D (V-B)), a catalytically inactive Cas4 (Fig. 5B (V-A), 5E (V-B)) and a Cas4 knockout (Fig. 5C (V-A)). For type V-A, spacer length ranged from 24-36 nt, with the most common spacer length being 29 nt (Fig. 5A, B and C). Type V-B spacer lengths were substantially longer ranging from 33-41 nt with 35 nt as the most common spacer length (Fig. 5D and E). The spacer length distribution is similar to those found in the native CRISPR arrays (43, 58). However, the spacer length distribution was not affected by Cas4, as both the knocked out or catalytically-inert Cas4 resulted in similar distributions (Fig. 5).

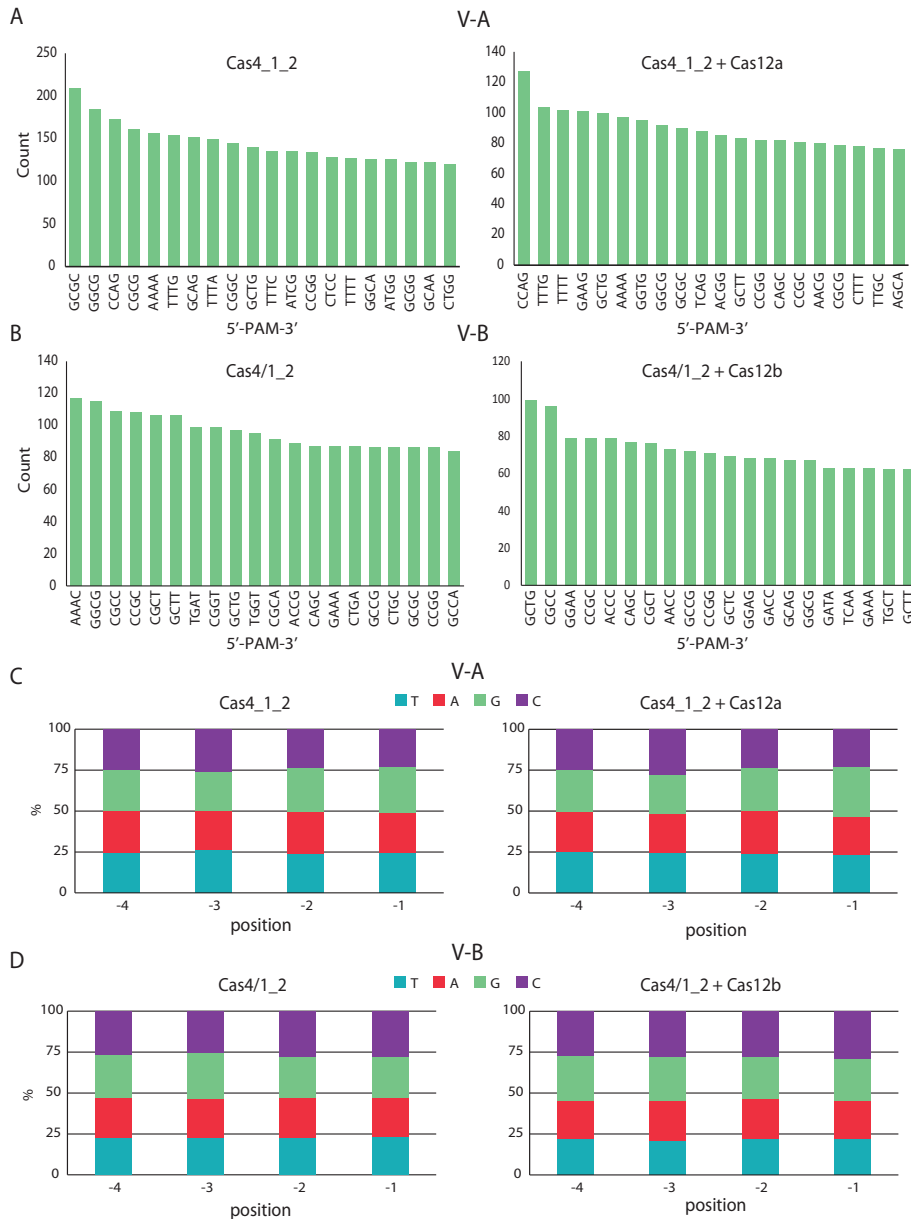


Figure 4 | 5'-PAM distribution of newly acquired spacer for type V-A and V-B. (A) Top 20 5'-PAM (NNNN) sequences for type V-A in cells expressing Cas4, Cas1 and Cas2 (Cas4_1_2) or cells expressing Cas4, Cas1, Cas2 and Cas12a (Cas4_1_2 + Cas12a). **(B)** Top 20 5'-PAM sequences for type V-B in cells expressing Cas4/1 and Cas2 (Cas4/1_2) and cells expressing Cas4/1, Cas2 and Cas12a (Cas4/1_2 + Cas12b). **(C)** 5'-PAM nucleotide distribution in the -4, -3, -2, and -1 position the protospacer for type V-A. **(D)** 5'-PAM nucleotide distribution in the -4, -3, -2, and -1 position the protospacer for type V-B.

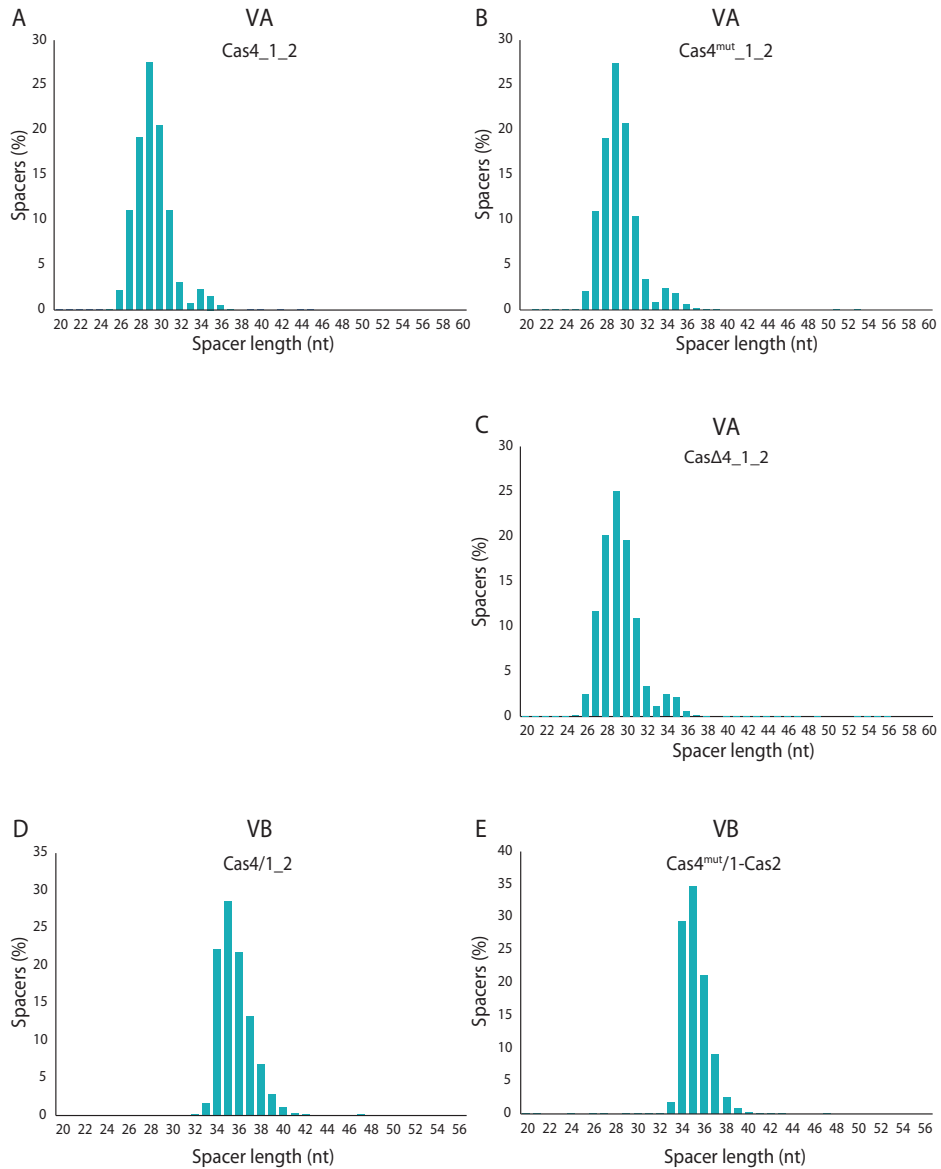


Figure 5 | Spacer length distribution of newly acquired spacers for type V-A and type V-B. Y-axis indicates spacers percentage (%) and x-axis indicates the spacer length (nt). **(A)** Spacer length distribution of V-A expressing Cas4, Cas1 and Cas2 (Cas4_1_2). **(B)** Spacer length distribution of V-A expressing catalytically Cas4, Cas1 and Cas2 (Cas4^{mut}_1_2). **(C)** Spacer length distribution of V-A expressing Cas1 and Cas2 (CasΔ4_1_2). **(D)** Spacer length distribution of V-B expressing Cas4/1 and Cas2 (Cas4/1_2). **(E)** Spacer length distribution of V-B expressing catalytically inactive Cas4 domain of Cas4/1 and Cas2 (Cas4^{mut}/1_2).

Spacer Mapping

Spacers obtained by primed adaptation are often acquired near the location of the previously acquired protospacer of the MGE. To assess whether type V-A or V-B also have primed adaptation, spacer mapping was done for the most wild type conditions, so conditions with and without Cas12a/b, but also for conditions with a protospacer (PS) or a primed protospacer (PPS). Spacers obtained from these conditions were mapped back to pCas_adaptation, pCas_effector, pTarget and the genome of BL21-AI (Fig. 6). In V-A systems, more spacers were acquired from the genome compared to V-B (Fig. 6A and B). Whereas in V-B, more spacers were acquired from pTarget compared to V-A (Fig. 6A and B). However, in both systems, spacers obtained from pTarget increased when Cas12 is expressed. This can be due to Cas12 selecting for spacers from pTarget as an outcome of target cleavage, since pTarget was not selected for in the growth medium. This effect is more pronounced in type V-B, which might indicate differences in cleavage efficiency. The addition of a protospacer (PS) also increased spacers being acquired from pTarget (Fig. 6A and B). This was also observed when a primed protospacer (PPS) is present in type V-A, but not in type V-B (Fig. 6A and B). When mapping the spacers onto the pTarget, no differences in spacer mapping were observed between the different conditions (Fig. 6C). A peak was to be expected ~3700 nt for pTarget, which corresponds with the protospacer location on pTarget (Fig. 6C). Also, no difference in spacer mapping was observed for pCas_adaptation, pCas_effector and the genome (Fig. 6C and S4). The peak locations found in pCas_effector differ in conditions expressing Cas4, Cas1 and Cas2 (Cas4_1_2; V-A) or Cas4/1 and Cas2 (Cas4/1_2; V-B) because in these conditions an empty pCas_effector plasmid was used as control, which is ~3.5 kb smaller (Fig. 6C). Correcting for this plasmid size difference leads to a similar a mapping pattern in all conditions. The first peak found in pCas_effector corresponds with the backbone of pCas_effector and the second peak corresponds with the *laci* gene. From these results, no primed adaptation was observed in either system.

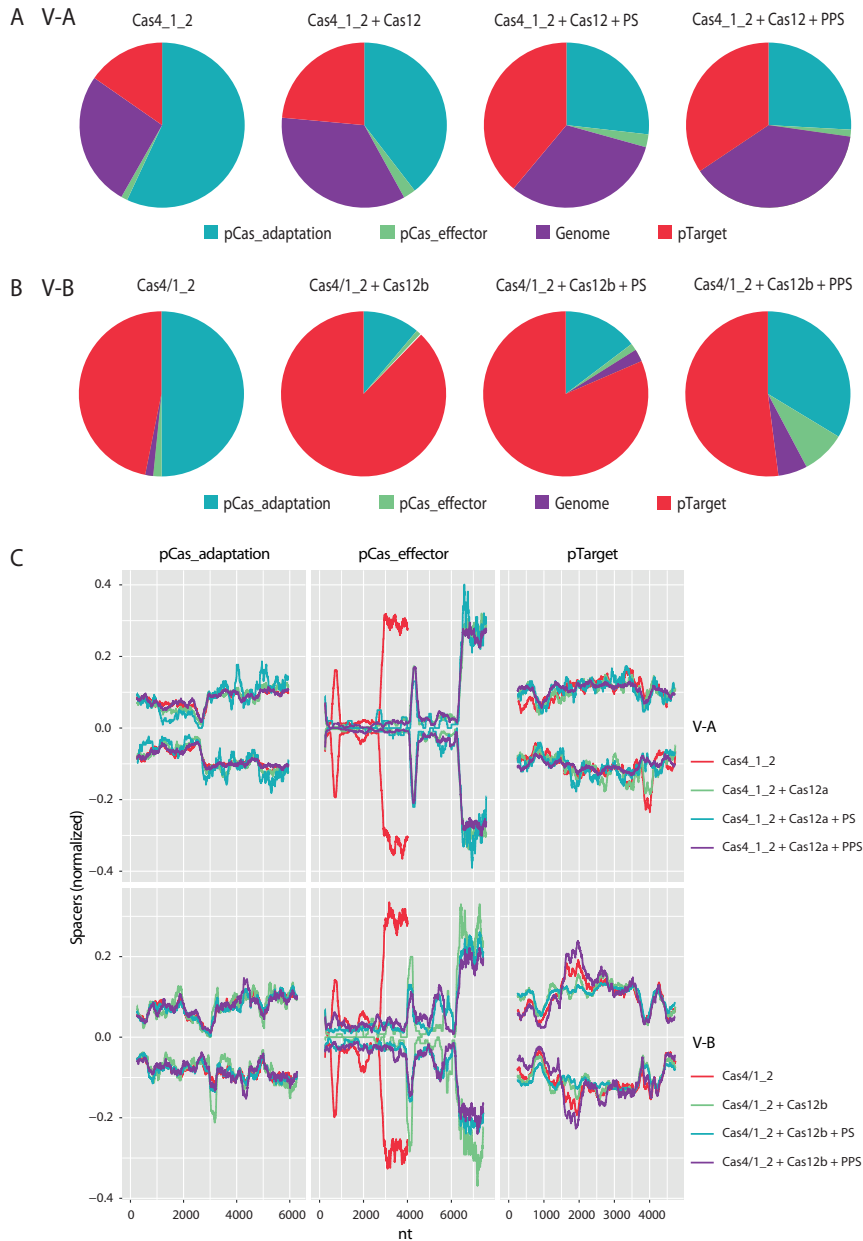
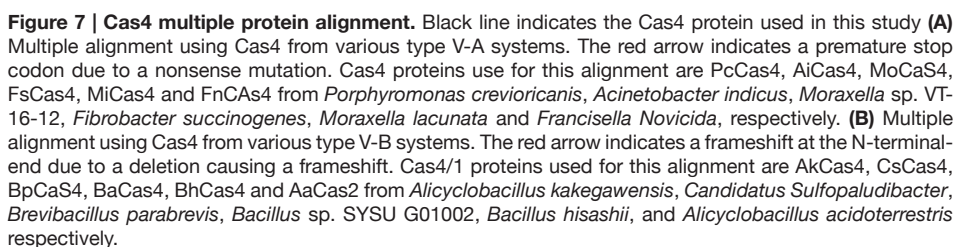


Figure 6 | Spacer mapping on pCas_adaptation, pCas_effector and pTarget. Data was obtained from deep sequencing using truncated Cas4, using all three replicates. Cas4_2_1 = Cas4, Cas1, Cas2, Cas4/1_2 = Cas4/1, PS = protospacer, PPS = primed protospacer containing a mismatch in the first nt of the protospacer. **(A)** Pie chart of the spacer distribution for type V-A. **(B)** Pie chart of the spacer distribution for type V-B. **(C)** Spacer mapping for type V-A and V-B on plasmids pCas_adaptation, pCas_effector and pTarget. X-axis indicates spacers normalized and y-axis indicate nucleotide (nt) position of the plasmid being mapped on.

Cas4 truncation

The absence of the expected consensus T-rich 5'-PAM and the lack of spacer length variation for the different Cas4 conditions for either type V systems, prompted us to speculate that no pre-spacer selection was taking place. Cas12 is only able to cleave PAM-containing protospacers. This means that the selection of functional spacers can occur at the interference stage and only cells containing spacers with the correct PAM would survive phage predation. However, this way of selecting for functional spacers is not a very efficient way of building up resistance within a population. To further investigate this phenomenon, a multiple alignment was made using the native FnCas4 (V-A) and aaCas4/1 (VB) sequence with other Cas4 and Cas4/1 variants from type V-A and type V-B systems, respectively (Fig. 7). The multiple alignment of Cas4 from type V-A revealed the presence of a nonsense mutation, which caused an early pre-mature stop codon encoded on amino acid position six of the protein TTG (Leu) → TAG (STOP) (Fig. 7A). We hypothesized that this would result in an incorrect translational start site of Cas4, expressing a shorter N-terminally-truncated Cas4 protein starting from isoleucine at amino acid position seven. The multiple alignment of Cas4/1 from type V-B revealed the presence of frameshift mutation, which was caused by a deletion GATG(Met) → GAT (Asp) (Fig. 7B). This led to an incorrect prediction of the start codon, which led to expressing a shorter N-terminally-truncated Cas4/1 starting from leucine (Fig. 7B). These results indicated that our spacer acquisition assays were performed by expressing N-terminally truncated Cas4 proteins for both type V-A and V-B, which could explain the absence of T-rich 5'-PAMs (Fig. 4) and the apparent lack in spacer length distribution (Fig. 5). Indeed, the N-terminal part of Cas4 has been reported previously to be important for binding to the Cas1-Cas2 complex (179).



To address whether the N-terminal Cas4 truncations were responsible for the lack of T-rich PAM containing spacers, we removed the pre-mature stop codon (type V-A) and restored the reading frame (type V-B) of the Cas4 ORFs (Fig. S2). Subsequently, a smaller scale spacer acquisition study was conducted by expressing Cas4, Cas1 and Cas2 in the presence or absence of Cas12a (type V-A) or Cas4/1, Cas2 in the presence or absence of Cas12b (type V-B). CRISPR arrays were amplified, gel extracted, cloned into a cloning vector (pJET 1.2) and sequenced Sanger sequencing. From the obtained spacers, the 5'-PAM was analyzed (Fig. 8). For type V-A, newly acquired spacers contained a canonical 5'-NTTV PAM in a Cas12a-independent manner (Fig. 8A, C and E) (43). For type V-B, a canonical 5'-NTTN PAM was also observed, but were more abundant in the absence of Cas12b (n= 15) (Fig. 8B, D and E) (58). This phenomenon might be explained by acquisition events from sources other than pTarget, resulting in cellular lethality due to the dsDNA cleavage by Cas12b of PAM-containing protospacers. This phenomenon was also expected for type V-A when expressing Cas12a. However, the reduction of a 5'-NTTV PAM in V-A in the presence of Cas12 is much lower than that of V-B (Fig. 8E). This can be due to the differences in cleavage efficiency between Cas12a and Cas12b. Apart from PAM analysis, a spacer length was also compared in conditions with and without Cas12 for both systems and found no difference, meaning Cas12 might not have a role in spacer trimming (Fig. S3). The role of Cas4 on spacer length distribution remains to be explored.

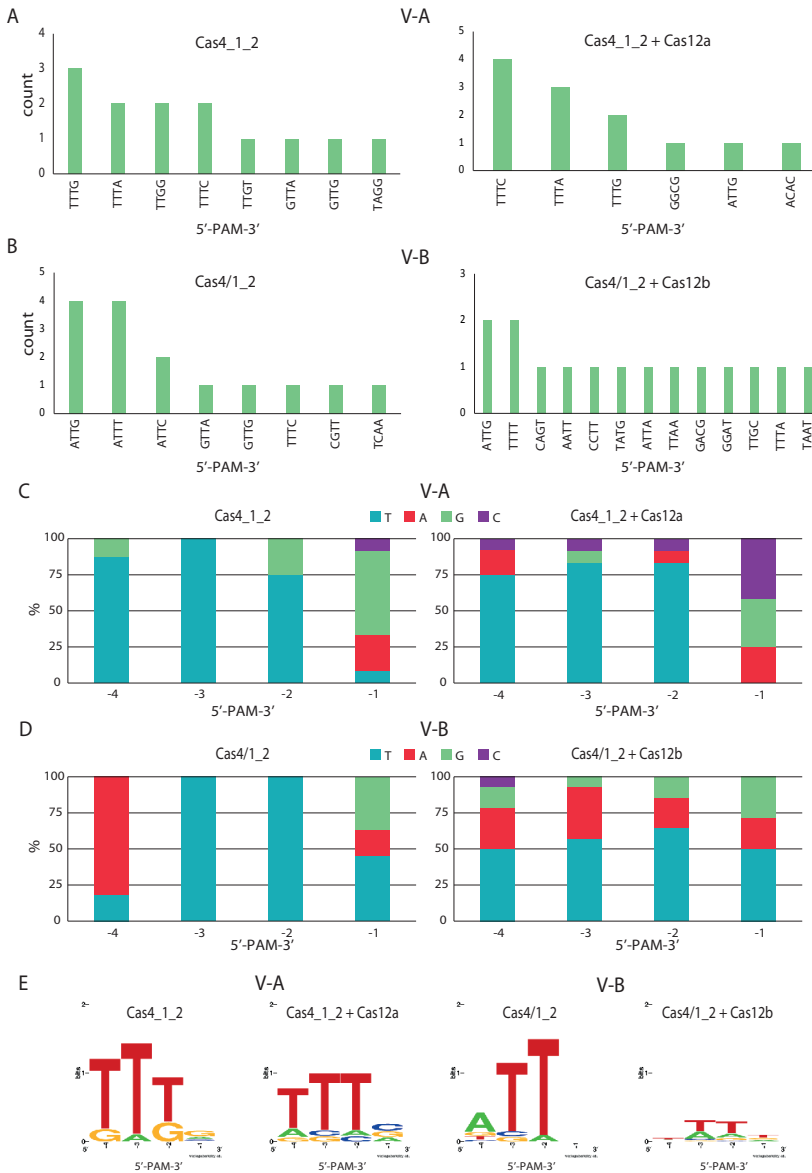


Figure 8 | 5'-PAM distribution of newly acquired spacer for type V-A and V-B with N-terminally-elongated Cas4. (A) 5'-PAM (NNNN) sequences for type V-A in cells expressing N-terminally-elongated Cas4, Cas1 and Cas2 (Cas4_1_2) (n=13) and cells expressing N-terminally-elongated Cas4, Cas1, Cas2 and Cas12a (Cas4_1_2 + Cas12a) (n=12). (B) 5'-PAM (NNNN) sequences for type V-B in cells expressing N-terminally-elongated Cas4/1 and Cas2 (Cas4/1_2) (n=15) and cells expressing N-terminally-elongated Cas4/1, Cas2 and Cas12a (Cas4/1_2 + Cas12b) (n=15). (C) 5'-PAM nucleotide distribution in the -4, -3, -2, and -1 position the protospacer for type V-A, using spacers from panel A. (D) 5'-PAM nucleotide distribution in the -4, -3, -2, and -1 position the protospacer for type V-B, using spacer from panel B (E) Web logo of 5'-PAMs for type V-A and type V-B, using spacer from panel A and B.

Spacer origin with N-terminally-elongated Cas4 or Cas4/1

Spacers obtained from the small-scale pilot experiment were mapped on the four possible sources, pCas_adaptation, pCas_effector, pTarget and the *E. coli* BL21-AI genome. Most of the spacers were acquired from pCas_adaptation, followed by pCas_effector, genome and lastly, pTarget (Fig. 9). The higher abundance of spacers acquired from pCas_adaptation, can be due to the result of a higher copy number, as spacers are often acquired from highly replicating replicons (180). Again, in type V-B, spacers are acquired more often from pTarget than V-A and this bias greater in the presence of Cas12b (Fig. 9 and 6B). Why this trend is not shown for V-A is unclear but can also be due to smaller sample size of spacer sequenced. Spacer mapping was not analyzed with the spacer obtained from the small-scale experiment, because no primed conditions were used in this experiment. Nonetheless, these results show that expression of an N-terminally truncated Cas4 proteins were indeed the cause for a lack of T-rich PAM containing spacers. This new setup will be used in future experiments, to obtain a large amount of correctly acquired spacers, which can be used for proper spacer mapping.

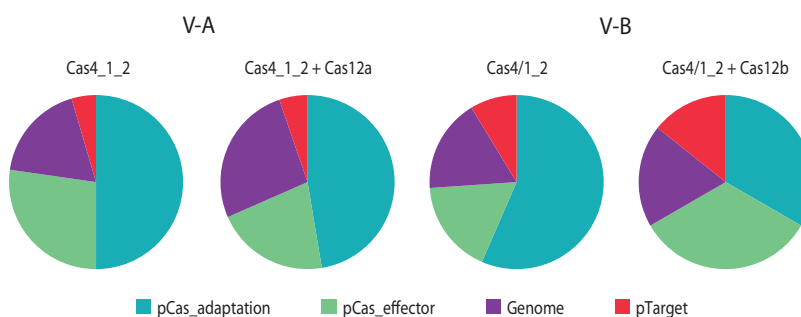


Figure 9 | Spacer origin from V-A and type V-B with N-terminally-elongated Cas4 or Cas4/1, respectively. Cas4_2_1 (n=22) = Cas4, Cas1 Cas2, Cas4_1_2 + Cas12a (n=19) = Cas4, Cas1 and Cas2 + Cas12a ; Cas4/1_2 (n=23) = Cas4/1 and Cas2 ; Cas4/1_2 + Cas12b (n=22) = Cas4/1, Cas2 + Cas12b. Data was obtained from Sanger sequencing using N-terminally-elongated Cas4 or Cas4/1.

Discussion and Conclusion

To obtain a thorough understanding of the adaptation mechanism in type V systems, an extensive study was set up to study both naïve and primed adaptation in type V-A and type V_B from *Francisella tularensis* subsp. *novicida* U112 and *Alicyclobacillus acidoterrestris* ATCC 49025, respectively. For type V-A, we showed that Cas1 and

Cas2 are sufficient for adaptation to occur, whereas Cas4 and Cas12a are not required. Similar results were found in type V-B, where the activity of the Cas4 domain and the Cas12b protein are not essential for adaptation. This resembles the situation in class 1 systems, but differs from other class 2 systems, such as type II-A, where Csn2 and Cas9 are strictly required for adaptation (29, 30). Contrary to II-A, in type V-C, a minimal system containing just Cas1 was shown to be sufficient for adaptation and acquisition of functional PAM containing spacers (174). This highlights that the adaptation machinery of type V-A and V-B systems are distinct to those of type V-C, because Cas1 and Cas2 alone are unable to acquire canonical 5'-PAM containing spacers. Although Cas4 is not required for adaptation in type V-A or type V-B, it supports adaptation by PAM-scanning and trimming pre-spacers. PAM-scanning by Cas4 has been previously reported in type I systems (26-28). The role of Cas4 in adaptation was discovered after correcting for a N-terminally-truncated *cas4* and *cas4/1* in type V-A and V-B, respectively. These mutations occurred in the genome of the host organism *Francisella Novicida* U112 and *Alicyclobacillus acidoterrestris* ATCC 49025. Having a mutation in the adaptation gene can be harmful, but not detrimental to a functional CRISPR-Cas system. This is because natural selection, such as that of a phage infection will select for cells that have randomly acquired a T-rich PAM containing spacer. To test whether these CRISPR-Cas systems are active in their native hosts, both organisms should be challenged by phages and acquired spacers should be analyzed. The Sanger approach resulted in the PAM-conclusion, but to fully address the mapping and trimming question a full-scale NGS experiment should be conducted. It is interesting to know, whether Cas4 from type II-B also contains this PAM-scanning activity since Cas9 scans for the PAM in type II-A systems (29, 30). As for the role of Cas12a in adaptation, preliminary data suggest that Cas12a does not seem to play a role in adaptation in either PAM selection or spacer length. Cas12a might play a role in the adaptation rate, when a protospacer or mismatch protospacer is already present, since there is strong evidence to suggest that class 2 systems also have primed adaptation (166). It is possible that a N-terminally truncated Cas4 or Cas4/1 prevented primed adaptation from occurring. Primed adaptation conditions will be tested in the follow-up study expressing the N-terminally elongated Cas4 or Cas4/1. To summarize, a large-scale experiment was set-up to study different aspects of adaptation in type V-A and type V-B. The NGS approach showed that expressing the Cas4 as they are natively encoded in their host organisms is not sufficient to obtain T-rich PAM containing spacers. This raises the question whether the adaptation activities are still of importance for immunity in their respective hosts. Correcting the Cas4 protein resulted in spacers targeting a canonical 5'T-rich PAM. Our results showed that Cas4 is involved in PAM-scanning in both type V-A and V-B, in contrast to other class 2 systems. Having expressed the corrected Cas4 protein is only just the start which enables us to dive deeper in the adaptation mechanism of type V-A and V-B.

Methods

Bacterial strains and growth conditions

For plasmid cloning *E. coli* strains DH5- α and DH10- β were used. As for the adaptation growth experiment, *E. coli* BL21-AI was used containing a T7 polymerase under an arabinose inducible promoter. Cells were grown in 37°C at 220 rpm in Luria Bertani (LB) medium consisting of 10 g/L peptone, 10 g/L NaCl and 5 g/L yeast extract. Ampicillin (Amp)(100 μ g/mL), spectinomycin (Spec)(100 μ g/mL), chloramphenicol (Cam)(35 μ g/mL), L-arabinose (2 g/L) and IPTG (0.5 mM) were added when required.

Plasmid construction

The adaptation experiment in *E. coli* consists of three-plasmids: pCas_adaptation, pCas_effector and pTarget. These plasmids have a spectinomycin, chloramphenicol and ampicillin resistance gene and a CloDF13(~20-40 copies), p15A (~10-12 copies) and ColE1 (~15-20 copies) ori, respectively. Details on each primer and plasmid use in this study can be found in supplementary Table S3 and Table S4. The initial pCas_adaptation plasmids (pCas4_1_2_VA_pre and pCas4/1_2_VB) were cloned by ligation independent cloning (LIC) using plasmid 13S-S (addgene #48329). pCas4_1_2_VA_pre was missing a cas4 domain, which was later added via PCR to create pCas4_1_2_VA. pCas4_1_2_VA was then used to create pCas4_1_2_VA, pCas4 Δ 1_2_VA and pCas-mut4_1_2_VA by three-point ligation using HindIII, BsmI and KpnI restriction sites. pCas-4/1 Δ 2_VB and pCas-mut4/1_2_VB were also created by three-point ligation but used AflIII - BsmI - HindIII, and AflIII - blunt - XmaI restriction sites, respectively. pCas_4(I-U)/1_2 was constructed by Gibson assembly using a linear fragment of pCas4/1_2_VB and a cas4 domain amplified out of the I-U system from pCas4/1-2LR (Almendros et al., 2019). To restore Cas4 truncated genes, pCas4_1_2_VA_elongated and pCas4_1_2_VB_elongated were constructed by around the horn PCR.

As for the pCas_effector plasmids, pCas12a was achieved by amplification of a pACYC-duet-Cas12a-Cas4_1_2 to remove cas4, cas1 and cas2. Afterwards pCas12a(RuvC) and pCas12a(PI) were constructed by restriction digestion using SapI and SpeI restriction enzymes and ligating an insert, digested with the same restriction enzyme, from pRham_Cas12a(RuvC) and pRham_Cas12a(PI), respectively. pCas12b was also constructed by restriction digestion and ligation to pACYC-duet using EcoRI and BamHI. pCas12b(RuvC) was then constructed by amplification of two fragments from pCas12b and ligating them together by blunt end ligation. Mutations were introduced in the 5'end of primers. pCas12b(PI) was constructed by GoldenGate and ligating the vector to a short insert create by two oligonucleotides annealed to each other.

pTarget_no_ps is a p2A-T plasmid from (addgene # 29665) the LIC collection. Protospacers for targeting and priming were introduced by amplification of the whole plasmid using primers containing the protospacer in the overhang, also known as around the horn PCR. pTarget_no_ps2 was later constructed to remove the T7 promoter by gibbon assembly using NEBuilder® HiFi DNA Assembly Master Mix.

Adaptation growth experiment

pTarget was transformed into BL21-AI strains containing pCas_adaptation and pCas_effector, then plated on agar plates containing Amp, Spec and Cm and incubated overnight. The following day, three colonies from each plate were inoculated into 2ml medium in a 15 ml falcon tube. Cells were grown for 3 hours in 37°C (shaking) and then cas genes expression was induced by the addition of L-arbinose and IPTG and grown for an additional 48 hours. Final OD₆₀₀ was measured and corrected to a OD₆₀₀ of 1. 200 ul of cells were harvested and transferred to 1.5 ml Eppendorf tube, centrifuged for 1 min at max speed, resuspended in 50ul MQ and stored at 4°C.

Population PCR

2 µl of cells was used in a 50 µl reaction using Q5® High-Fidelity 2X Master Mix from new England biolabs. Degenerate 3' phosphorothioated primers ordered from IDT were used in PCR (Table S5). PCR reactions and thermocycling conditions were carried out according to manufacturer's protocol. Initial denaturation was 10 min, extension time 30 sec and annealing temperature was 67 °C for V-A and 70 °C for V-B. Amplified products were separated by gel electrophoresis using a 3% agarose gel and visualized using a Bio-Rad imager.

Sample preparation for sequencing

To prepare the samples for deep sequencing, samples were pooled, and PCR purified, and library preparation was done using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® according to manufacturer's protocol.

For Sanger sequencing, amplified +1 arrays were gel extracted and cloned into pJET2.1 using CloneJET PCR Cloning Kit (Thermo Scientific™) according to manufacturer's protocol. Individual colonies were then analysed by Sanger sequencing (Eurofins).

Multiple alignment of Cas4 and Cas4/1

Amino acid sequences were obtained from UniProt and multiple sequence alignment was done using ClustalO.

Spacer mapping

Most spacers mapped on multiple locations, due to partial sequence similarities between the plasmids and chromosome (e.g. the *lacI* gene, which is present in pCas_adaptation, pCas_effector and the *E. coli* genome). These spacers are undistinguishable as to which source it was obtained from. For both the NGS experiment and the smaller-scale experiment, spacers mapping back to multiple plasmids were all counted as one hit.

Acknowledgments

We would like to thank Carina Nieuwenweg, Meral Türen, Alexander Bartels, Laure de Nies, Pilar Bobadilla Ugarte for their technical assistance. J.v.d.O is supported by the NWO/TOP grant 714.015.001.

Author contributions

W.Y.W., C.A., S.A.J., J.v.d.O., S.J.J.B. and R.H.J.S. conceived this study and the experimental design. W.Y.W., S.A.J., C.A., S.Y., R.J., C.N., conducted the experimental work. W.Y.W. and R.H.J.S. supervised this project. W.Y.W. and R.H.J.S. wrote the manuscript.

Corresponding author

Correspondence should be addressed to raymond.staals@wur.nl

Competing interest

No potential conflict of interest is reported by the authors

Supplementary Figures and Tables

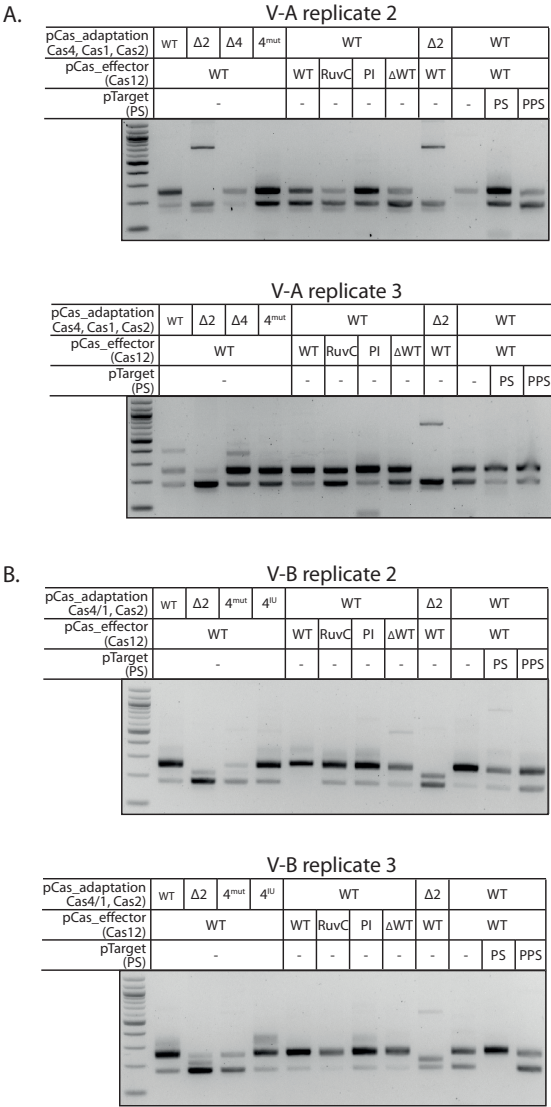


Figure S1 | Population PCR of cells expressing type V-A or V-B Cas genes (and variation hereof) in biological replicates two and three. CRISPR arrays were amplified after 48 hours of cas genes expression and visualized by agarose gel electrophoresis. Plasmids variants are indicated on top of the gel. WT pCas_adaptation= Cas4, Cas1 and Cas2 (V-A) or Cas4/1 and Cas2 (V-B), Δ2 = ΔCas2, Δ4 = ΔCas4, 4^{mut} = catalytically inactive Cas4, ΔWT pCas_effector = ΔCas12a/b, RuvC = catalytically inactive Cas12a/b, PICas12a/b containing mutation in the PI domain, - = Δprotospacer, PS= with protospacer, PPS = priming protospacer containing a mismatch in the 1st nt of the protospacer. +1 spacer expanded is indicated by a black arrow. **(A)** Replicates of V-A. **(B)** Replicates of V-B.

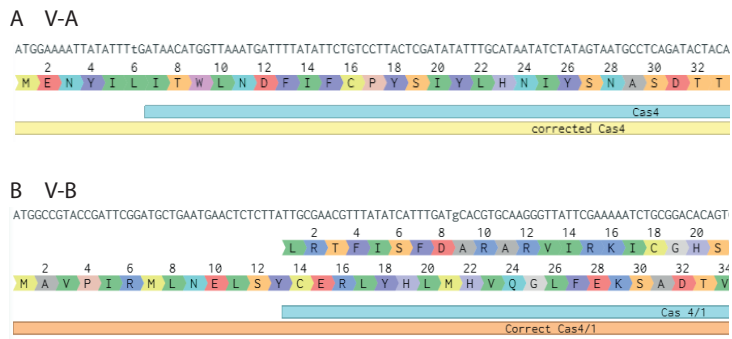


Figure S2 | Corrected N-terminal Cas4 amino acid sequence. (A) DNA and amino acid sequenced of corrected Cas4 for type V-A. Lower case “t” indicates A → T mutation in Leu6. Blue box indicates previously expressed cas4. Yellow box indicates corrected cas4. **(B)** DNA and amino acid sequenced of corrected Cas4 for type V-A. Lower case “g” indicates G insertion causing a frameshift. Blue box indicates previously expressed cas4/1. Orange indicate corrected cas4/1.

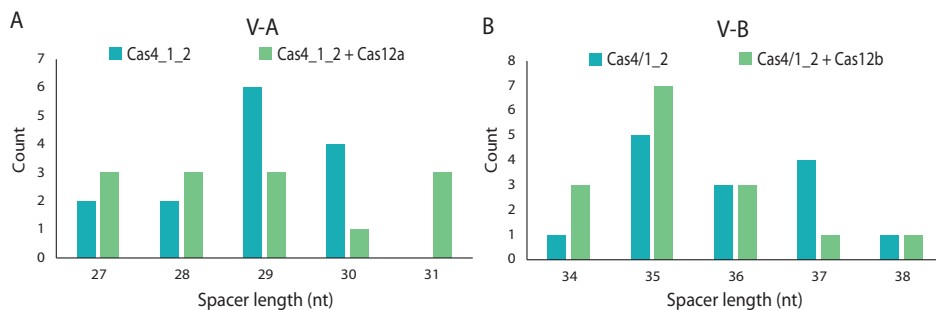


Figure S3 | Spacer length distribution of spacers after restoring Cas4 for type V-A and type V-B. Data was obtained from Sanger sequencing. **(A)** Spacer length distribution for V-A. Cas4_2_1 (n=14) = Cas4, Cas1 Cas2, Cas4_1_2 + Cas12a (n=13) = Cas4, Cas1 and Cas2 + Cas12a ; Cas4/1_2 (n=14) = Cas4/1 and Cas2 ; **(B).** Spacer length distribution V-B. Cas4/1_2 + Cas12b (n=15) = Cas4/1, Cas2 + Cas12b.

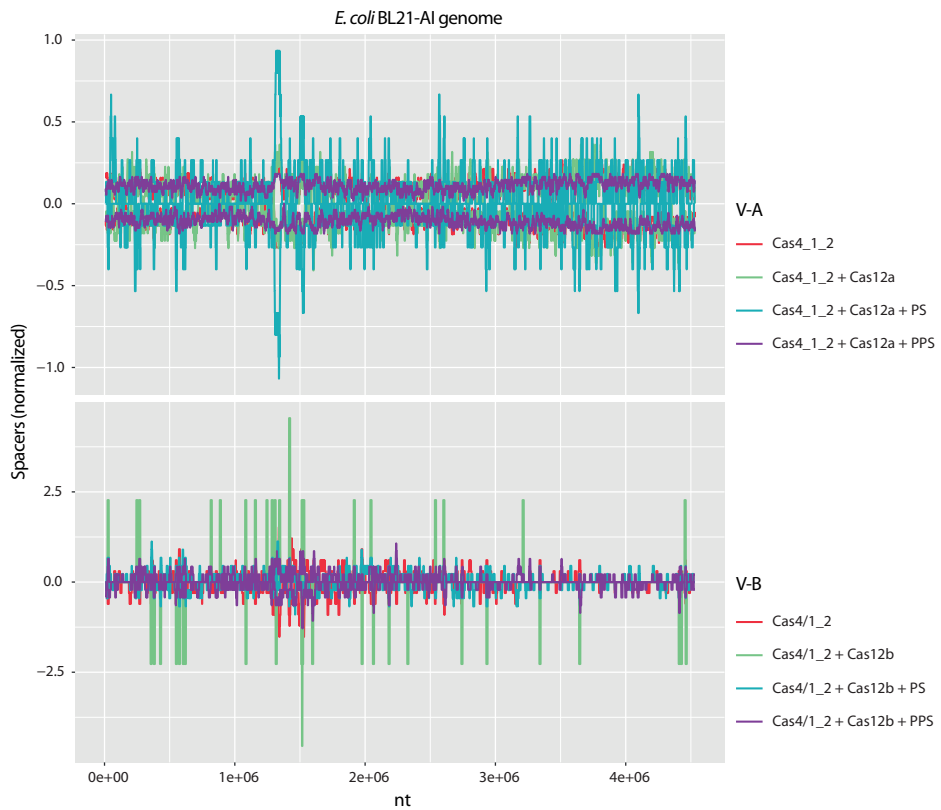


Figure S4 | Spacer mapping on BL21-AI genome. X-axis indicates spacers normalized and y-axis indicate nucleotide (nt) position of the plasmid being mapped on PS = protospacer PPS = primed protospacer.

Table S1 | NGS sequencing summary type V-B

Names	Sample #	replicate	Total reads	Total spacers	Total unique spacers
Cas412	01	rep1	4164	3246	2990
Cas412	01	rep2	15778	10704	9605
Cas412	01	rep3	6887	5991	5314
Cas41Δ2	02	rep1	169	128	122
Cas41Δ2	02	rep2	2337	952	905
Cas41Δ2	02	rep3	55	48	48
CasΔ412	03	rep1	697	666	612
CasΔ412	03	rep2	66030	30524	27571
CasΔ412	03	rep3	3545	1834	1674
Casmut412	04	rep1	2644	2405	2267
Casmut412	04	rep2	2170	2060	1859
Casmut412	04	rep3	15480	10866	9796
Cas412 + Cas12a	05	rep1	265	260	182
Cas412 + Cas12a	05	rep2	10790	9082	7014
Cas412 + Cas12a	05	rep3	3170	2456	1762
Cas412 + dCas12a	06	rep1	4951	2723	1923
Cas412 + dCas12a	06	rep2	1461	1346	992
Cas412 + dCas12a	06	rep3	1770	1620	1113
Cas412 + PlmutCas12a	07	rep1	1903	1490	1154
Cas412 + PlmutCas12a	07	rep2	2038	1969	1501
Cas412 + PlmutCas12a	07	rep3	888	862	593
Cas412 + Empty pCas	08	rep1	5033	3764	2913
Cas412 + Empty pCas	08	rep2	4823	4302	3098
Cas412 + Empty pCas	08	rep3	20583	14336	10701
Cas41Δ2 + Cas12a	09	rep1	22	21	16
Cas41Δ2 + Cas12a	09	rep2	618	419	312
Cas41Δ2 + Cas12a	09	rep3	19	18	16
Cas412 + Cas12a-naive	10	rep1	1602	1474	1107
Cas412 + Cas12a-naive	10	rep2	86487	35506	27065
Cas412 + Cas12a-naive	10	rep3	524	481	329
Cas412 + Cas12a-target	11	rep1	2355	2197	1623
Cas412 + Cas12a-target	11	rep2	987	954	768
Cas412 + Cas12a-target	11	rep3	209	207	179
Cas412 + Cas12a-priming	12	rep1	788	703	539
Cas412 + Cas12a-priming	12	rep2	24743	17934	14723
Cas412 + Cas12a-priming	12	rep3	31070	15928	12399

Table S2 | NGS sequencing summary type V-A

Names	Sample #	replicate	Total reads	Total spacers	Total unique spacers
Cas412	01	rep1	2700	2483	2369
Cas412	01	rep2	5515	4744	4274
Cas412	01	rep3	6699	5737	5242
Cas41Δ2	02	rep1	84	27	27
Cas41Δ2	02	rep2	2	2	2
Cas41Δ2	02	rep3	390	309	297
Casmut412	03	rep1	1482	915	880
Casmut412	03	rep2	537	431	412
Casmut412	03	rep3	552	498	482
Cas4(IU)12	04	rep1	1064	1005	958
Cas4(IU)12	04	rep2	18073	8099	7854
Cas4(IU)12	04	rep3	2587	2169	2076
Cas412 + Cas12b	05	rep1	9183	7522	6780
Cas412 + Cas12b	05	rep2	1097	1064	962
Cas412 + Cas12b	05	rep3	967	932	848
Cas412 + dCas12b	06	rep1	20633	14788	11345
Cas412 + dCas12b	06	rep2	1573	1527	1104
Cas412 + dCas12b	06	rep3	7567	6422	4888
Cas412 + PlmutCas12a	07	rep1	1511	1455	1123
Cas412 + PlmutCas12a	07	rep2	7362	6333	4921
Cas412 + PlmutCas12a	07	rep3	5697	4988	3855
Cas412 + Empty pCas	08	rep1	3662	3376	2605
Cas412 + Empty pCas	08	rep2	7106	5996	4853
Cas412 + Empty pCas	08	rep3	4394	3940	3003
Cas41Δ2 + Cas12	09	rep1	32	29	22
Cas41Δ2 + Cas12	09	rep2	24	21	17
Cas41Δ2 + Cas12	09	rep3	693	276	209
Cas412 + Cas12b-naive	10	rep1	9037	7501	6557
Cas412 + Cas12b-naive	10	rep2	24447	16446	14647
Cas412 + Cas12b-naive	10	rep3	2167	1946	1614
Cas412 + Cas12b-target	11	rep1	2265	1771	1450
Cas412 + Cas12b-target	11	rep2	10885	6511	4870
Cas412 + Cas12b-target	11	rep3	52990	23628	21422
Cas412 + Cas12b-priming	12	rep1	149	148	125
Cas412 + Cas12b-priming	12	rep2	7293	5611	4432
Cas412 + Cas12b-priming	12	rep3	56065	6919	5435

Table S3 | Plasmids constructed in this study and their cloning strategy. BB = backbone

Name	cloning strategy	template
pCas_adaptation		
pCas4_1_2_VA_pre	Ligation In dependent Cloning (LIC)	pY002 (addgene #69975)
pCas4_1_2_VA	Round-the-horn PCR	pCas4_1_2_VA_pre
pCas4_1_Δ2_VA	Three point ligation	pCas4_1_2_VA
pCasΔ4_1_2_VA	Three point ligation	pCas4_1_2_VA
pCas-mut4_1_2_VA	Three point ligation	pCas4_1_2_VA
pCas4/1_2_VB	Ligation In dependent Cloning (LIC)	pZ001 (addgene #70166)
pCas-4/1Δ2_VB	Three point ligation	pCas4/1_2_VB
pCas-mut4/1_2_VB	Three point ligation	pCas4/1_2_VB
pCas_4(l-U)/1_2	Gibson	pCas4/1-2LR (Almendros et al., 2019) pCas4/1_2_VB
pCas4_1_2_VA_Cas4_elongated	Round-the-horn PCR	pCas4_1_2_VA
pCas4/1_2_VB_Cas4_elongated	Round-the-horn PCR	pCas4/1_2_VB
pCas_effector		
pCas12a	Round-the-horn PCR	pACYC_Cas12a_Cas412
pCas12a(RuvC)	Digestion and ligation	pCas12a pRham_dCpf1
pCas12a(PI)	Digestion and ligation	pCas12a pRham_Cpf1PI
pCas12b	Digestion and ligation	pACYC-duet pZ001 (addgene #70166)
pCas12b (RuvC)	blunt-end ligation	pCas12b
pCas12b (PI)	Goldengate with inserted created by two oligos	pCas12b oligo inserts annealed
pTarget		
PS_VA	Round-the-horn PCR	pTarget/p2A-T (addgene # 29665)
Priming_VA	Round-the-horn PCR	pTarget/p2A-T (addgene # 29665)
PS_VB	Round-the-horn PCR	pTarget/p2A-T (addgene # 29665)
Priming_VB	Round-the-horn PCR	pTarget/p2A-T (addgene # 29665)
pTarget2	Gibson	pTarget/p2A-T (addgene # 29665) pUA66

Description	
	PCR insert using BN1034, BN1035
	PCR insert using BG14039, BG14040
	Digested BB with HindIII & KpnI
	PCR part 1 (BG14426 and BG14427) digested with HindIII & BsmI
	PCR part 2 (BG14428 and BG14429) and digested with BsmI & KpnI
	Digested BB with AflII & HindIII
	PCR part 1 (BG14434 and BG14435) and digested with AflII & BsmI
	PCR part 2 (BG14436 and BG14437) and digested with BsmI & HindIII
	Digested BB with AflII & HindIII
	PCR part 1 (BG14438 and BG14439) and digested with AflII & BsmI
	PCR part 2 (BG14440 and BG14441) and digested with BsmI & HindIII
	PCR insert using BN1043, BN1044
	Digested BB with XmaI & BamHI
	PCR part 1 (BG14430 and BG14431) and digested with XmaI & BsmI
	PCR part 2 (BG14432 and BG14433) and digested with BsmI & BamHI
	Digested BB with AflII & XmaI
	PCR part 1 (BG14438 and BG14439) and digested with AflII
	PCR part 2 (BG14440 and BG14441) and digested with XmaI
	PCR insert (BG14442 and BG14443)
	PCR backbone (BG14444 and BG14445)
	PCR backbone (BG22746 and BG22850)
	PCR backbone (BG22746 and BG22851)
	PCR backbone (BG15060 and BG15061)
	Digested with SapI & SpeI
	Digested with SapI & SpeI
	Digested with SapI & SpeI
	Digested with SapI & SpeI
	Digested with EcoRI & BamHI
	PCR (BN1050 & BN1051) and digested with EcoRI and BamHI
	PCR part 1 (BG15486 and BG15487)
	PCR part 2 (BG15488 and BG15489)
	PCR (BG15490 & BG15491) and digested with BsaI
	BG15492 & BG15493
	PCR (BG15483 & BG15485)
	PCR (BG15484 & BG15485)
	PCR (BN1211 & BN505)
	PCR (BN1320 & BN505)
	PCR
	PCR

Table S4 | Oligonucleotides used in this study

oligo ID	sequence (5'-3')
BG14039	ATTCTGTCTTACTCGATATATTTGC
BG14040	ATAAAATCATTTAACCATGTTATTGCATTGG
BG14426	TGGAGTGAACTTAGTCTATCATTG
BG14427	ATAGAATGCGTCATAACTGACTATCAACATACACC
BG14428	TATGAATGCGAAAATTGCAAACCTTAGTCTTTATGTT
BG14429	TCCTTTCGGGCTTTGTTAG
BG14430	AACCGGTACGTGTCGGAGC
BG14431	ATAGAATGCGACATATCGGCGCATCGCCA
BG14432	TATGAATGCGAGGATTGATGTAACTTCAAATACG
BG14433	TTCTTTCGGGCTTTGTTAG
BG14434	TACATATGAAATCTTCTCACC
BG14435	ATAGAATGCGGAATATAAAATCATTTAACCATGTTAT
BG14436	TATGAATGCGGTGATAAAACGGACTTGTTAG
BG14437	CTACGATAAACAATGCAAGA
BG14437	CTACGATAAACAATGCAAGA
BG14438	GATGGAAGCGTTCGCTAAAAGAC
BG14439	TGCCGCTTCAACGGGCTCC
BG14440	CATTCAAGCGCACCAACG
BG14441	GGTCGATTCCATCCCATCTTTC
BG14442	ATCCAATGCAATGGCTGAGACAGACGG
BG14443	CGGACTCGTGAGAAAGCGCACTTCATC
BG14444	TCTCAGCCATTGCATTGGATTGGAAGTACAGG
BG14445	GCTTCTCAGGAGTCCGGACGGAGAG
BG14858	TACATATGAAATCTTCTCACCATCACC
BG14859	ATAGAATGCGCTCTCTCCACAAGTAAGCCC
BG14860	TATGAATGCGAGGCAATCAAGACTATCTAT
BG15060	GCTGCTGCCACCGCTGAG
BG15061	GCAAGCTTGTCGACCTGCAGG
BG15483	AATTTAGAGAAGTCATTTAATAAGGCCACTGTTAAAAGCTGATCC GGCTGCTAAC
BG15484	AATTTACAGAAGTCATTTAATAAGGCCACTGTTAAAAGCTGATCC GGCTGCTAAC
BG15485	TTATGGAGTTGGGATATCTATATCTCC
BG15486	GCCTTGTTGCAAAATTGGCAGAG
BG15487	GCCAGCAGGATGAGCTGG
BG15488	GGAATTGAGCGAGTACCAGTTC
BG15489	ATGTGATCCGAATCGTCTCG

description

Fw V-A Cas4 SDM 2

Rv V-A Cas4 SDM 2

Fw VA ΔCas2 US

Rv BsmI VA ΔCas2 US

Fw BsmI VA ΔCas2 DS

Rv VA ΔCas2DS

Fw VB ΔCas2 US

Rv BsmI VB ΔCas2 US

Fw VB ΔCas2 del DS

Rv VB ΔCas2 del DS

Fw VA ΔCas4 US

Rv BsmI VA ΔCas4 US

Fw BsmI VA ΔCas4 DS

Rv VA ΔCas4 DS

Rv VA ΔCas4 DS

Fw VB K81ACas4 US

Rv VB K81ACas4 US

Fw VB K81ACas4 DS

Rv VB K81ACas4 DS

Fw Cas4 I-U Gibson

Rv Cas4 I-U Gibson

Fw vector 4/1_2 VB Gibson

Rv vector 4/1_2 VB Gibson

Fw VA K70ACas4 US

Rv BsmI VA K70ACas4 US

Fw BsmI VA K70ACas4 DS

Fw pACYC_Cas12a

Rv pACYC_Cas12a

Fw VA target ps

Fw VA priming ps

Rv VA ps

Fw dCas12b D570A

Rv dCas12b E848A

Fw dCas12b E848A BB

Rv dCas12b D570A BB

oligo ID	sequence (5'-3')
BG15490	GCCAAATTTTGTAGCCCTTGGCCGACAAGGACGCAGTTGGTGGGCTTGGAATCGCGAAGG
BG15491	GGCGCCTTCGCGATTCCAAGCCCACTGCGTCCTTGTGCGCCAAGGGGCTCAAAAATT
BG15492	ATAGGTCTCGCGCCGAACAAACCGCGGTGGGTTC
BG15493	TATGGTCTCTTGGCGCAATTGCTGCGCGTC
BG22746	CATATGTATATCTCCTTCTTAAAGTTAAAC
BG22850	GAAAATTATATTTTGATAACATGGTTAAATGATTTTATATTTCTGTC
BG22851	GCCGTACCGATTTCGGATGCTGAATGAACTCTCTTATTGCGAACGTTTATATCATTTGATGCACGTGCAAGGGT TATTCGAAAAATC
BG22880	GACGAAAGGGCCTCGTGATACG
BG22881	TCATGCAACTCGTAGGACAGGTG
BG22882	CTACGAGTTGCATGATCACTGATAGATACAAGAGCCATAAGAAC
BG22883	CGAGGCCCTTTCGTCAAAGCAAAATGAACTAGCGATTAGTCG
BN1034	TACTTCCAATCCAATGCATTGCATAATATCTATAGTAATGCCT
BN1035	GTCATTTAATAAGGCCACTGTTAAAATAACATTGGAAGTGGATAA
BN1043	TACTTCCAATCCAATGCATTGCGAACGTTTATATCATTTGATC
BN1044	GGTAAAAAGACGAATGATGCATCCTAACATTGGAAGTGGATAA
BN1050	AAGGATCCTTAGGAGGGCGCTAGATGCGCTCCATCCCCCATC
BN1051	TTGAATTCTTAAATATCCCCGTGTTTTCAC
BN1211	AATTTTCGTTTGGTAAAGGTAAAAAGACGAATGATGCATCCGCTGATCCGGCTGCTAAC
BN1320	AATTTTGTGTTGGTAAAGGTAAAAAGACGAATGATGCATCCGCTGATCCGGCTGCTAAC
BN505	TTATGGAGTTGGGATATCTATATCTCC
-	CGACTCACTATAGGAGAGCGGC
-	AAGAACATCGATTTTCCATGGCAG

Table S5 | Barcode primers use for PCR amplification of expanded array

Name	sequence (5'-3')
Fw VA_deg	NNNNNNGGTCTAAGAACCTTAAATAATTTCTACTGTTGTAGAT* H
Fw VB_deg	NNNNNNGCGATCTGAGAAGTGGCAC* V
Rv pCas_adaptation	NNNNNNAACTCAGCTTCCTTTCGGGCTTT*G

description
R122A G143P oligo Top
R122A G143P oligo Bot
Fw Bsal Cas12b Plmut BB
Rv Bsal Cas12b Plmut BB
Rv pCas_adaptation_phos
Fw pCas_adaptation_V-A_Cas4
Fw pCas_adaptation_V-B_Cas4
Fw p2A_T gibson
Rv p2A_T gibson
Fw pUA66 insert gibson
Rv pUA66 insert gibson
Fw Cas4_1_2 VA LIC
Rv Cas4_1_2 VA LIC
Fw Cas4/1_2 VB LIC
Rv Cas4/1_2 VB LIC
Fw BamHI Cas12b
Rv EcoRI Cas12b
Fw VB_PS SEED T1C
Fw VB PS
Rv pTarget
Fw pJet 2.1
Rv pJet 2.1

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
G	T	C	T	A	A	G	A
T	A	A	T	T	T	C	T
T	A	G	A	T	G	T	C

A T A A T T T C
T C H A P T E R
A C T T T A A A
A C T G T T G 4
T A A A A C T T

Multiplex gene editing by CRISPR-Cas12a (Cpf1) using a single crRNA array

Bernd Zetsche*, Matthias Heidenreich*, Prarthana Mohanraju*, Iana Fedorova, Jeroen Kneppers, Ellen M DeGennaro, Nerges Winblad, Sourav R Choudhury, Omar O Abudayyeh, Jonathan S Gootenberg, Wen Y. Wu, David A. Scott, Konstantin Severinov, John van der Oost† & Feng Zhang†

*These authors contributed equally to this work.

†To whom correspondence should be addressed: F.Z. zhang@broadinstitute.org) or J.V.D.O. (john.vanderoost@wur.nl)

This chapter has been published as:

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E. M., Winblad N., Choudhury S.R., Abudayyeh O.O., Gootenberg J.S., Wu, W. Y., Scot D.A., Severinov K., van der Oost J. and Zhang F., (2017). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nature biotechnology, 35(1), 31-34.

Abstract

Targeting of multiple genomic loci with Cas9 is limited by the need for multiple or large expression constructs. Here we show that the ability of Cas12a to process its own CRISPR RNA (crRNA) can be used to simplify multiplexed genome editing. Using a single customized CRISPR array, we edit up to four genes in mammalian cells and three in the mouse brain, simultaneously.

Results

Although multiplex gene editing is possible with Cas9 nuclease, it requires relatively large constructs or simultaneous delivery of multiple plasmids (181–185), both of which are problematic for multiplex screens or *in vivo* applications. By contrast, the Cpf1 nuclease requires only one Pol III promoter to drive several small crRNAs (39 nucleotide (nt) per crRNA).

We confirmed *in vitro* that Cas12a (hereafter referred to by its previous name, Cpf1) alone is sufficient for maturation of crRNAs (43, 167) (Figure. 1a) using an artificial CRISPR pre-crRNA array consisting of four spacers separated by direct repeats from the CRISPR locus of *Francisella novicida* (FnCpf1) and two Cpf1 orthologs with activity in mammalian cells, *Acidaminococcus* Cpf1 (AsCpf1) and *Lachnospiraceae* Cpf1 (LbCpf1) (Figure. 1b and Supplementary Figure. 1). Small RNA-seq showed that AsCpf1 cleavage products correlate to fragments resulting from cuts at the 5' end of direct repeat hairpins, identical to the cleavage pattern we observed in *Escherichia coli* heterologously expressing FnCpf1 CRISPR systems (43) (Figure. 1c).

We further validated these results by generating AsCpf1 mutants that are unable to process arrays. Guided by the crystal structure of AsCpf1 (186), we mutated five conserved amino acid residues likely to disrupt array processing (H800A, K809A, K860A, F864A, and R790A)⁽¹⁸⁶⁾. All mutations interfered with pre-crRNA processing but not DNA cleavage activity *in vitro* (Figure. 1d and Supplementary Figure. 2a, b), an effect that was also observed for FnCpf1 (167). AsCpf1 recognizes specific nucleotides at the 5' flank of the direct repeat stem loop. Substitution of these nucleotides weakened or abolished RNA cleavage (Supplementary Figure. 3a). Dosage tests with the five AsCpf1 mutants revealed that mutants K809A, K860A, F864A, and R790A show pre-crRNA processing when used at high concentration (Supplementary Figure. 3b) or for extended incubation times (Supplementary Figure. 3c), but H800A was inactive regardless of dose and incubation time.

We next tested whether this mutant retains DNase activity in human embryonic kidney (HEK) 293T cells using three guides. Insertion/deletion (indel) frequency at the *DNMT1* and *GRIN2b* loci were identical for wild-type and H800A AsCpf1, and only slightly higher at the *VEGFA* locus in cells transfected with wild-type AsCpf1, demonstrating that the RNA and DNA cleavage activity can be separated in mammalian cells (Figure. 1e).

Cpf1-mediated RNA cleavage needs to be considered when designing lentivirus vectors for simultaneous expression of nuclease and guide (Figure. 1f). Lentiviruses carry a (+) strand RNA copy of the DNA sequence flanked by long terminal repeats, including the pre-crRNA, allowing Cpf1 to bind and cleave at the direct repeat sequence. Hence, reversing the orientation of the direct repeat is expected to result in (+) strand lentivirus RNAs not susceptible to Cpf1-mediated cleavage.

We designed a lentivirus encoding AsCpf1 and a crRNA expression cassette. We transduced HEK293T cells with a MOI (multiplicity of infection) of <0.3 and analysed indel frequencies in puromycin-selected cells 10 d after infection. Using guides encoded on a reversed expression cassette targeting *DNMT1*, *VEGFA*, or *GRIN2b* resulted in robust indel formation for each targeted gene (Figure 1g).

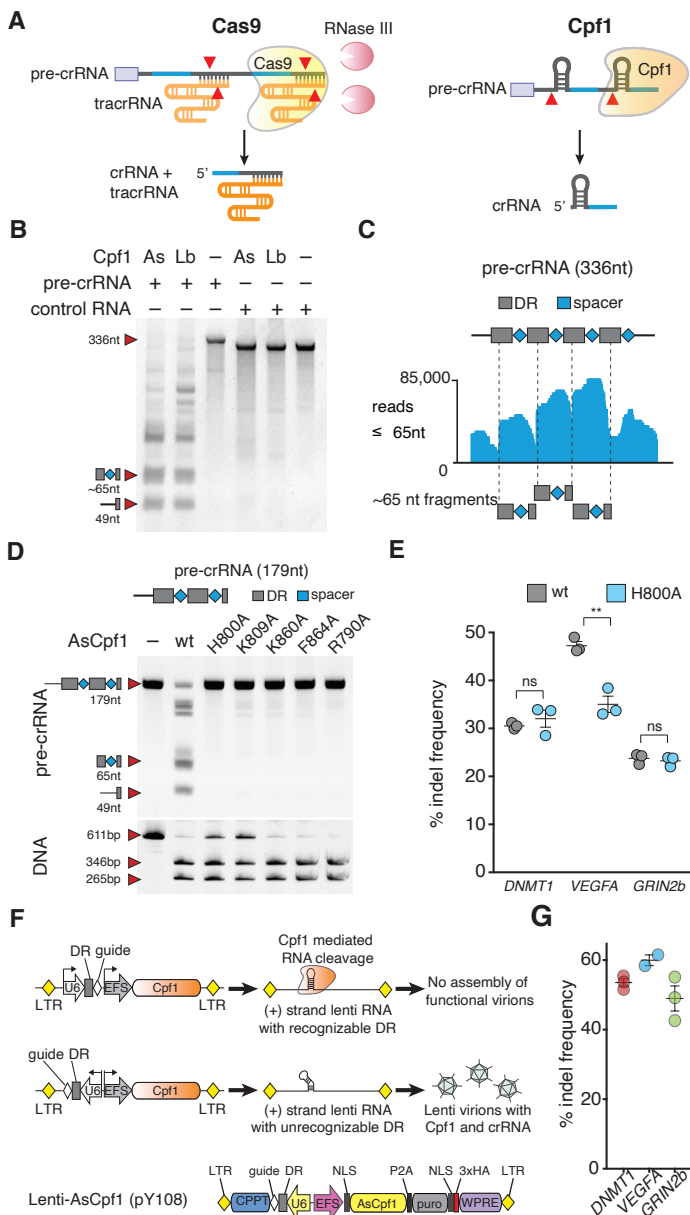


Figure 1 | Cpf1-mediated processing of pre-crRNA is independent of DNA cleavage. (a) Schematic of pre-crRNA processing for Cas9 and Cpf1. Cleavage sites indicated with red triangles. Trans-activating crRNA (tracrRNA). (b) *In vitro* processing of FnCpf1 pre-crRNA transcript (80 nM) with purified AsCpf1 or LbCpf1 protein (~320 nM), cropped gel image. (For full gel, see Supplementary Fig. 1.) (c) RNA-seq analysis of FnCpf1 pre-crRNA cleavage products, as shown in b. A high fraction of sequence reads smaller than 65 nt are cleavage products of spacers flanked by direct repeat sequences, cropped gel images. (d) Pre-crRNA (top) and DNA cleavage (bottom) mediated by AsCpf1 point mutants. H800A, K809A, K860A, F864A, and R790A fail to process pre-crRNA but retain DNA cleavage activity *in vitro*. 330 nM pre-crRNA was cleaved with 500 nM Cpf1 in 15 min and 25 nM DNA was cleaved with 165 nM Cpf1 in 30 min. (For full gels, see Supplementary Fig. 2.) (e) Indel frequencies mediated by AsCpf1H800A are comparable to wt AsCpf1, bars are mean of 3 technical replicates from one experiment, error bars are s.e.m. (Student t-test; n.s., not significant; **P = 0.003). (f) Schematic of lentivirus Cpf1 construct with the U6::direct repeat cassette in different orientations (top and middle), (+)-strand RNA copy with recognizable direct repeats are susceptible to Cpf1-mediated degradation, preventing functional virion formation. Schematic of AsCpf1 (pY108) construct (bottom). (g) Indel frequencies analysed by SURVEYOR nuclease assay after puromycin selection 10 d after transduction with lentivirus AsCpf1 in HEK cells. Horizontal bars are mean of 2 or 3 individual infections; error bars are mean \pm s.e.m. U6, Pol III promoter; CMV, cytomegalovirus promoter; NLS, nuclear localization signal; HA, hemagglutinin tag; DR, direct repeat sequence; P2A, porcine teschovirus-1 2A self-cleaving peptide; LTR, long terminal repeat; WPRE, woodchuck hepatitis virus post-transcriptional regulatory element.

We leveraged the simplicity of Cpf1 crRNA maturation to achieve multiplex genome editing in HEK293T cells using customized CRISPR arrays. We chose four guides targeting different genes (*DNMT1*, *EMX1*, *VEGFA*, and *GRIN2b*) and constructed three arrays with variant direct repeat and guide lengths for expression of pre-crRNAs (array 1, 19 DR with 23 nt guide; array 2, 19 nt DR with 30 nt guide; array 3, 35 nt DR with 30 nt guide; Figure. 2a). Indel events were detected at each targeted locus in cells transfected with array 1 or array 2. However, the crRNA targeting *EMX1* resulted in indel frequencies of <2% when expressed from array 3. Overall, array 1 performed best, with all guides showing indel levels comparable to those mediated by single crRNAs (Figure. 2b). Furthermore, small RNA-seq confirmed that autonomous, Cpf1-mediated pre-crRNA processing occurs in mammalian cells (Figure. 2c). Using arrays with guides in different orders resulted in similar indel frequencies, suggesting that positioning within an array is not crucial for activity (Supplementary Figure. 4a, b).

To confirm that multiplex editing occurs within single cells, we generated AsCpf1-P2A-GFP constructs to enable fluorescence-activated cell sorting (FACS) of transduced single cells (Figure. 2d) and clonal expansion. We used next-generation deep sequencing (NGS) to compare edited loci within clonal colonies derived from cells transfected with either pooled single guides or array 1. Focusing on targeted genes edited at every locus (indels \geq 95%) shows that multiplex editing occurs more frequently in colonies transfected with array 1 (6.4% all targets, 12.8% three targets, 48.7% two targets) than in pooled transfection (2.4% all targets, 3.6% three targets, 11.9% two targets) (Figure. 2e).

We next tested multiplex genome editing in neurons using AsCpf1. We designed a gene-delivery system based on adeno-associated viral vectors (AAVs) for expression of AsCpf1. We generated a dual vector system in which AsCpf1 and the CRISPR-Cpf1 array were cloned separately (Figure. 2f). We constructed a U6-promoter-driven

Cpf1 array targeting the neuronal genes *Mecp2*, *Nlgn3*, and *Drd1*. This plasmid also included a green fluorescent protein (GFP), fused to KASH nuclear transmembrane domain (187), in order to enable FACS of targeted cell nuclei (188).

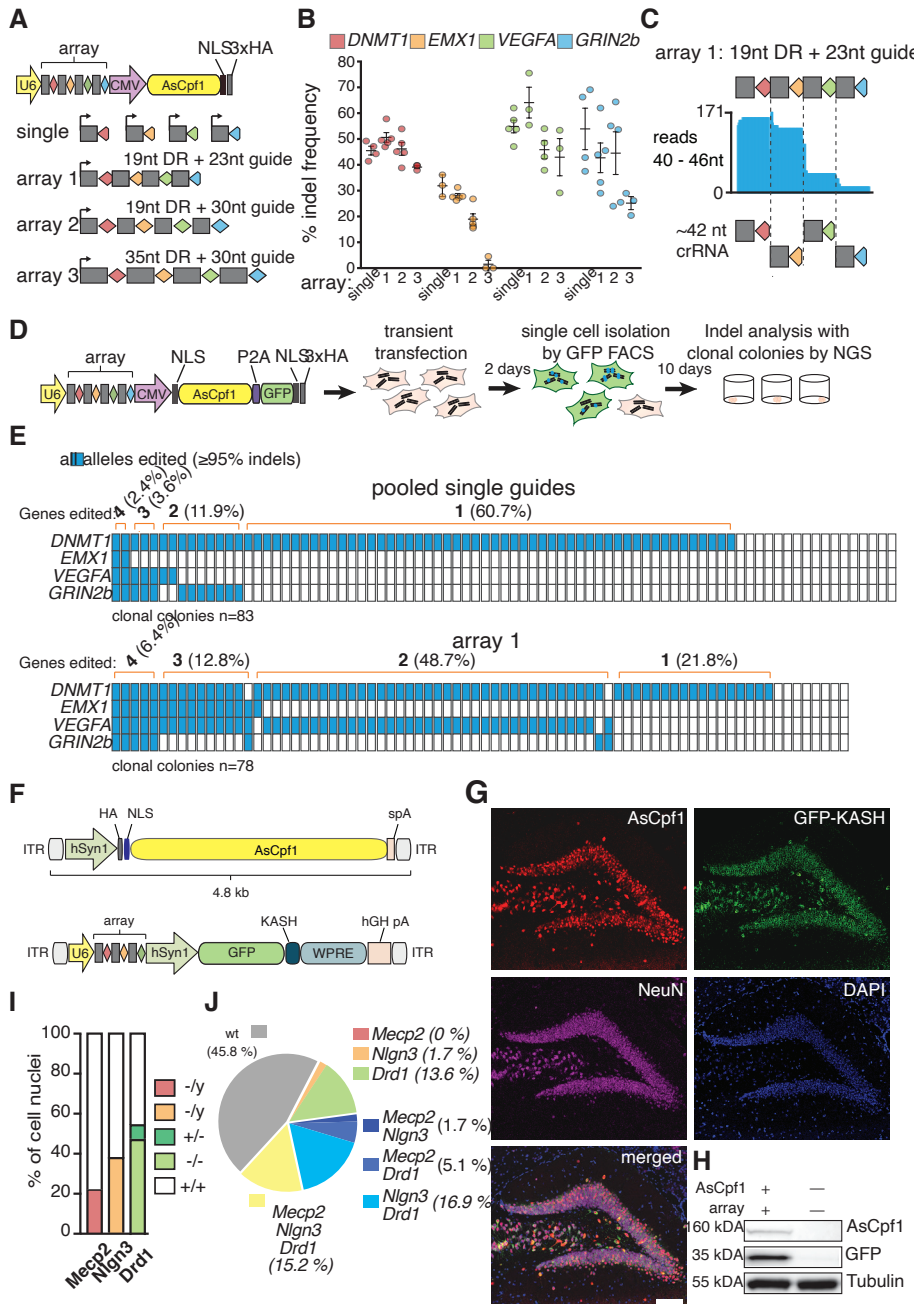


Figure 2 | Cpf1-mediated multiplex gene editing in mammalian cells and mouse brain. (a) Schematic of multiplex gene editing with AsCpf1, using a single plasmid approach. (b) Genome editing at four different genomic loci mediated by AsCpf1 with different versions of artificial CRISPR arrays (array 1, crRNAs in their mature form (19-nt DR with 23-nt guide); array 2, crRNAs are in an intermediate form (19-nt DR with 30-nt guide); array 3 crRNAs are in their unprocessed form (35-nt DR with 30-nt guides)). Indels were analysed by SURVEYOR nuclease assay 3 d after transfection. Horizontal bars are the means of two individual experiments with three to five technical replicates; error bars are mean \pm s.e.m. (c) Small RNA-seq reads from HEK cells transfected with AsCpf1 and array 1 show fragments corresponding to mature crRNA for each of the four guides. (d) Schematic for analysis of indel events in clonal colonies 48 h after transient transfection. (e) Quantification of indel events measured by NGS in clonal colonies from HEK cells transiently transfected with pooled single-guide-RNA plasmids or plasmid carrying array 1. Colonies were expanded for 10 d after sorting. Each column represents one clonal colony; blue rectangles indicate target genes with all alleles edited. (f) Schematic of AAV vector design for multiplex gene editing. Bottom: grey rectangles, direct repeat; diamonds, spacer (red: *Mecp2*, orange: *Nlgn3*, green: *Drd1*). (g) Immunostaining of dorsal DG 4 weeks after stereotactic AAV injection (representative image of $n = 4$ mice). Brain sections were co-stained with anti-HA (red), anti-GFP (green) and anti-NeuN (magenta) antibodies. Nuclei were labelled with DAPI (blue). Scale bar, 100 μ m. (h) Western blot analysis of DG expressing HA-AsCpf1 and GFP-KASH (representative blot from $n = 4$ mice). (i) Fraction of mono- (–/+), bi- (–/–) or maternal (–/y) allele editing for *Drd1* (autosomal), *Mecp2* and *Nlgn3* (x-chromosomal). (j) Analysis of multiplexing efficiency in individual cells. ITR, inverted terminal repeat; spA, synthetic polyadenylation signal; hSyn1, human synapsin 1 promoter; KASH, *K*larsicht *A*NC1 *S*yne1 *h*omology nuclear transmembrane domain; hGH pA, human growth hormone polyadenylation signal.

We first transduced mouse primary cortical neurons *in vitro* and observed robust expression of AsCpf1 and GFP-KASH 1 week after viral delivery. A SURVEYOR nuclease assay run on purified neuronal DNA confirmed indel formations in all three targeted genes (Supplementary Figure. 5). Next, we tested whether AsCpf1 could be expressed in the brains of living mice for multiplex genome editing *in vivo*. We stereotactically injected our dual vector system in a 1:1 ratio into the hippocampal dentate gyrus (DG) of adult male mice. Four weeks after viral delivery we observed robust expression of AsCpf1 and GFP-KASH in the DG (Figure. 2g, h). Consistent with previous studies (188, 189), we observed ~75% co-transduction efficiency of the dual viral vectors (Supplementary Figure. 5c). We isolated targeted DG cell nuclei by FACS (Supplementary Figure. 6) and quantified indel formation using NGS. We found indels in all three targeted loci with ~23%, ~38%, and ~51% indel formation in *Mecp2*, *Nlgn3*, and *Drd1*, respectively (Supplementary Figure. 5d, e). We quantified the effectiveness of bi-allelic disruption of the autosomal gene *Drd1* and found that ~47% of all sorted nuclei (i.e., ~87% of all *Drd1*-edited cells) harboured bi-allelic modifications (Figure. 2i). Next, we quantified the multiplex targeting efficiency in single neuronal nuclei. Our results show that ~15% of all transduced neurons were modified in all three targeted loci (Figure. 2j). Taken together, our results demonstrate the effectiveness of AAV-mediated delivery of AsCpf1 into the mammalian brain and simultaneous multi-gene targeting *in vivo* using a single array transcript.

Taken together, these data highlight the utility of Cpf1 array processing in designing simplified systems for *in vivo* multiplex gene editing. This system should simplify guide RNA delivery for many genome editing applications in which targeting of multiple genes is desirable.

Experimental procedures

Cpf1 protein purification

Humanized Cpf1 were cloned into a bacterial expression vector (6-His-MBP-TEV-Cpf1, a pET-based vector kindly given to us by Doug Daniels). Two litres of Terrific Broth growth media with 100 $\mu\text{g mL}^{-1}$ ampicillin was inoculated with 10 mL of an overnight culture of Rosetta (DE3) pLyseS (EMD Millipore) cells containing the Cpf1 expression construct. Growth media plus inoculant was grown at 37 °C until the cell density reached 0.2 OD₆₀₀, then the temperature was decreased to 21 °C. Growth was continued until OD₆₀₀ reached 0.6 when a final concentration of 500 μM IPTG was added to induce MBP-Cpf1 expression. The culture was induced for 14–18 h before harvesting cells and freezing at –80 °C until purification. Cell paste was resuspended in 200 mL of lysis buffer (50 mM HEPES pH 7, 2 M NaCl, 5 mM MgCl₂, 20 mM imidazole) supplemented with protease inhibitors (Roche cOmplete, EDTA-free) and lysozyme. Once homogenized, cells were lysed by sonication (Branson Sonifier 450), then centrifuged at 10,000g for 1 h to clear the lysate. The lysate was filtered through 0.22- μm filters (Millipore, Stericup) and applied to a nickel column (HisTrap FF, 5 mL), washed, and then eluted with a gradient of imidazole. Fractions containing protein of the expected size were pooled, TEV protease (Sigma) was added, and the sample was dialyzed overnight into TEV buffer (500 mM NaCl, 50 mM HEPES pH 7, 5 mM MgCl₂, 2 mM DTT). After dialysis, TEV cleavage was confirmed by SDS-PAGE, and the sample was concentrated to 500 μL before loading on a gel filtration column (HiLoad 16/600 Superdex 200) via FPLC (fast protein liquid chromatography, AKTA Pure). Fractions from gel filtration were analysed by SDS-PAGE; fractions containing Cpf1 were pooled and concentrated to 200 μL (50 mM Tris-HCl pH 7.5, 2 mM DTT, 5% glycerol, 500 mM NaCl) and either used directly for biochemical assays or frozen at –80 °C for storage.

In vitro synthesis of pre-crRNA arrays

Pre-crRNA arrays were synthesized using the HiScribe T7 High Yield RNA Synthesis Kit (NEB). PCR fragments coding for arrays, with a short T7-priming sequence on the 5' end, were used as templates for *in vitro* transcription reaction (Supplementary Table 1). T7 transcription was performed for 4 h and then RNA was purified using the MEGAclean Transcription Clean-Up Kit (Ambion).

In vitro cleavage assay

In vitro cleavage was performed with purified recombinant proteins for AsCpf1 and LbCpf1. Cpf1 protein and *in vitro*-transcribed pre-crRNA arrays were incubated at 37 °C in cleavage buffer (20 mM Tris HCl, 50 mM KCl supplemented with RNase Inhibitor Murine (NEB)) for 5 min to 1 h, as indicated in figure legends. Each cleavage reaction contained 20–630 nM of Cpf1 protein and 165 or 330 nM of synthesized pre-

crRNA array, as indicated in figure legends. For DNA cleavage, 25 nM of target was cleaved with 165 nM Cpf1 and 340 nM crRNA for 30 min at 37 °C. Reactions were stopped with proteinase K (Qiagen), heat denaturation and run on 10% TBE-Urea polyacrylamide gels. Gels were stained with SYBR Gold DNA stain (Life Technologies) for 10 min and imaged with a Gel Doc EZ gel imaging system (Bio-Rad).

Pre-crRNA array design and cloning

Guide sequences targeting human genes are listed in Supplementary Table 2. crRNAs were designed as four oligos (IDT) consisting of direct repeats, each one followed by a crRNA (Supplementary Table 3). The oligos favoured a one-directional annealing through their sticky-end design. The oligonucleotides (final concentration 10 µM) were annealed in 10×T4 ligase buffer (final concentration 1×; NEB) and T4 PNK (5 units; NEB). Thermocycler conditions were adjusted to 37 °C for 30 min, 95 °C for 5 min followed by a –5 °C/min ramp down to 25 °C. The annealed oligonucleotides were diluted 1:10 (final concentration 1 µM) and ligated into *BsmBI*-cut pcDNA-huAsCpf1-U6 (pY26), using T7 DNA ligase (Enzymatics), in room temperature for 30 min. The constructs were transformed into STBL3 bacteria and plated on ampicillin-containing (100 g mL⁻¹) agar plates. Single colonies were grown in standard LB media (Broad Facilities) for 16 h. Plasmid DNA was harvested from bacteria according to QIAquick Spin Miniprep protocol (QIAGEN).

Cell culture and transfection

Human embryonic kidney 293T (HEK293T) cell line (Life Technologies) were maintained in Dulbecco's modified Eagle's Medium (DMEM) + GLUTAMAX (Gibco) supplemented with 10% FBS (HyClone) at 37 °C with 5% CO₂ incubation. HEK293FT cells were seeded onto 24-well plates (Corning) 24 h before transfection. Cells were transfected using Lipofectamine 2000 (Life Technologies) at 70–80% confluency following the manufacturer's recommended protocol. For each well of a 24-well plate, a total of 500 ng plasmid DNA was used; each well represents one technical replicate.

Surveyor nuclease assay for genome modification

HEK293T cells were transfected with DNA, as described above. Cells were incubated at 37 °C for 72 h after transfection before genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were suspended in QuickExtract solution and incubated at 65 °C for 15 min, 68 °C for 15 min, and 98 °C for 10 min. The genomic region flanking the CRISPR target site for each gene was PCR amplified (primers listed in Supplementary Table 4), and products were purified using QIAquick PCR purification Kit (Qiagen) following the manufacturer's protocol. 200 ng total of the purified PCR products were mixed with 1 µl 10× Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 10 µl, and subjected

to a re-annealing process to enable heteroduplex formation: 95 °C for 10 min, 95 °C to 85 °C ramping at –2 °C/s, 85 °C to 25 °C at –0.25 °C/s, and 25 °C hold for 1 min. After re-annealing, products were treated with Surveyor nuclease and Surveyor enhancer S (IDT) following the manufacturer's recommended protocol and analysed on 10% Novex TBE polyacrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 10 min and imaged with a Gel Doc gel imaging system (Bio-Rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, $100 \times (1 - (1 - (b + c) / (a + b + c))^{1/2})$, where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Small RNA extraction from cells

HEK293T cells were harvested 48 h after transfection and the total RNA was extracted with the miRNeasy mini kit (Qiagen) according to manufacturer's conditions. rRNA was removed using the bacterial Ribo-Zero rRNA removal kit (Illumina).

NGS analysis of in vitro and in vivo cleavage pattern

RNA-seq libraries were prepared using a derivative of a previously described method (190). N. Dugar, G. Vogel, J. Sharma, C. M. Institute for Molecular Infection Biology (IMIB). Briefly, after PNK treatment in the absence and presence of ATP (enrichment of 5'OH and 3'P, respectively) RNA cleavage products were poly-A tailed with *E. coli* Poly(A) Polymerase (NEB), ligated to 5' RNA adapters using T4 RNA ligase I (NEB) and reverse transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent Technologies). cDNA was amplified by a fusion PCR method to attach the Illumina P5 adapters as well as unique sample-specific barcodes to the target amplicons (191). PCR products were purified by gel-extraction using QiaQuick PCR purification Kit (Qiagen) following the manufacturer's recommended protocol. DNA samples from single nuclei were pre-amplified with SURVEYOR primers (Supplementary Table 4) and nested-PCR was performed with NGS primers (Supplementary Table 5) before Illumina barcodes were added. Finally, barcoded and purified DNA samples were quantified by Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equi-molar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

RNA-sequencing analysis

The prepared cDNA libraries were pooled and sequenced on a MiSeq (Illumina). Pooled sequencing reads were assigned to their respective samples on the basis of their corresponding barcodes and aligned to the proper CRISPR array template sequence using BWA 3. Interval lists were generated using the paired-end alignment coordinates and the intervals were used to extract entire transcript sequences using Galaxy tools (<https://usegalaxy.org/>) (192). The extracted transcript sequences were analysed using Geneious 9.

AAV DNA constructs

The AAV hSyn1-HA-NLS-AsCpf1-spA vector was generated by PCR amplifying the AsCpf1 encoding sequence using forward PCR primer including HA-NLS (5'-aacacaggaccggtgccaccatgtaccatagatgttccagattacgcttcgccgaagaaaaagcgcaaggtcgaagcgtccacacagttcagggccttaccaacctgtatcaggtgagc-3')

and reverse PCR primer including a short poly A signal(spA) (5'-gcgggccgcacacaaaaaaccaacacacagatctaataaaataaagatctttattgaattcttagtgccagctcctggatgtaggccagcc-3') (188), and cloning of the resulting PCR template into AAV backbone under control of the human *Synapsin 1* promoter (hSyn1). For the generation of AAV U6-DR(*SapI*)-hSyn1-GFP-KASH-hGH (not shown) and U6-*Mecp2-Nlgn3-Drd1* array-hSyn1-GFP-KASH-hGH vectors, gene blocks (Integrated DNA Technologies) encoding U6-DR(*SapI*) and U6-*Mecp2-Nlgn3-Drd1* array, respectively, have been cloned into AAV hSyn-GFP-KASH-hGH backbone (188). All constructs were verified by Sanger sequencing.

Production of AAV vectors

AAV1 particles in DMEM culture medium were produced as described previously (189). Briefly, HEK293FT cells were transfected with transgene plasmid, AAV1 serotype plasmid and pDF6 helper plasmid using polyethylenimine (PEI). DMEM culture medium containing low-titre AAV1 particles was collected after 48 h and sterile filtered. For high-titre AAV1/2 production, HEK293FT cells were transfected with AAV1 and AAV2 serotype plasmids in equal ratios, transgene plasmid and pDF6 helper plasmid. 48 h after transfection, cells were harvested, and high-titre AAV1/2 virus was purified on heparin affinity column (189). The titre of AAV vectors was determined by real-time quantitative PCR (qPCR) using probe and primers specific for the hSyn1 promoter sequence (Integrated DNA Technologies).

Primary cortical neuron culture

Mice used to obtain neurons for tissue cultures were euthanized according to the protocol approved by the Broad's Institutional Animal Care and Use Committee (IACUC). Primary neurons were prepared from postnatal day P0.5 mouse brains and plated on laminin/poly-D-lysine-coated coverslips (VWR). Briefly, cortices were dissected in ice-cold HBSS (Sigma) containing 50 $\mu\text{g mL}^{-1}$ penicillin/streptomycin (Thermo Fisher) and incubated for 10 min at 37 °C with HBSS containing 125 Units papain (Worthington Biochemical) and 400 Units DNase I (Sigma). After enzymatic digestion, the tissues were washed twice in HBSS and gently triturated with a fire-polished Pasteur pipette. Cells were then transferred into neuronal growth medium (Neurobasal A medium, supplemented with B27, Glutamax (Life Technologies) and penicillin/streptomycin) and grown at 37 °C and 5% CO₂. For inhibition of glia cell proliferation, cytosine-beta-D-arabinofuranoside (AraC, Sigma) at a final concentration of 10 μM was added to the culture medium after 48 h and replaced by fresh culture medium after 72 h. For AAV1 transduction, cultured neurons were infected with low-

titre AAV1 as described previously (189). One week after transduction, neurons were harvested for isolating genomic DNA [QuickExtract DNA extraction buffer (Epicentre)] or fixed in 4% paraformaldehyde (PFA) for immunofluorescence staining.

Stereotactic injection of AAV1/2 into the mouse brain

The Broad's Institutional Animal Care and Use Committee (IACUC) approved all animal procedures described here. Craniotomy was performed on adult (12–16 weeks) male C57BL/6N mice according to approved procedures, and 1 μ l of 1:1 AAV mixture (AAV hSyn1-HA-NLS-AsCpf1-spA: 2.25×10^{12} Vg mL⁻¹; AAV U6-*Mecp2-Nlgn3-Drd1* array-hSyn1-GFP-KASH-hGH: 9.7×10^{12} Vg mL⁻¹) was injected into the dorsal dentate gyrus (anterior/posterior: -1.7; mediolateral: +/-0.6; dorsal/ventral: -2.15). The pipette was held in place for 3–5 min after injection to prevent leakage. After injection, the incision was sutured, and post-operative analgesics were administered according to approved IACUC protocol for 3 d following surgery.

Purification of cell nuclei from intact brain tissue

Cell nuclei from AAV1/2-injected hippocampal tissue were purified as described previously (188). Briefly, dissected tissue was homogenized in ice-cold homogenization buffer (HB) (320 mM sucrose, 5 mM CaCl₂, 3 mM Mg(Ac)₂, 10 mM Tris pH7.8, 0.1 mM EDTA, 0.1% NP40, 0.1 mM PMSF, 1 mM β -mercaptoethanol) using 2 ml type A and B Dounce homogenizer (Sigma). For gradient centrifugation, OptiPrep density gradient medium (Sigma) was used. Samples were centrifuged at 10,100g (7,500 r.p.m.) for 30 min at 4 °C (Beckman Coulter, SW28 rotor). Cell nuclei pellets were resuspended in 65 mM β -glycerophosphate (pH 7.0), 2 mM MgCl₂, 25 mM KCl, 340 mM sucrose, and 5% glycerol. Number and quality of purified nuclei was examined using bright-field microscopy.

FACS of cell nuclei

Purified cell nuclei were co-labeled with Vybrant DyeCycle Ruby Stain (1:500, Life Technologies) and sorted using a Beckman Coulter MoFlo Astrios EQ cell sorter (Broad Institute Flow Cytometry Core). Single and population (250–500 nuclei) GFP-KASH⁺ and GFP-KASH⁻ nuclei were collected in 96-well plates containing 5 μ l of QuickExtract DNA extraction buffer (Epicentre) and spun down at 2,000g for 2 min. Each 96-well plate included two empty wells as negative control.

Western blot analysis

AAV-injected dentate gyrus tissues were lysed in 100 μ l of ice-cold RIPA buffer (Cell Signalling Technologies) containing 0.1% SDS and protease inhibitors (Roche, Sigma) and sonicated in a Bioruptor sonicator (Diagenode) for 1 min. Protein concentration was determined using the BCA Protein Assay Kit (Pierce Biotechnology, Inc.). Protein samples were separated under reducing conditions on 4–15% Tris-HCl gels (Bio-

Rad) and analysed by western blotting using primary antibodies: mouse anti-HA (Cell Signalling Technologies 1:500), mouse anti-GFP (Roche, 1:500), rabbit anti-Tubulin (Cell Signalling Technologies, 1:10,000) followed by secondary anti-mouse and anti-rabbit HRP antibodies (Sigma-Aldrich, 1:10,000). Blots were imaged with Amersham Imager 600.

Immuno-fluorescent staining

4 weeks after viral delivery, mice were transcardially perfused with PBS followed by PFA according to approved IACUC protocol. 30 µm free-floating sections (Leica, VT1000S) were boiled for 2 min in sodium citrate buffer (10 mM tri-sodium citrate dehydrate, 0.05% Tween20, pH 6.0) and cooled down at RT for 20 min. Sections were blocked with 4% normal goat serum (NGS) in TBST (137 mM NaCl, 20 mM Tris pH 7.6, 0.2% Tween-20) for 1 h. Primary antibodies were diluted in TBST with 4% NGS and sections were incubated overnight at 4 °C. After three washes in TBST, samples were incubated with secondary antibodies for 1 hour at RT. After washing three times with TBST, sections were mounted using VECTASHIELD HardSet Mounting Medium including DAPI and visualized with confocal microscope (Zeiss LSM 710, Ax10 ImagerZ2, Zen 2012 Software). Following primary antibodies were used: mouse anti-NeuN (Millipore, 1:400); chicken anti-GFP (Aves Labs, 1:200–1:400); rabbit anti-HA (Cell Signalling Technologies, 1:100). Anti-HA signalling was amplified using biotinylated anti-rabbit (1:200) followed by streptavidin AlexaFluor 568 (1:500) (Life Technologies). Anti-chicken AlexaFluor488 and anti-mouse AlexaFluor647 secondary antibodies (Life Technologies) were used at 1:1,000.

Randomization and blinding

Neither randomization nor blinding were used in these experiments.

Accession codes

SRA: PRJNA354073

Acknowledgements

We would like to thank F.A. Ran for helpful discussions and overall support, and B. Cartigny and J. van den Bogaerde for technical assistance, and the entire Zhang laboratory for support and advice. 6-His-MBP-TEV, a pET-based vector, was kindly given to us by Doug Daniels of the Broad Institute. M.H. was supported by the Human Frontiers Scientific Program. O.A.A. is supported by a Paul and Daisy Soros Fellowship and a Friends of the McGovern Institute Fellowship. J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. E.M.D.G. is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), of the National Institutes of Health (5T32EB1680). K.S. is supported by an NIH grant GM10407, Russian Science Foundation grant 14-14-00988, and Skoltech. J.v.d.O. is supported by Netherlands Organization for Scientific Research (NWO) through a TOP grant (714.015.001). F.Z. is supported by the NIH through NIMH (5DP1-MH100706 and 1R01-MH110049); by the New York Stem Cell, Poitras, Simons, Paul G. Allen Family, and Vallee Foundations; and by David R. Cheng, Tom Harriman, and B. Metcalfe. F.Z. is a New York Stem Cell Foundation Robertson Investigator. The authors plan to make the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab website (<http://www.genome-engineering.org/>).

Author contributions

B.Z., M.H., J.v.d.O., and F.Z. conceived this study and designed the experiments. B.Z., M.H., P.M., I.F., J.K., E.M.D., N.W., S.R.C., O.O.A., J.S.G., W.Y.W. and D.A.S. conducted the experiments. K.S., J.v.d.O., and F.Z. supervised this project. B.Z., M.H., J.v.d.O., and F.Z. wrote the manuscript with input from all authors.

Competing interests

A patent has been filed relating to the presented data. F.Z. is a founder and scientific advisor for Editas Medicine, and a scientific advisor for Horizon Discovery.

Supplementary Figures & Tables

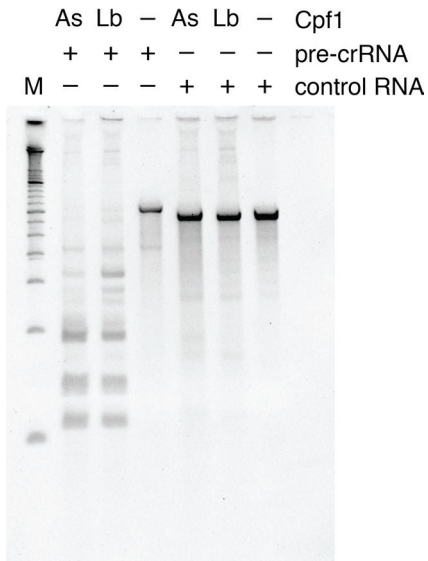


Figure S1 | Full gel image of figure 1b. Full gel image for in vitro processing of FnCpf1 pre-crRNA transcript with purified AsCpf1 or LpCpf1 protein. M = DNA standard.

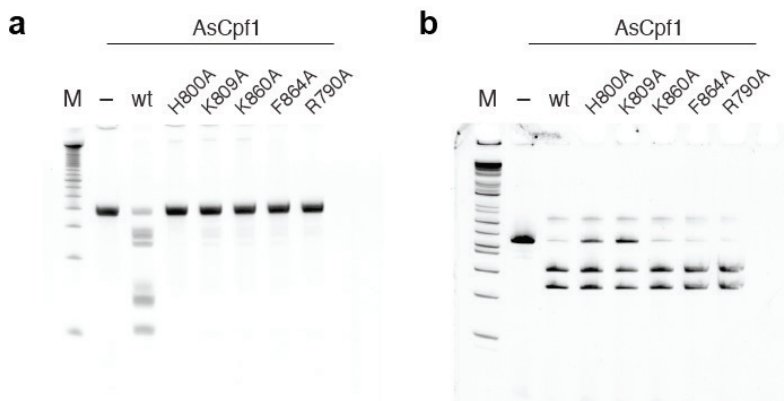


Figure S2 | Full gel images of figure 1d. (a) Full gel image for pre-crRNA cleavage. (b) Full gel image for DNA cleavage. M = DNA standard.

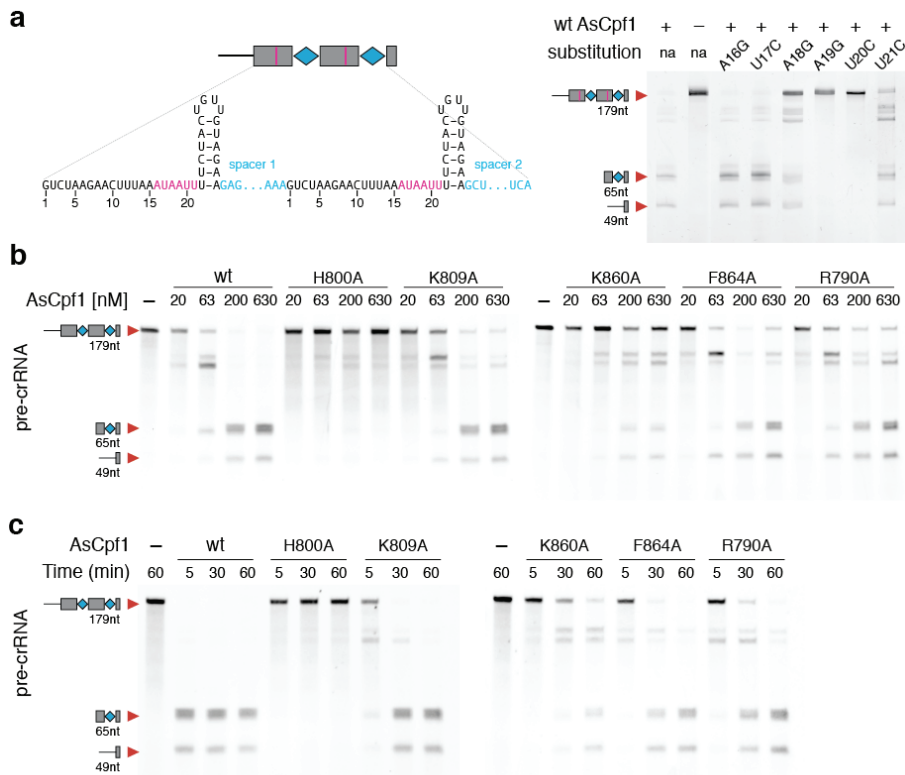


Figure S3 | Cpf1 mediated pre-crRNA cleavage is sequence and dose dependent. (a) Cpf1 mediated pre-crRNA processing is sequence dependent. Single nucleotide substitutions at position A19 and U20 abolish RNA cleavage *in vitro*. 200 nM pre-crRNA was cleaved with 500 nM Cpf1 in 1 hour. **(b, c)** AsCpf1 point mutants, with the exception of H800A, are active at high dose. **(c)** Titration of AsCpf1 mutants reveals pre-crRNA processing at high AsCpf1 protein concentration. **(d)** Prolonged incubation time allows pre-crRNA processing by AsCpf1 point mutants. Only H800A does not process pre-crRNA to mature crRNA at high dose. 165 nM pre-crRNA was incubated with the indicated concentration **(c)** or with 500 nM AsCpf1 protein **(d)** for 30 min.

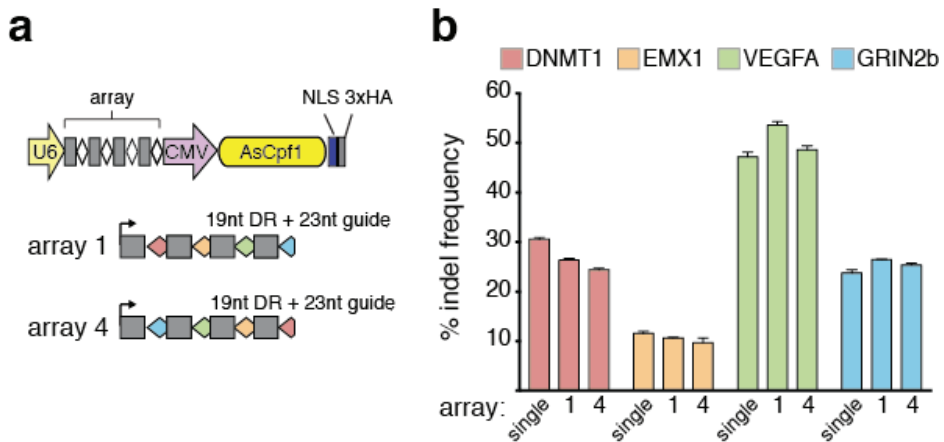


Figure S4 | Indel levels are not influenced by guide order. (a) Schematic of multiplex gene editing with AsCpf1, using a single plasmid approach. Two arrays with guides in reversed order are compared (array-1 and array-4). (b) Quantification of indel frequencies measured by Surveyor nuclease assay. Guides expressed from array-1 and array-4 result in similar indel frequencies for each targeted gene.

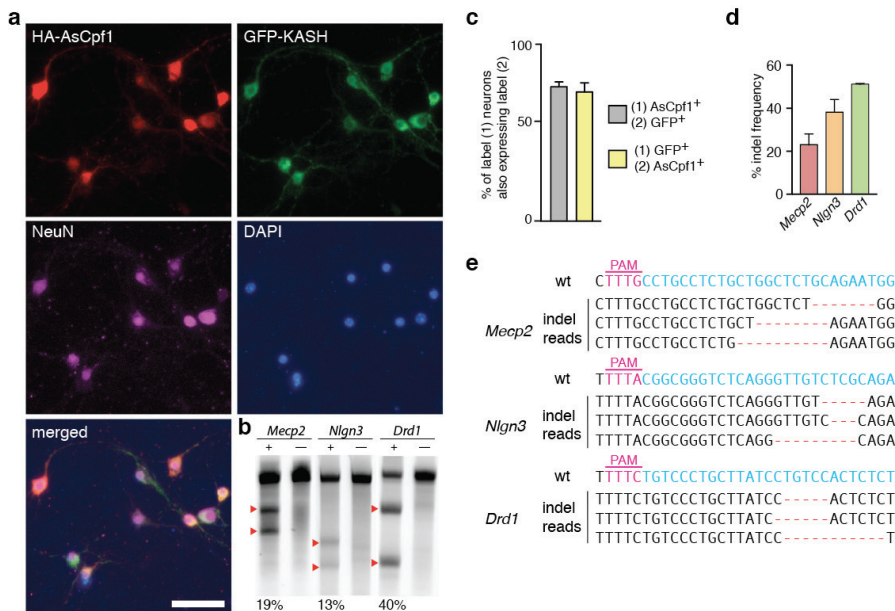


Figure S5 | AAV delivery of AsCpf1 and multiplex gene editing in primary neurons and mouse brain. (a) Immunostaining of AsCpf1 (anti-HA antibody, red) and GFP-KASH (anti-GFP antibody, green) in primary cortical neurons (anti-NeuN antibody, magenta) 7 days after viral infection with dual vector system. Nuclei were labelled with DAPI (blue). Scale bar: 25 μ m. (b) SURVEYOR nuclease assay showing indel formations (+) in all 3 targeted loci. Control neurons (-) were infected with AsCpf1 only (Bottom: Indel percentage; representative images from $n = 3$ independent experiments). (c)

Quantification of dentate gyrus neurons (DG) efficiently transduced by the dual- vector system *in vivo* ($n = 581$ nuclei from 3 mice). **(d)** NGS indel analysis of modified *Mecp2*, *Ngn3* and *Drd1* loci in single DG nuclei ($n = 59$ cells from 2 male mice, error bars represent mean \pm SEM). **(e)** Representative mutation patterns detected by NGS. Blue, wild-type (wt) sequence; red dashes, deleted bases; PAM sequence marked in magenta.

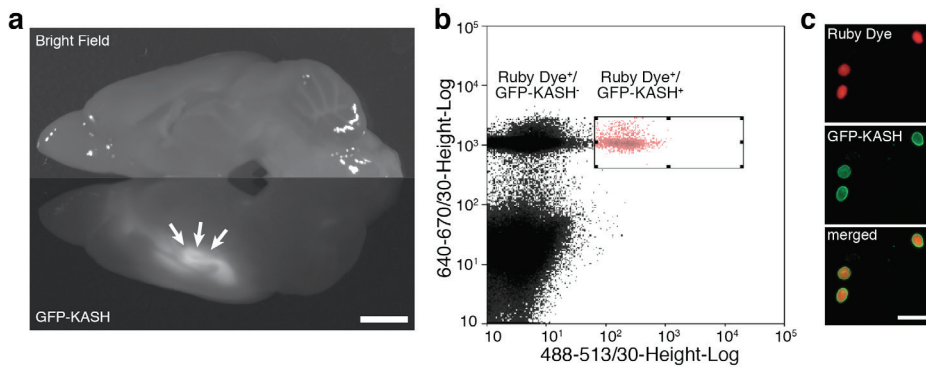


Figure S6. *In vivo* delivery of AAV dual vector system and sorting of targeted cell nuclei from intact brain. **(a)** Sagittal dissection of adult mouse brain 4 weeks after viral delivery shows infected hippocampal formation (bottom). **(b)** Representative FACS plot showing Ruby Dye⁺/GFP-KASH⁻ and Ruby Dye⁺/GFP-KASH⁺ nuclei populations. **(c)** Representative images of sorted Ruby Dye⁺/GFP-KASH⁺ nuclei used for NGS indel analysis. Scale bars: 2 mm in (a), 25 μ m in (c).

Table S1. Sequences of pre-crRNA arrays used for *in vitro* cleavage reaction.

4 spacer pre- crRNA	GGGGGUCUUUUUUGCUGAUUUUAGGC AAAACGGGUCUAAGAACUUUAAAUAUUUCUACU GUUGUAGAUGAGAAGUCAUUUAAUAAGGCCACUGUUAAAAGUCUAAGAACUUUAAAUAUU UCUACUGUUGUAGAUGCUCUAUUCUGUGCCUUCAGAUAAUUCAGUCUAAGAACUUUAAA UAAUUUCUACUGUUGUAGAUGCUCUAGAGCCUUUUGAUUAGUAGCCGGUCUAAGAACUUUA AAUAAUUUCUACUGUUGUAGAUAUAGCGAUUUUAUGAAGGUCAUUUUUUUGUCUAGCUUUAAU GCGGUAGUUUAUCACAGUUAAAUUGCUAACG
2 spacer pre-crRNA	UAGGUCUUUUUUGCUGAUUUUAGGC AAAACGGGUCUAAGAACUUUAAAUAUUUCUACUG UUGUAGAUGAGAAGUCAUUUAAUAAGGCCACUGUUAAAAGUCUAAGAACUUUAAAUAUUU CUACUGUUGUAGAUGCUCUAUUCUGUGCCUUCAGAUAAUUC
control RNA	UACGCCAGCUGGCGAAAGGGGAUGUGCUGCAAGGCGAUUAGUUGGGUAACGCCAGGGUU UUCCAGUCACGACGUUGUAAAACGACGGCCAGUGAAUUCGAGCUCGGUACCGGGNNNNN NNNGAGAGUCAUUUAAUAAGGCCACUGUUAAAAGCUUGGCGUAAUCAUGGUCUAGCUG UUUCCUGUGUAGAAUUGUUAUCCGCUCACAAUCCACACAUAUACGAGCCGAAGCAUAA AGUGUAAAGCCUGGGUGGCCUAAUGAGUGAGCUAACUCACAUUAAUUGCGUU

Table S2. Cpf1 guide sequences used for single and pre-crRNA array expression.

DNMT1 23 nt guide	CTGATGGTCCATGTCTGTTACTC
EMX1 23 nt guide	TGGTTGCCACCCCTAGTCATTGG
VEGFA 23 nt guide	CTAGGAATATTGAAGGGGGCAGG
GRIN2b 23 nt guide	GTGCTCAATGAAAGGAGATAAGG
DNMT1 30 nt guide	CTGATGGTCCATGTCTGTTACTCGCCTGTC
EMX1 30 nt guide	TGGTTGCCACCCCTAGTCATTGGAGGTGAC
VEGFA 30 nt guide	CTAGGAATATTGAAGGGGGCAGGGGAAGGC
GRIN2b 30 nt guide	GTGCTCAATGAAAGGAGATAAGGTCCCTGA

Table S3. DNA oligonucleotides for array cloning.

array 1 T1	AGATCTGATGGTCCATGTCTGTTACTCAATTTCTACTCTTGTAGATTGGTTGCCAC
array 1 T2	CCTAGTCATTGGAATTTCTACTCTTGTAGATCTAGGAATATTGAAGGGGCAGGAATTTCTACTCTTGTAGA TGTGCTCAATGAAAGGAGATAAGG
array 1 B1	AAAACCTTATCTCCTTTTCATTGAGCACATCTACAAGAGTAGAAATTCCTGCCCTT
array 1 B2	CAATATTCCTAGATCTACAAGAGTAGAAATTCGAATGACTAGGGTGGGCAACCAATCTACAAGAGTAGAAAT TGAGTAACAGACATGGACCATCAG
array 2 T1	AGATCTGATGGTCCATGTCTGTTACTCGCCTGTCAATTTCTACTCTTGTAGATTGGTTGCCACCCCTAGTC
array 2 T2	TGAAGGGGGCAGGGGAAGGCAATTTCTACTCTTGTAGATGTGCTCAATGAAAGGAGATAAGGTCTTGA
array 2 B1	AAAATCAAGACCTTATCTCCTTTTCATTGAGCACATCTACAAGAGTAGAAATTCCTTCCCTGCCCTT
array 2 B2	CAATATTCCTAGATCTACAAGAGTAGAAATGTCACCTCCAATGACTAGGGTGGGCAACCAATCTACAAGA GTAGAAATGACAGCGAGTAACAGACATGGACCATCAG
array 3 T1	AGATGTCAAAGACCTTTTAAATTTCTACTCTTGTAGATCTGATGGTCCATGTCTGTTACTCGCCTGTGTC AAAAGACCTTTTAAATTTCTACTCTTGTAGATTGGTTGCCACCCCTAGTCATTGGAGGTGACGTCAAAGA CCTTTTAAATTTCTACTCTTGTAGATCTAGGAATATT
array 3 T2	GAAGGGGGCAGGGGAAGGCGTCAAAGACCTTTTAAATTT CTACTCTTGTAGATGTGCTCAATGAAAGG AGATAAGGTCTTTGAGTCAAAGACCTTTTAAATTTCTACTCTTGTAGAT
array 3 B1	AGAAATTAAGAGGTCTTTTGACGCTTCCCTGCCCTTCAATATTCCTAGATCTACAAGAGTAGAAAT TAAAAAGGTCTTTTGACGTACCTCCAA
array 3 B2	TGACTAGGGTGGGCAACCAATCTACAAGAGTAGAAATTAAGAGGTCTTTTGACGACAGCGAGTAACA GACATGGACCATCAGATCTACAAGAGTAGAAATTAAGAGGTCTTTTGAC
array 4 T1	AGATGTGCTCAATGAAAGGAGATAAGGAATTTCTACTCTTGTAGATCTAGGAATATT
array 4 T2	GAAGGGGGCAGGAATTTCTACTCTTGTAGATTGGTTGCCACCCCTAGTCATTGGAATTTCTACTCTTGTGA GATCTGATGGTCCATGTCTGTTACTC
array 4 B1	AAAAGAGTAACAGACATGGACCATCAGATCTACAAGAGTAGAAATTCGAATGACTAG
array 4 B2	GGTGGGCAACCAATCTACAAGAGTAGAAATTCCTGCCCTTCAATATTCCTAGATCTACAAGAGTAGAA ATTCCTTATCTCCTTTTCATTGAGCAC

Table S4. PCR primers for amplification of DNA regions for SURVEYOR nuclease assay.

DNMT1 FW	CTGGGACTCAGGCGGGTCAC
DNMT1 RV	CCTCACACAACAGCTTCATGTCAGC
EMX1 FW	CCATCCCCTTCTGTGAATGT
EMX1 RV	GGAGATTGGAGACACGGAGA
VEGFA FW	CTCAGCTCCACAACTTGGTGCC
VEGFA RV	AGCCCGCCGCAATGAAGG
GRIN2b FW	GCATACTCGCATGGCTACCT
GRIN2b RV	CTCCCTGCAGCCCCTTTTA
Mecp2 FW	GGTCTCATGTGTGGCACTCA
Mecp2 RV	TGTCCAACCTTCAGGCAAGG
Nlgn3 FW	GTAACGTCTGGACACTGTGG
Nlgn3 RV	TTGGTCCAATAGGTCATGACG
Drd1 FW	TGGCTAAGCCTGGCCAAGAACG
Drd1 RV	TCAGGATGAAGGCTGCCTTCGG

Table S5. PCR primers for amplification of DNA regions for next generation sequencing.

NGS DNMT1 FW	CCATCTCATCCCTGCGTGTCTCCTGAACGTTCCCTTAGCACTCTGCC
NGS DNMT1 RV	CCTCTCTATGGGCAGTCGGTGATGCCTTAGCAGCTTCCTCCTCC
NGS EMX1 FW	CCATCTCATCCCTGCGTGTCTCCGGGCTCCCATCACATCAACCG
NGS EMX1 RV	CCTCTCTATGGGCAGTCGGTGATGCCAGAGTCCAGCTTGGGCCC
NGS VEGFA FW	CCATCTCATCCCTGCGTGTCTCCAGGGGTCAGTCCAGGATTCCA
NGS VEGFA RV	CCTCTCTATGGGCAGTCGGTGATGCATTGGCGAGGAGGAGCAG
NGS GRIN2b FW	CCATCTCATCCCTGCGTGTCTCCGTTCAAGGATTTCTGAGGCTTTTGAAAG
NGS GRIN2b RV	CCTCTCTATGGGCAGTCGGTGATGGGGCTTCATCTTCAACTCGTCGAC
NGS Mecp2 FW	CCATCTCATCCCTGCGTGTCTCCGAAAAGTCAGAAGACCAGG
NGS Mecp2 RV	CCTCTCTATGGGCAGTCGGTGATGGTGGGGTCATCATACATAGG
NGS Nlgn3 FW	CCATCTCATCCCTGCGTGTCTCCACCCCGAGGATGGTGTCTCG
NGS Nlgn3 RV	CCTCTCTATGGGCAGTCGGTGATGGGTAGAAGGCGTAGAAGTAGG
NGS Drd1 FW	CCATCTCATCCCTGCGTGTCTCCAAGCCACCGAAGTGCTTTCC
NGS Drd1 RV	CCTCTCTATGGGCAGTCGGTGATGCACAGCTTCCAGGGCATGACC

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
T	C	G	A	T	G	G	G
C	C	C	T	C	C	A	G
G	A	A	C	C	T	G	G

A T A A T T T C
T C H A P T E R
A A A C C T T A
C T C C A C T 5
G G A A T A T T

Cut and paste: genome editing of *E. coli* using Cas12a and T4 ligase

Wen Y. Wu, Jorik Bot, Sjoerd C.A. Creutzburg, Raymond H.J. Staals, John van der Oost[†]

[†]To whom correspondence should be addressed: J.V.D.O. (john.vanderoost@wur.nl)

Abstract

Genome editing methods for *Escherichia coli* are of high importance for both fundamental and applied research. Currently, various methods exist for genome editing in *E. coli* using homologous recombination. Here, we demonstrate the proof of concept of a novel genome editing method, termed “cut & paste”, which utilizes the Cas12a nuclease of the type V-A CRISPR-Cas system. Cas12a targets and cleaves at two selected locations within the genome. Cleavage by Cas12a generates double-stranded DNA breaks with 4-5 nt compatible staggered ends that can be repaired by ligation using T4 ligase. As a prove of concept, a genomic deletion in *E. coli* by cut & paste was successfully achieved in this study, however, with a relatively low editing efficiency. Further improvements of the system are required to make cut & paste an efficient editing tool to generate accurate genomic deletions in prokaryotes.

Introduction

For many years *Escherichia coli* has been a convenient model organism for both fundamental and applied research. Therefore, precise, fast and efficient genome editing techniques for *E. coli* are essential. Until now, various methods have been developed for genome editing in *E. coli*, such as group II intron retro-homing, cre-lox recombination and lambda red mediated recombineering (193-196). The latter is currently the most applied method, as it allows for easy and efficient insertions and deletions using either a dsDNA (PCR product) or ssDNA recombination template (oligo). Lambda red recombineering functions by having the ssDNA repair template anneal to the lagging strand during replication (197). Although lambda red recombineering made genome editing more efficient, the editing efficiency remains low (<1%). In addition, to find the correct edited clone, a large amount of cells need to be screened, for instance by PCR (198). One solution is to include an antibiotic marker in the recombination template, so selection for correct recombinants is based on selection for antibiotic resistance (199). However, to make a markerless strain, an additional recombination step is needed to remove the antibiotic marker.

Another recombination approach utilizes the sequence-specific nucleases of the powerful Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) – CRISPR associated (Cas) system. CRISPR-Cas systems are divided in two classes. Class 1 systems utilize proteins complexes consisting of multiple subunits to mediate the sequence-specific DNA cleavage, whereas class 2 use a single-subunit effector protein (18). The first CRISPR-Cas protein used in combination with the lambda red system was Cas9, a class 2, type II CRISPR-Cas system (200). The Cas9 nuclease uses two RNAs, a CRISPR RNA (crRNA) guide and a trans-activating CRISPR RNA (tracrRNA) anchor, which can be synthetically fused together to form a single-guide RNA (sgRNA) (168). A Cas9-sgRNA complex uses its guide to find a complementary dsDNA target, also known as a protospacer. Upon the initial recognition of a protospacer adjacent motif (PAM), local unwinding of the upstream DNA sequence occurs that allows the spacer part of the crRNA to displace the non-target DNA strand, and to base pair with the target DNA strand. Complete base pairing of guide and target trigger activation of the two nuclease domains, resulting in cleavage of both DNA strands to generate blunt ends at the PAM proximal end (168). In eukaryotes and only a few prokaryotes, double stranded breaks can be repaired by non-homologous end joining, which ligates both DNA ends together with small insertions or deletions at the ligated location (201). However, *E. coli* does not contain an endogenous NHEJ system, and hence is unable to repair double-stranded breaks, implying that they will not survive the cleavage by Cas9. So, when used in combination with the lambda red system, the nuclease activity of Cas9 allows for counterselection by cleaving the unmodified, wild-type DNA, thereby selecting for the desired recombinant DNA. Indeed, it has been demonstrated that Cas9 in combination with lambda red allows for fast, efficient and markerless genome editing (78, 202).

Subsequently, other CRISPR-Cas systems have emerged, such as the Cas12a (Cpf1) nuclease of the class 2, type V-A system (79, 203). Cas12a has also been used as a counterselection strategy in combination with the lambda red system. This yielded similar editing efficiencies as Cas9 ranging from 80 to 100% for integration at a single locus and around 20% for integration at three different loci simultaneously (204, 205). Although Cas9 and Cas12a function similarly in terms of providing counter selective pressure, their mechanistic properties differ substantially. Cas12a does not require a tracrRNA since Cas12a is able to autonomously process its pre-crRNA (the precursor transcript of a CRISPR array) into mature crRNA guides. In addition, Cas12a recognizes a 5' T-rich PAM and cleaves dsDNA to generate 5 nt staggered ends or "sticky ends" at the PAM distal end, at position 18-23 of the protospacer (53, 79). This latter property can be exploited for a new and alternative genome editing method in *E. coli*. Just like transferring DNA fragments in and out a plasmid using appropriate restriction enzymes, a similar approach could be executed using Cas12a but then, because of the 20 nt recognition site, at genome level. This was done in eukaryotes, which showed increased precise targeted integrations compared to Cas9 (206). For generating a genomic knock-out, Cas12a can generate two double strand breaks with compatible sticky ends at selected genomic locations, that can be recombined and covalently linked using a ligase, either from the host or a heterologous one (T4 ligase) (207, 208). To generate a genomic knock-in, on the other hand, DNA templates supplied in trans can be designed to have sticky-ends compatible to one or two breaks introduced by Cas12a in the genome. Here, we describe and show a proof of a concept of "cut & paste", as a new genome editing method in *E. coli* using Cas12a and T4 ligase for generating accurate genomic deletions.

Results

Plasmid reconstruction in vivo by the cut & paste system

The cut & paste method was first tested in a three-plasmid setup: pCas, pDonor and pAcceptor (Fig. 1A). pCas contains an operon consisting of *cas12a*, T4 ligase (*ligT4*) and the CRISPR array under a single rhamnose inducible promoter. The CRISPR array consists of two spacers, "Sp1" and "Sp2", which target "PS1" (pAcceptor) and "PS2" (pDonor), respectively. pAcceptor contains the 5' half of a chloramphenicol resistance gene (*cat-M*) and a *lacZα* gene, flanked by identical protospacers. "PS1", whereas pDonor contains the 3' half of the *cat* gene (*cat-C*), flanked by another set of identical protospacers, "PS2". Cleavage of PS1 and PS2 by Cas12a generates compatible sticky ends consisting of 5'CTCCA (top strand) and 5'TGGAG (bottom strand), respectively. Both sticky ends of the donor fragment are compatible to the generated gap in pAcceptor, so ligation of *cat-C* into pAcceptor can occur in either

orientation. This means that 50% of *cat-C* insertion will restore chloramphenicol resistance (Fig. 1A). *E. coli* cells harboring all three plasmids were cultured for 5 days, expressing Cas12a and T4 ligase with rhamnose induction and selecting for pCas (KanR) and pAcceptor (AmpR). Each day, cultures were inoculated into fresh medium containing kanamycin, ampicillin and rhamnose, but also in medium containing chloramphenicol to select for correctly edited pAcceptor plasmids (Fig. 1B). After 5 days, no growth was observed in any of the erlenmeyers with chloramphenicol-containing medium, meaning no successful cut & paste had taken place. To check whether cleavage by Cas12a was occurring, an *in vivo* plasmid loss and an *in vitro* cleavage assay (using purified Cas12) were performed (Fig. 1C, D) showing that both plasmids were indeed successfully cleaved by Cas12a. However, the pAcceptor is cleaved more slightly efficiently than pDonor, with this effect being more pronounced *in vitro* (Fig. 1D), potentially explaining the absence of successful recombinants.

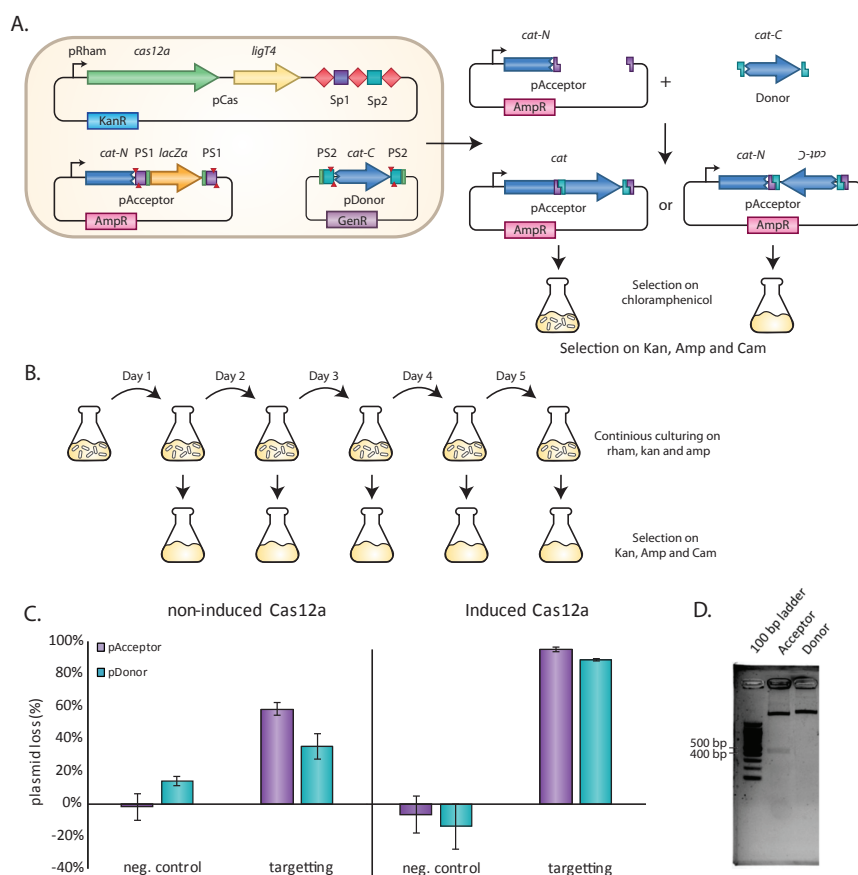


Figure 1 | Plasmid engineering using the cut & paste system. (A) Schematic of the cut & paste system using the three-plasmid setup, consisting of pCas, pAcceptor and pDonor. pCas expresses *cas12a*, T4 ligase (*ligT4*) and the CRISPR array under a single rhamnose inducible promoter (pRham). The CRISPR array contains two spacers, “Sp1” and “Sp2”, complementary to protospacer 1 and 2 (“PS1” and “PS2”),

respectively. pAcceptor has an incomplete chloramphenicol resistance gene (*cat-N*) and a *lacZα* flanked by two PS1 sequences. pDonor contains the other half of the chloramphenicol resistance gene (*cat-C*) flanked by two PS2 sequences. **(B)** Workflow of cut & paste plasmid reconstruction. Cells were cultured in medium containing rhamnose (rham), kanamycin (kan) and ampicillin (amp) and re-inoculated in fresh culture every day for five days. Each day, cells were also inoculated in medium containing kanamycin (kan), ampicillin (amp) and chloramphenicol (cam) to select for correctly ligated pAcceptor. **(C)** Plasmid loss assay of pAcceptor and pDonor, in conditions with Cas12a (rhamnose induced) and without (non-induced). pCas containing no CRISPR array was used as negative control (neg. control). Y-axis represents plasmid loss in %. Error bars were calculated using three biological replicates (n=3). **(D)** *In vitro* cleavage assay of pAcceptor and pDonor. Black arrow indicates cleavage product. 100bp NEB ladder was used. Expected products for pAcceptor are 450 bp and 2670 bp. Expected products for pDonor are 364 bp and 2968 bp.

Selecting spacers with high cleavage efficiency

Unsuccessful *in vivo* plasmid reconstruction by the cut & paste was possibly due to the lower cleavage efficiency of pDonor, meaning a different spacer sequence should be used. To obtain a spacer with a high cleavage efficiency, six randomly generated spacers (spacers 3-8) containing similar GC content, were tested *in vitro*. The *in vitro* cleavage assay was measured in a time series of 0, 10, 20 and 50 min, then visualized on agarose gel (Fig. 2A). Subsequently, cleavage of a linear fragment containing the protospacer was quantified (Fig. 2B). Out of the six spacers, spacer 3 performed the best with complete cleavage observed around ten minutes, followed by spacer 5 and spacer 6. Spacer 4 had a moderate cleavage efficiency with only half its targets cleaved after 50 minutes, whereas spacer 7 and spacer 8 had the lowest cleavage efficiencies with little to no targets cleaved after 50 minutes. Since spacer 3 had the highest cleavage efficiency, it was used to replace spacer 2 in subsequent “cut & paste” assays.

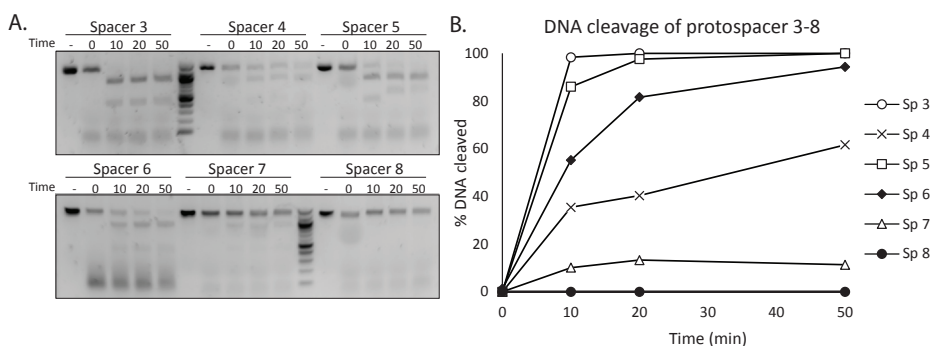


Figure 2 | *In vitro* cleavage assay of 6 different spacers. **(A)** Agarose gel electrophoresis analysis of cleavage products generated by Cas12a loaded with crRNA containing a spacer variant (3-8) using 3nM linear targets (1334 bp) in a time series of 0, 10, 20 and 50 minutes. Cleavage by Cas12a results in products consisting of 419 bp and 916 bp long. 100bp NEB ladder was used **(B)** Quantification of the results presented in panel A, % of cleaved DNA with the time presented on the X-axis (in minutes).

Gene deletion by cut & paste

As described above, attempts to reconstruct a plasmid *in vivo* by cut & paste were not successful. To reduce the editing complexity of the system, we first need to reduce the number of targets, i.e. the copy number. Therefore, the experimental design shifted to a genomic deletion. A deletion excludes the need for a compatible DNA fragment to be inserted, which further reduces the editing complexity. To increase T4 expression, the T4 ligase was placed under the control of a constitutive promoter (Pbla). In addition, an internal terminator-like sequence was found within the T4 ligase gene using ARNold (209). The internal terminator like-sequence is found at position 514-534 of the open reading frame, which could have hindered expression. The internal terminator-like sequence was removed by silent mutations (T513G, T534C and T538G) (T4 Δ term). A strain was used in which a genomic *gfp* sequence, flanked by protospacers 1 and 3 ("PS1" and "PS3"), was inserted into the *thyA* gene, thereby disrupting its reading frame. ThyA catalyzes the conversion of 2'-deoxyuridine-5'-monophosphate (dUMP) to 2'-deoxythymidine-5'-monophosphate (dTMP). Without ThyA expression, the cells become auxotrophic to thymidine, and hence are unable to grow in the absence of this compound. Cells were transformed either with a plasmid (pCasII) expressing Cas12a together with a CRISPR array containing spacers 1 and 3 ("SP1" and "SP3"), or with similar plasmids that either contained the wild-type T4 ligase (pCall+T4) or the T4 ligase where the internal terminator-like sequence was removed (pCasII+T4 Δ term) (Fig. 3A). Cas12a targets protospacer 1 and 3 and generates compatible sticky ends, 5'CTCCA (top strand) and 5'TGGAG (bottom strand), respectively. If both sticky ends are hybridized and ligated, the reading frame of *thyA* is restored, enabling growth in the absence of thymidine (Fig. 3A). Cells containing pCasII were grown for three days in medium selecting for pCasII and supplemented with thymidine. Each day cells were re-inoculated in fresh medium and plated on agar selecting for pCasII, with and without thymidine. Based on colony forming unit (CFU) counts of each plate in two independent experiments, the fraction of edited cells was calculated (Fig. 3B). In addition, colonies grown on plates without thymidine were used in a colony PCR and analyzed by Sanger sequencing to confirm editing (Figure. S1). On day one, approximately 1 out of 5,000 cells were correctly edited by cut & paste in the pCasII+T4 Δ term transformed cells, whereas only 1 in a million showed correct editing in the pCasII or pCasII+T4 cells. WT T4 (pCasII+T4) had similar editing compared to cells without T4 ligase (pCasII), strongly suggesting that WT T4 ligase is not functional in the cell. The similar editing efficiencies of pCasII and pCasII+T4 suggest that a host ligase (probably LigA) might be responsible for this phenomenon. Expression of T4 Δ term (pCasII+T4 Δ term) increased editing by ~200 fold, demonstrating that the internal terminator like sequence was indeed limiting its expression and more importantly facilitated cut & paste genome editing (Fig. 3B). Editing efficiency increased over the course of the experiment in the pCasII+T4 Δ term transformed cells, reaching to 1 out of 160 cells (0.6%) on day 3 (Fig. 3C), whereas the fraction of correctly edited cells for pCasII and pCasII+T4 remained the same. To investigate Cas12a escape mutants found during cut & paste, i.e. colonies grown on plates without thymidine, the pCasII plasmids in a few

colonies were sequenced and mutations were found either in the CRISPR array or the *cas12a* gene. The CRISPR array was mutated to contain only one repeat, instead of a repeat-spacer-repeat CRISPR array. As for *cas12a*, deletion of the RBS was found in most cases, and in one instance a transposon appeared to be integrated within the coding sequence. Mutations in the cut & paste systems most likely led to a higher survival rate of wildtype cells (in the presence of thymidine) and reduced the overall editing efficiency. Even though the highest fraction of genome editing was 0.6%, this study still demonstrates that Cas12a can be used in combination with T4 ligase to generate specific genomic deletion in *E. coli*. Improving the editing efficiency should be addressed in follow up studies.

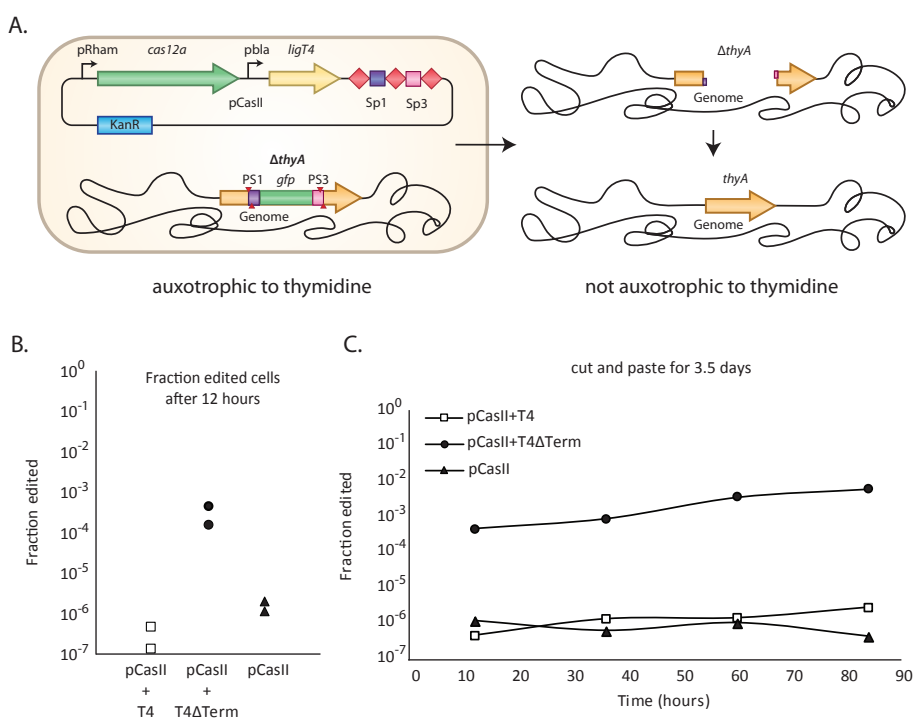


Figure 3 | Gene deletion using cut & paste. (A) Schematic showing how cut & paste can be used to generate a precise genomic deletion in *E. coli*. pCasII expresses Cas12a under control of a rhamnose inducible promoter, T4 ligase and the CRISPR array under a single constitutive promoter (Pbla). The CRISPR array contains spacer 1 and spacer 3. Within *E. coli*'s *thyA* gene is a genomic insertion of *gfp* which knocks out *thyA* (Δ thyA), causing thymidine auxotrophy. Cas12a targets protospacer 1 (purple) and 3 (pink) and generates compatible sticky ends. If sticky ends are hybridized and ligated, *thyA* is restored. (B) Fraction edited cells after 12 hours for cells containing a wild type T4 ligase (WT T4), a modified T4 ligase without an internal terminator-like sequence (T4 Δ Term) and without T4 ligase (Δ T4). Data was collected using two independent experiments. Y-axis represents fraction edited cells (the CFUs of the plates with thymine divided by the CFUs of the plates without thymine). (C) Fraction edited cells by cut & paste in prolonged incubation ($n=1$) of one independent experiment of B. Cells were incubated in a total of 84 hours and plates at time points, 12, 36, 60 and 84 hours. Y-axis represents fraction edited cells and x-axis represents time in hours.

Discussion

In this work, we tested a novel genome editing approach in *E. coli* termed “cut & paste”, which utilizes the staggered ends generated by Cas12a cleavage and ligates them together with T4 ligase *in vivo*. Using cut & paste, precise genomic deletion in *E. coli* was achieved. However, editing efficiency remains low for deletions. The starting editing efficiency was around 1 out of 2000–7000 cells (0.14–0.05%) in 12 hours and reaches up to 1 out of 160 cells (0.62%) when incubated for 84 hours. The highest editing efficiency noted for the cut & paste tool is in the same order of that reported for an overnight (16hr) lambda red recombination system without a selection marker, where editing was reported in 1 out of 90–260 (0.4–1.1%) tested cells/colonies (198). This low editing efficiency is caused by the occurrence of escape mutants and/or a low cut & paste efficiency. Mutation in either the CRISPR array or Cas12a nuclease, will remove the counterselection of Cas12a and will allow, in the presence of thymidine, wild type cells with a split *thyA* gene to survive within the population. Equally important is the cut & paste efficiency, which is dictated by several factors, as discussed below.

The recombination through the here described cut & paste approach relies on several factors, such as cleavage efficiency by Cas12a, sticky ends annealing and gap ligation efficiency. Cleavage efficiency is crucial for cut & paste genome editing as it generates sticky ends and counterselects wild type sequences. Apart from appropriate functional expression of the Cas12a nuclease, cleavage efficiency heavily relies on saturation with functional crRNA guides. Such as the importance of the design of the CRISPR array, especially of the spacer sequences used, as exemplified in this study and earlier work (46, 210). More specifically, it was found that folding of the crRNA greatly influences cleavage efficiency of Cas12a. This can be resolved by changing the spacer order in the CRISPR array or by modifying the non-base pairing region of the spacer (21–23) to enforce a more favorable crRNA structure (53, 210, 211). Therefore, crRNA structure should be taken into consideration by using an RNA structure prediction tool when designing spacers for genome editing. Unlike Cas9, Cas12a does not consistently cleave at the same position of the protospacer, the cleavage position deviates with 1 nt, causing sticky ends to vary between 4–6 nt in length, with the majority being 5nt long (79). Incompatible sticky ends can lead to improper sticky end annealing, which will reduce the editing efficiency. Although cleavage efficiency is certainly an important aspect for our tool, our results indicate that in this setup the subsequent ligation reaction was the major bottleneck, as removal of the terminator-like sequence within T4 ligase substantially increased the number of edited cells. Increasing ligase activity even further could therefore be beneficial to enhance the efficiency of our tool even further. However, cloning for a stronger constitutive promoter for T4 ligase (Ptac) was unsuccessful and resulted in mutations in the promoter region (not shown), indicating toxicity for high T4 ligase expression. Possible explanation can be that the overexpression of T4 ligase may have a similar effect as LigB, which blocks DNA replication and reduces cell viability

if overexpressed (212). Next, T4 ligation of double stranded breaks produced by Cas9 cleavage led to large deletions (>6.9 kbp) in the genome (213). This is most likely caused by RecBCD linear DNA degradation in combination with Cas9 cleavage after perfect re-ligation of the double stranded break. Within the cut & paste system, sticky ends can also be degraded by RecBCD, which can further lower the editing efficiency. However, deletion of RecBCD or addition of the lambda Gam protein was able to reduce the deletion size (213). Another way to increase cut & paste editing efficiency is by inducing cell dormancy during the editing process to halt replication and buy time for *in vivo* digestion and ligation to take place. Halting replication also delays sensing of DNA damage by RecA-LexA, which delays induction of apoptotic-like death (ALD) pathways, an extreme SOS response (214, 215).

To improve the editing efficiency of the cut & paste systems, a Cas12a-T4 ligase fusion protein can be used instead of two separate proteins. A Cas12a-T4ligase fusion will allow T4 to be constantly in proximity of the cleaved DNA for quick repairs. Also fine tuning the expression of T4 ligase can be done by testing different promoter strength, e.g. the Anderson promoter library (216). Another is to reduce the spacer length to 17-19 nt, since spacer length < 20 nt generates more consistent staggered end length of 8 nt long (217).

Currently, the lambda red recombination system in combination with Cas nuclease-mediated counterselection has been shown to be the most efficient in genome editing tool for *E. coli*. However, the efficiency is reduced when the length of the to-be-inserted fragment is more than 1 kb (218). This limitation might be caused by the activity of lambda exonuclease (exo) degrading dsDNA (197). Long dsDNA repair templates for longer inserts, require more time to be degraded by lambda red, meaning less ssDNA repair templates available. For cut & paste, however, it is hypothesized that editing the efficiency would not decrease with increasing insertion size. This is because Cas12a cleavage and T4 ligation efficiency are not influenced by repair template size.

All in all, this work has shown a proof of concept of a novel recombination approach, cut & paste, as a genome editing tool for *E. coli*, at least suitable for generating precise deletions. Admittedly, at present the efficiency of cut & paste is lower than the most used current technology in which CRISPR-Cas is combined with lambda red. Still, by further optimizing, the cut & paste approach has the potential to become a new addition to the genome editing toolbox.

Acknowledgments

The authors would like to thank Tess Hogeboom and Prarthana Mohanraju for their technical assistance. J.v.d.O is supported by the Dutch Research Council (NWO) through a TOP grant (714.015.001).

Author contributions

W.Y.W, S.C.A.C and J.v.d.O conceived this study and the experimental design. J.B. conducted the experimental work. W.Y.W, S.C.A.C and J.v.d.O supervised this project. W.Y.W., R.H.J.S. and J.v.d.O wrote the manuscript.

Competing interest

No potential conflict of interest is reported by the authors

Materials and Methods

Bacterial strains and growth conditions

For plasmid cloning, *E. coli* strains DH5- α and DH10- β were used. For testing genomic deletions, the *E. coli* Δ ThyA strain was used containing a *gfp* sequence flanked by PS1 and PS3 in the *thyA* gene. Δ ThyA was created with λ -red recombination using a PCR fragment as a template. Cells were grown at 37°C in Luria Bertani (LB) liquid medium (10 g/L peptone, 10 g/L NaCl and 5 g/L yeast extract) at 220 rpm. Ampicillin (100 μ g/mL), kanamycin (50 μ g/mL), gentamycin (30 μ g/mL), chloramphenicol (35 μ g/mL), rhamnose (2 g/L) and thymidine (100 μ g/mL in liquid, 20 μ g/mL in plates) were added where required.

Plasmid construction

The plasmid insertion by cut and paste consists of a three plasmids system: pCas, pAcceptor and pDonor. The three plasmids are resistant to kanamycin, ampicillin and gentamycin, respectively. Moreover, all three plasmids have compatible origin of replications consisting of pBR322, p15A and pBBR1, respectively.

Construction of pCas starts with pRham-Cas12, which was constructed using ligation independent cloning (LIC). T4 ligase was then inserted into pRham-Cas12a by digestion (BamHI and Sall) and ligation with a digested PCR amplified T4 ligase (BamHI and Sall), to create pCas_no_array. pCas was then constructed by digesting pCas_no_array with NotI and SpeI and ligated with a digested CRISPR array containing Spacer 1 and Spacer 2. The CRISPR array was obtained by digestion of pMA-RQ_Cas12a_array_Sp1_Sp2 with NotI and SpeI. pMA-RQ_Cas12a_array_Sp1_Sp2 is an entry vector for Cas12a spacer cloning.

pCas2_no_array was constructed by digestion of pCas_no_array with BamHI and a bla promoter was ligated in. The bla promoter was created by ligating two oligo's together to create an adapter. To add the CRISPR array containing spacer 4, spacer 2 was removed from pMA-RQ_Cas12a_array_Sp1_Sp2 by digestion of NcoI and NheI. Spacer 4 was created by annealing two oligo's together and then ligated into the digested vector to construct pMA-RQ_Cas12a_array_Sp1_Sp4. pCas2 was then constructed by digestion and ligation of pCas2_no_array and pMA-RQ_Cas12a_array_Sp1_Sp4 using (NotI and SpeI). A frameshift was introduced into T4 ligase to construct pCas2ΔT4_no_array and pCas2ΔT4. The frameshift was introduced by a SacI digestion of pCas2_no_array and pCas2. Sticky ends of the digested fragments were then filled in by a Klenow reaction and ligated together by blunt end ligation. pCas2_T4Δterm_no_array and pCas2_T4Δterm were constructed by a three-point ligation. pCas2_no_array and pCas2 were digested with SpeI and AflII. Two fragments were amplified by PCR, of which one contained an SpeI site upstream and the other contained a AflII site downstream. A three-point ligation was done using SpeI, blunt and AflII sites.

pAcceptor was constructed by a three-point ligation using pWUR873 vector, digested with KpnI and SpeI. Both *cat* and *lacZα* were PCR amplified to contain a KpnI site upstream and SpeI site downstream of the gene, respectively. Then a three-point ligation was by using KpnI, blunt and SpeI.

pDonor was constructed by digestion and ligation of pSEVA631 and a PCR amplified '*cat*'. pSEVA631 was digested with AvrII and NotI whereas '*cat*' was digested with SpeI and NotI. Both fragments were ligated by a NotI site and the compatible overhang of SpeI and AvrII.

pTarget 3 – pTarget 8 used for testing cleavage efficiency of protospacers 3-8 were constructed by digestion of PSC033 with Scal and ligated with a PCR amplified kanamycin resistance gene containing a protospacer upstream.

Generating DH10-β ΔThyA strain

A DH10-β harboring pSC020, was transformed with a *gfp* fragment amplified using TH004 as template. pSC020 contains lambda red under an arabinose inducible promoter and was induced with 50nM L-arabinose during preparation of competent cells. In the recovery phase of the transformation, LB supplemented with thymidine was used. Cells were incubated for 2-3 hours at 30 °C at 750 rpm. Then 250 μl was plated on LB agar containing thymidine and trimethoprim (5mg/L). Colonies were confirmed by colony PCR and sanger sequencing.

Plasmid loss assay

DH10-β harboring pCas, pAcceptor and pDonor were grown overnight in medium selecting for pCas (kanamycin) and pDonor (gentamycin) or pCas (kanamycin) and pAcceptor (ampicillin) to investigate plasmid loss for pAcceptor or pDonor, respectively. Cells were then plated on plates containing all three antibiotics (kanamycin, gentamycin and ampicillin) and also kanamycin and gentamycin or kanamycin and ampicillin for plasmid loss of pAcceptor or pDonor, respectively.

Cut and paste insertion (triple plasmid system)

DH10-β cells harboring pCas, pAcceptor and pDonor were grown overnight in medium selecting for pCas (kanamycin) and pAcceptor (ampicillin). In addition, cells were also grown in medium with and without 0.2 g/L L-rhamnose. Cells were then inoculated in medium containing chloramphenicol (1:100) to select for correctly modified pAcceptor plasmids. For a prolonged experiment, apart from inoculating in medium containing chloramphenicol, cells were inoculated in fresh medium containing kanamycin, ampicillin and +/- rhamnose.

Cut and paste genomic deletion

DH10-β ΔThyA was transformed using the following plasmids: pCas2, pCas2_ΔT4 and pCas2_T4Δterm. Transformants were inoculated in 10 ml LB (1:100) containing kanamycin, rhamnose and thymidine and grown overnight. 1 mL of cells were sampled, centrifuged for 5 min at 3000 g and resuspended in LB medium to remove residual thymidine. Resuspended cells were then plated on agar containing kanamycin and +/- thymidine. When plating with thymidine, 50 μL of 10E6 diluted cells were used. When plating without thymidine 50 μL of a serial dilution (10-3000) was used. Cells growing on plates without thymidine were confirmed by colony PCR and sanger sequencing.

Calculating fraction edited cells

$$\text{Fraction edited cells} = \frac{\frac{\text{CFU}}{\text{mL}} \text{ no thymidine plate}}{\frac{\text{CFU}}{\text{mL}} \text{ + thymidine plate}}$$

***in vitro* cleavage assay**

Cas12a proteins were expressed and purified according to Mohanraju et al., 2018 (219). crRNA was generated by in vitro transcription (IVT) using a dsDNA template, obtained from either PCR amplification or annealing two oligo's together. The IVT reaction consisted of template (25ng/μL), T7 RNA polymerase (10 U/μL), NEB 5x reaction buffer T7 RNA polymerase and rNTP's (1mM each). Reaction was incubated at 37°C for 2.5 hours. 2xRNA loading dye fortified with 500nM EDTA was added to the sample, and sample was loaded on at 5% acrylamide gel. RNA band corresponding the size of the transcript was cut out of gel and incubated overnight in buffer (50mM Tris, 1nM EDTA, 10 mM DTT) at 37°C at 900 rpm. Amicon Ultra 0.5 ml 10K centrifuge were then used to purify the RNA according to manufacturer's protocol.

dsDNA targets were generated by PCR amplifying pTarget 3 – pTarget 8. Cas12a (60nM) and crRNA (120nM) were pre-incubated for 0.5 hour at 37°C then linear dsDNA targets (3nM) were added to a final volume of 100 μL. At time point 0, 10, 20 and 50 minutes, 20 μL were taken and added to 5 μL purple loading dye (NEB).

Supplementary Figures and Tables

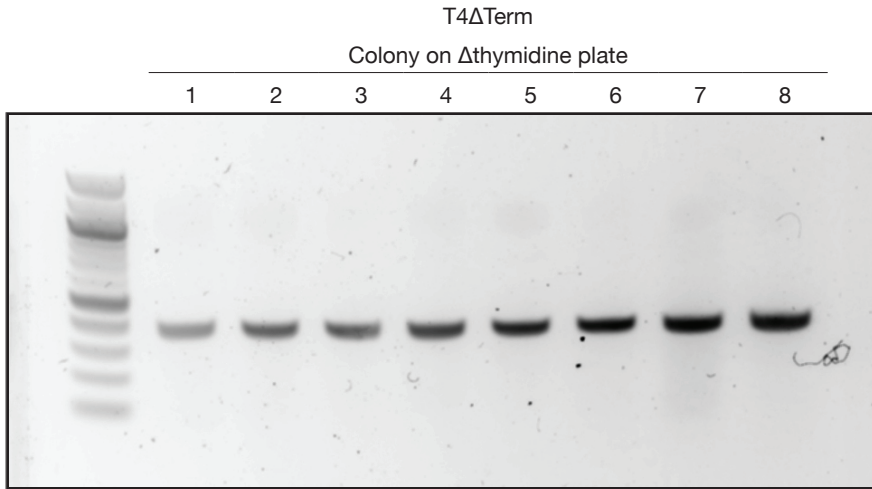


Figure S1 | Colony PCR of T4 Δ term on Δ thymidine LB plate using 12hr samples. 100bp ladder (NEB) was used as a marker. Correct deletion would lead to a band of 377bp and the wild type of 1148bp.

Table S1 | List of primers used in this study.

oligo ID	sequence (5'-3')
pAcceptor	
BG8291	TACTGGTACCCCGCTTCGCGGGGTTTTTCAAGTTTACACTTTATGCTTCCGGCTCGTATAATTGAGTTATCGAGATTTTCAGGAGC
BG8292	TTGTTCGATGGGAAACCTTACCCTCCAGAGCGATGAAAACGTTTC
BG8293	GCCGTAGATAAACAGGCTGG
BG8294	TACTACTAGTCTGGAGGTAAGGTTCCCATCGACAATTAGCGCCATTTCGCCATTCAGG
pDonor	
BG8295	TCATAC TAGTGGTACCTTGGACTTACCAATGAGCACGTGGAGTGAATACCACGACGATTTC
BG8296	TACTGCGGCCGCTTGGACTTACCAATGAGCACGTGGAGTTTACGCCCGCCCTGCCA
pCas	
BG7802	ACTCCAACCTCCATAAGGATCCTAGAGCGGCCGCCAC
BG7803	ACTTATATCTCCTTCTTAAAGTTAAACAAAATTATTCTAGAGG
BG7709	TTTAAGAAGGAGATATAAGTCATGTCAATTATCAAGAATTTGTTAATAAATATAG
BG7710	TTATGGAGTTGGAGTCTTATTATTAGTTATTCTATTCTGCACG
BG8101	GGATCCTTTGTTAACTTTAAGAAGGAGATATAAGTATGATTCCTTAAATTCCTGAACG
BG8102	GTCGACTCATAGACCAGTTACCTCA
pCas2	
BG8832	GATCCTTTACACTTTATGCTTCCGGCTCGTATAAT
BG8833	GATCATTGTCTCATGAGCGGATACATATTTGAAG
BG8728	CATGGGTCTAAGAACTTTAAATAATTTCTACTGTTGTAGATACACACTGCAATTCAGGTTGGAGTG
BG8729	CTAGCACTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTATTTAAAGTTCTTAGACC
BG8998	CATGGCGAATGTATCAAAGCAGC
BG8999	CATACTTATATCTCCTTCTTAAAGTTAAACAAAAGGATCATTATACG
BG9000	CATCACCATCACCATCACATTCTTAAATTCCTGAACGAAATAGCATCTATTGG
BG9006	CTTCCTTAAAGAAGTCTAATCTTAGATATTCATTACCAGCTCGTGATAACAGGCGCACATCATCTAATTCATCACCTCTAACTTCAG
BG8906	AAAACGCGCACCTCGGGCCAGTGTATGGTAAACAGTGGCGCGCCTGGCCAACTCCAACCTGAATTGCAGTGTG
BG8907	GTTTTTCAGCTGGTTAGTACCGTAGTGATCTGGTCAATATGACGACCATCTGGAGGGTAAGGTTTCCCATC
ΔThyA deletion colony PCR	
BG4794	ATGCGTCGACTATCCGGGTCGTTTTTCAGCTGG
BG6627	GATAACATATGAAACAGTATTAGAACTGATG
in vitro transcription	
BG8625	AAGTAATACGACTCACTATAGGGTGTGGCTGATTAGGCAAAAACG
BG8626	CGAAGCGGGGAGACAG
BG8665	ACTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG8667	ACTCCACAATGATCTCGTAGGCGTATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG8668	ACTCCAGCTAGTGTTACGGGAGCAATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG8669	ACTCCAGTAAGCGATTAGACTGGATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG8670	ACTCCAAGCTCCGGTGATATAGTATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG8671	ACTCCATTGGGACCGGTAATTGTGATCTACAACAGTAGAAATTCATAGTGAGTCGTATTACTT
BG4925	AAGTAATACGACTCACTATA
protospacer targets 3-8	
BG8655	ACTTGTTCGATGGGAAACCTTACCCTCCAGGATAAAGCGGGCCATGTTAAGG
BG8657	ACTTGACACACTGCAATTCAAGTTGGAGTAGGAGCTATGAGCCATATTCAACG
BG8659	ACTTGAGCCCTACGAGATCATTTGTGGAGTAGGAGCTATGAGCCATATTCAACG
BG8660	ACTTGTGCTCCCGTAACACTAGCTGGAGTAGGAGCTATGAGCCATATTCAACG
BG8661	ACTTGCCAGTCTAAATCGCTTACTGGAGTAGGAGCTATGAGCCATATTCAACG
BG8662	ACTTGACTATATGCACCGGAGCTTGGAGTAGGAGCTATGAGCCATATTCAACG
BG8663	ACTTGACACAATTACCGGTCCCAATGGAGTAGGAGCTATGAGCCATATTCAACG
BG5393	TATACATATGTCAAAGAGACGTCTTTTGTAAAGAATG

discription
Fw cat KpnI - terminator- pLacUV4
Rv cat protospacer 1
Fw lacza
Rv lacza SpeI protospacer 1
Fw cat SpeI KpnI protospacer 2
Rv cat NotI protospacer 2
Fw pRham LIC cloning
Rv pRham LIC cloning
Fw Cas12a LIC cloning
Rv Cas12a LIC cloning
Fw T4 BamHI RBS
Rv T4 SalI
Fw BamHI pbla adapter top
Rv BamHI pbla adapter bottom
FW repeat Spacer 2 adapter top
Rv repeat Spacer 2 adapter bottom
Fw Cas12a
Rv T4 ligase front
Fw T4 lihase 6x his
Rv T4 Ligase dTerm
Fw GFP ThyA homologous arm
Rv GFP ThyA homologous arm
Fw ThyA
Rv ThyA
Fw T7 CRISPR array pMA-RQ_Cas12a_array_Sp1_Sp2
Rv CRISPR array pMA-RQ_Cas12a_array_Sp1_Sp2
Rv PS3 IVT Template
Rv PS4 IVT Template
Rv PS5 IVT Template
Rv PS6 IVT Template
Rv PS7 IVT Template
Rv PS8 IVT Template
Fw PT7
Rv KanR (used for plasmid and linear target construction)
Fw KanR PS3
Fw KanR PS4
Fw KanR PS5
Fw KanR PS6
Fw KanR PS7
Fw KanR PS8
Fw Eco147I NdeI (for linear target construction)

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
G	T	G	T	C	A	T	A
T	G	G	C	G	G	G	C
C	A	A	G	A	C	C	T

A T A A T T T C
T C H A P T E R
G C C C A G C T
C G A A G G C 6
A A A A C T T T

Characterizing a compact CRISPR-Cas12u1 enzyme

Wen Y. Wu*, Prarthana Mohanraju*, Sjoerd C. A. Creutzburg, Karlijn Keessen, Tahseen S. Khan, Stijn Prinsen, Winston X. Yan, Chunyu Liao, Kira S. Makarova, David A. Scott, Chase L. Beisel, Charlie Laffeber, Joyce H.G. Lebbink, Eugene V. Koonin & John van der Oost†

Manuscript in preparation

*These authors contributed equally to this work.

†To whom correspondence should be addressed: J.V.D.O. (john.vanderoost@wur.nl)

Abstract

CRISPR-Cas systems are prokaryotic heritable adaptive immune systems that have been repurposed as powerful genome editing tools in a wide range of organisms. These tools use RNA-guided Cas nucleases (757 to 1400 amino acids) for their specific DNA- or RNA-targeting activities. Here we present a novel Cas protein from the CRISPR-Cas type V-U1 system from *Mycolicibacterium mucogenicum* CCH10 (596 amino acids). Despite its small size, MmuCas12u1 seems to be able to process its own pre-crRNA. In addition, MmuCas12u1 is capable of targeting and binding of double-stranded DNA (dsDNA). Akin to most type V proteins, MmuCas12u1 recognizes a 5'-TTN-3' PAM on the non-target strand of a dsDNA. Unexpectedly, MmuCas12u1 enzyme does not cleave dsDNA, and analysis in *E. coli* indicates a crRNA-guided MmuCas12u1-mediated transcriptional silencing. By leveraging this property, MmuCas12u1 has been used for single- and multiplex- transcriptional silencing in *E. coli*. Finally, *in vivo* experiments suggest that the RuvC-dependent ribonuclease activity of MmuCas12u1 enhances the silencing effect.

Main text

The everlasting biological arms-race between bacteria and archaea and viruses has resulted in the evolution of remarkably diverse CRISPR-Cas defense systems in these prokaryotes against their invaders (19, 220, 221). The key players of the CRISPR-Cas systems are the Cas proteins that catalyze crRNA-guided interference of DNA or RNA targets (15). Based on the unique Cas effector complexes, CRISPR-Cas systems are currently grouped into two classes that are each subdivided into three types. Class 1 systems use multi-protein effector complexes to achieve target recognition and interference, while class 2 systems use a single protein with multiple functional domains for target recognition and interference (222, 223). The facile programmability and the successful heterologous expression of class 2 CRISPR-Cas nucleases has allowed for their repurposing for genome editing, transcriptional regulation, and diagnostics (224, 225).

Class 2 includes types II, V and VI, represented by the signature nucleases Cas9, Cas12 and Cas13, respectively. Cas9 cleaves double-stranded (ds) DNA using its HNH and RuvC nuclease domains, while the first characterized Cas12 variants (subtypes V-A and V-B) have been demonstrated to cleave dsDNA specifically and single-stranded (ss) DNA non-specifically using a single RuvC domain (43, 226-228). Both Cas9 and Cas12a cleave dsDNA adjacent to a short sequence, termed the Protospacer Adjacent Motif (PAM) (43, 227). Cas13 is the only known Cas nuclease to exclusively cleave RNA using two HEPN ribonuclease domains (229, 230). Although these nucleases have been widely used for genome engineering, the large size of Cas9, Cas12a and Cas13 (900-1630 amino acids) places constraints on some cellular delivery approaches that may limit certain applications including therapeutics (231, 232). By screening rapidly growing genomic and metagenomic databases, partly as a quest for potential novel genome editing tools, eight new functionally different Cas12-like systems have recently been identified and characterized: type V-C to V-J (Cas12c-j) (67, 69, 233-236). Some of these effectors are nearly half the size of the smallest Cas9 or Cas12a proteins potentially making them highly appealing for packaging in FDA approved safe-to-use Adeno-Associated Viruses (AAVs) for *in vivo* genome engineering applications and therapeutics (231, 232). Thus, discovery and unravelling the mechanism of novel and compact CRISPR-Cas systems is not only interesting for fundamental reasons, but also holds great potential for new and improved technological advancements.

Using CRISPR arrays as the search seed in the computational class 2 discovery pipeline (223) yielded several variants of type V loci, tentatively called uncharacterized (U) subtypes V-U1, -U2, U3, U-4 and -U5 (223, 237). These type V-U proteins show highly significant similarity to the TnpB-like proteins and appear to have evolved independently from distinct TnpB families (223) (Fig. 1A). The resemblance of type V-U1 proteins to type V nucleases suggests that they may have existed as an ancestral class 2 CRISPR system. They most likely evolved from a distinct, “domesticated”

TnpB-like transposase that gained domains over time, resulting in Cas12 variants with different features, eventually leading to the large type V nucleases like Cas12a (223). With sizes between 500 and 700 amino acids, the putative effector proteins of the type V-U loci are much smaller than the archetypal class 2 effectors, but larger than the transposon-encoded TnpB proteins (Fig. 1B). Despite the occurrence of the characteristic bacterial RuvC-like domains found in the type V-U1 proteins, their small size and the absence of other cas genes near the CRISPR array suggested it is unlikely for these systems to function as stand-alone CRISPR effectors (Fig. S1, S2) (223). Nonetheless, at least some of them were predicted to be active based on their respective CRISPR arrays which contain spacers homologous to phage genome sequences (223). Recently, the type V-U5 effector, Cas12k (formerly, C2c5), containing a naturally inactivated RuvC-like nuclease domain was shown to be hijacked by Tn7-like transposons to allow for directed DNA transposition via crRNA-guided targeting (69). However, the functionality of the other four subtype V-U systems remains to be uncovered.

Of the five V-U variants, subtype V-U1 is the most prevalent in different bacteria, whereas the remaining subtypes are largely limited in their spread to particular bacterial taxa (223). The evolutionary stability in terms of sequence conservation and consistent association with CRISPR arrays with diverse spacers (223), led us to hypothesize that these type V-U1 loci encode biologically functional enzymes with nucleic acid targeting activity despite their small size. To test the hypothesis, we studied the type V-U1 CRISPR-associated nuclease, Cas12u1 from *Mycolicibacterium mucogenicum* CCH10-A2 (MmuCas12u1) (Fig. 1A). MmuCas12u1 contains a RuvC-like nuclease domain near the C-terminal end, with an organization reminiscent to that found in other type V nucleases (Fig. 1B).

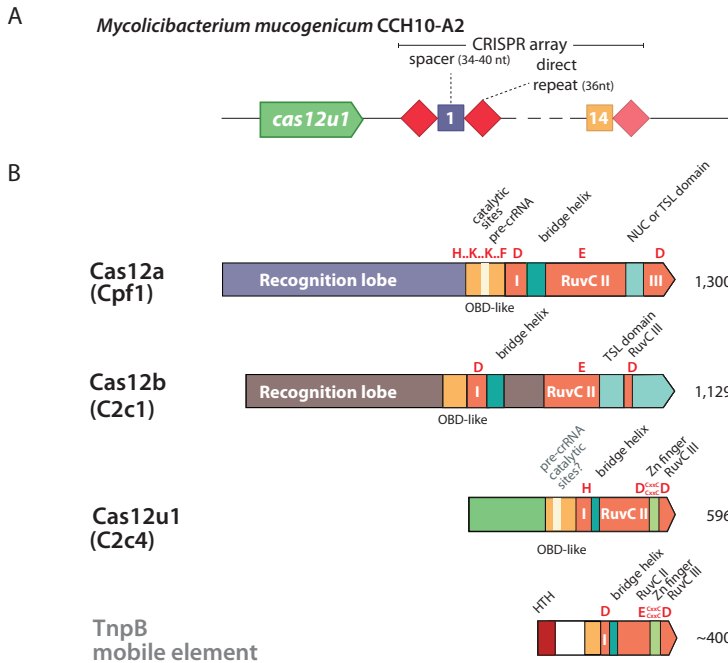


Figure 1 | Type V-U1 CRISPR-Cas system in *Mycobacterium mucogenicum* CCH10-A2
(A) Organization of the CRISPR-Cas locus on the genome of *Mycobacterium mucogenicum* CCH10-A2. Red diamonds are perfect direct repeats; the lighter red diamond at the right side of the CRISPR array indicates a slightly degenerated repeat, generally indicative of the 3' end of the transcribed precursor-crRNA **(B)** Domain architectures of Cas12a, Cas12b, Cas12u1 and TnpB proteins are compared. Protein lengths are drawn to scale. Amino acid lengths are based on *Francisella novicida* Cas12a, *Alicyclobacillus acidiphilus* Cas12b and *Mycobacterium mucogenicum* Cas12u1.

As some of the characterized type V variants use a second RNA (tracrRNA, scoutRNA) in addition to its crRNA, we initially performed an *in-silico* analysis for the presence or absence of a tracrRNA-like sequence in the MmuCas12u1 loci using a previously described prediction approach (238). Using this approach, no tracrRNA-like sequences have been detected in the adjacent DNA sequences. This is in line with the fact that the Cas12u1 CRISPR arrays have partial palindromic sequences (Fig. S3), a feature that appears to correlate with tracrRNA-independent guide processing systems.

To functionally characterize the type V-U1 protein, we transformed *Escherichia coli* cells with a plasmid containing the (*E.coli*) codon-harmonized *mmucas12u1* gene and a minimal CRISPR array (repeat-spacer-repeat), with a spacer targeting the lac promoter (Plac). After purification of the MmuCas12u1 protein to homogeneity (Fig. S4A), subsequent analysis revealed the presence of co-purified RNAs, that are presumably around the size of the mature crRNAs (Fig. S4B). This strongly suggests that MmuCas12u1 associates with a crRNA (see below, determination

of PAM sequence). However, sequencing of the small RNAs must be performed to corroborate this finding and to map the exact cleavage sites on the crRNA (Fig. S4B). The absence of a predicted tracrRNA and the size of the co-purified RNAs, suggested a potential for crRNA biogenesis by the effector protein itself, as has been reported for type V-A and V-H/I/J (167, 230, 233, 239). Therefore, we performed an *in vitro* pre-crRNA processing assay using purified recombinant MmuCas12u1 protein and a minimal pre-crRNA (repeat-spacer-repeat-spacer-repeat). Processing of the pre-crRNA to intermediates and seemingly mature guides was observed (Fig. S4C). pre-crRNA processing in an *in vitro* transcription and translation system (TXTL) followed by Northern blot analysis showed that MmuCas12u1 performs autonomous pre-crRNA processing, and that this activity is independent of the presence of an active RuvC-like domain (Fig S5) (240).

The PAM sequence plays a central role in self/non-self target selection in dsDNA cleaving CRISPR systems. In the absence of a PAM, the Cas nucleases cannot stably bind a potential target, even if it is perfectly complementary to the spacer (241). To test whether MmuCas12u1 requires a PAM and can conduct crRNA-guided dsDNA interference, we adapted the previously developed PAM-SCANR assay (242), a high-throughput *E. coli*-based positive and tunable screen for assessment of PAM specificity (Fig. 2A). It is based on a catalytically inactive crRNA-guided Cas effector blocking the -35 element within the promoter upstream of *lacI*. In the absence of binding (due to a non-functional PAM) by the inactive Cas effector, the expressed LacI repressor blocks the *lac* operator in the promoter of the green fluorescent protein (GFP) gene. In the case of binding of the inactive Cas nuclease (due to a functional PAM), *lacI* expression will be inhibited, hence resulting in expression of GFP (242). We generated an effector plasmid (pCas-MmuCas12u1) encoding a catalytically inactive *mmuCas12u1* gene [single mutant of one of the RuvC-II active site residues (D485A)], a CRISPR array plasmid (pCRISPR-PS), with a spacer targeting a 5'-NNNN-3' PAM library placed upstream of the -35 element of the promoter of *lacI* in the target PAM-SCANR plasmid (pTarget-PS). A CRISPR array plasmid with a non-targeting spacer (CRISPR-NT) was used as a negative control. Gene repression of the LacI repressor by crRNA-guided MmuCas12u1 binding of the dsDNA containing a functional PAM would lead to the expression of the GFP reporter. *E. coli* cells were transformed with the pCas-MmuCas12u1, either pCRISPR-PS or -NT and pTarget-PS plasmids, and after cultivation, GFP fluorescent cells were isolated through fluorescence-activated cell sorting (FACS) (Fig. 2B). Comprehensive screening based on next-generation sequencing of the pre-sorted and post-sorted PAM libraries and analyses of the target-flanking sequences revealed that the binding of target dsDNA by MmuCas12u1 depends on a 5'-NTTM-3' PAM (Fig. 2C). Weak functional PAMs were also detected by titrating the Isopropyl β -D-1-thiogalactopyranoside (IPTG) levels to downregulate the strength of LacI repression (Fig. S6A). The presence of a T nucleotide at the -2 position of the 5'-PAM appears most crucial. Thus, the PAM recognized by MmuCas12u1 is similar to that of the other characterized type V effector proteins (43, 213, 233, 243). To validate the PAM and to clarify the ambiguity at the -1 and -4 PAM positions, we generated a set of 16 different plasmids (pTarget-

GFP) containing a protospacer adjacent to a 5'-NTTN-3' PAM sequence on the promoter upstream of the *gfp* target gene (Fig. S6B). *E. coli* cells harboring the CRISPR array plasmid with a spacer targeting the promoter (pCRISPR-promoter) and pCas-MmudCas12u1 were transformed with the pTarget-GFP plasmids and assessed for silencing of GFP fluorescence, as a result of efficient dsDNA binding (Fig. S6B). As a control, we also analyzed the catalytically inactive type V-A effector (dCas12a) of *Francisella tularensis* subsp. *novicida* U112 (pCas-FndCas12a), with its corresponding crRNA guide targeting the same protospacer (Fig. S6C). Efficient GFP repression was observed for all the tested PAM variants, confirming the PAM sequence being 5'-(N)TTN-3' for MmudCas12u1 and 5'-TTTV-3' for FndCas12a. In addition, this analysis revealed robust *in vivo* crRNA-guided dsDNA binding by both MmudCas12u1 and FnCas12a (Fig. S6C).

Characteristic to most DNA-targeting class 2 interference complexes is their ability to recognize, bind and cleave both dsDNA and single-stranded DNA (ssDNA) substrates (43, 227, 228, 244, 245). Therefore, to test crRNA-guided dsDNA interference, a target plasmid containing a 5'-CTTA PAM adjacent to the previously used PAM-SCANR protospacer (pTarget-CTTA) was generated. It was transformed into *E. coli* cells harboring the effector plasmid encoding the wild-type MmuCas12u1 protein (pCas-MmuCas12u1) with either the pCRISPR-PS or the control pCRISPR-NT plasmid. Notably, upon transformation with the pTarget-CTTA plasmid, no depletion in the number of transformants was observed for the cells harboring the pCas-MmuCas12u1 and pCRISPR-PS, as compared with the strain harboring the pCas-MmuCas12u1 and the control pCRISPR-NT plasmids (Fig. 2D). In contrast, the dsDNA targeting Cas12a control did result in substantially lower number of transformants (Fig. 2D). This indicates that, at least under the tested conditions, CRISPR-MmuCas12u1 does not cleave dsDNA in the heterologous *E. coli* host (Fig. 2D).

To confirm the inability of MmuCas12u1 to cleave dsDNA, we repeated the same experiment, but with a plasmid (pCRISPR-GFP) containing a different spacer targeting the end of the *gfp* gene. Again, we did not observe any decrease in the number of the transformants as compared to the non-target control. Interestingly, however, we did observe a drop in the GFP fluorescence signal for the cells harboring the pCas-MmuCas12u1, pCRISPR-GFP and pTarget-GFP plasmid (Fig. 2E). This GFP repression activity was much lower or undetectable in the cells harboring the pCas-MmudCas12u1, pCRISPR-GFP and pTarget-GFP plasmids (Fig. 2E). This suggests that the observed silencing of gene expression by MmuCas12u1 is, at least to some extent, RuvC-dependent, possibly through cleavage of the mRNA transcript (Fig. 2E). Although the RuvC domain generally catalyzes cleavage of DNA, the recently characterized Cas12g (type V-G) nuclease mediates *in vitro* cleavage of both RNA and ssDNA (246). In addition, the RuvC domain has similar folds to the PIWI domain found in Argonautes, of which DNA- and RNA-cleaving variants are known (247-249). Thus, based on these observations, we hypothesized that MmuCas12u1 might possess target-activated (specific or non-specific) ssRNA cleavage activity.

To test this possibility, we incubated a purified MmuCas12u1 protein first with a crRNA guide, and then with a complementary dsDNA plasmid target, and eventually with a either a target or a non-target RNA. However, under these *in vitro* conditions, MmuCas12u1 appeared to be incapable to cleave either of the RNAs (Fig. S7).

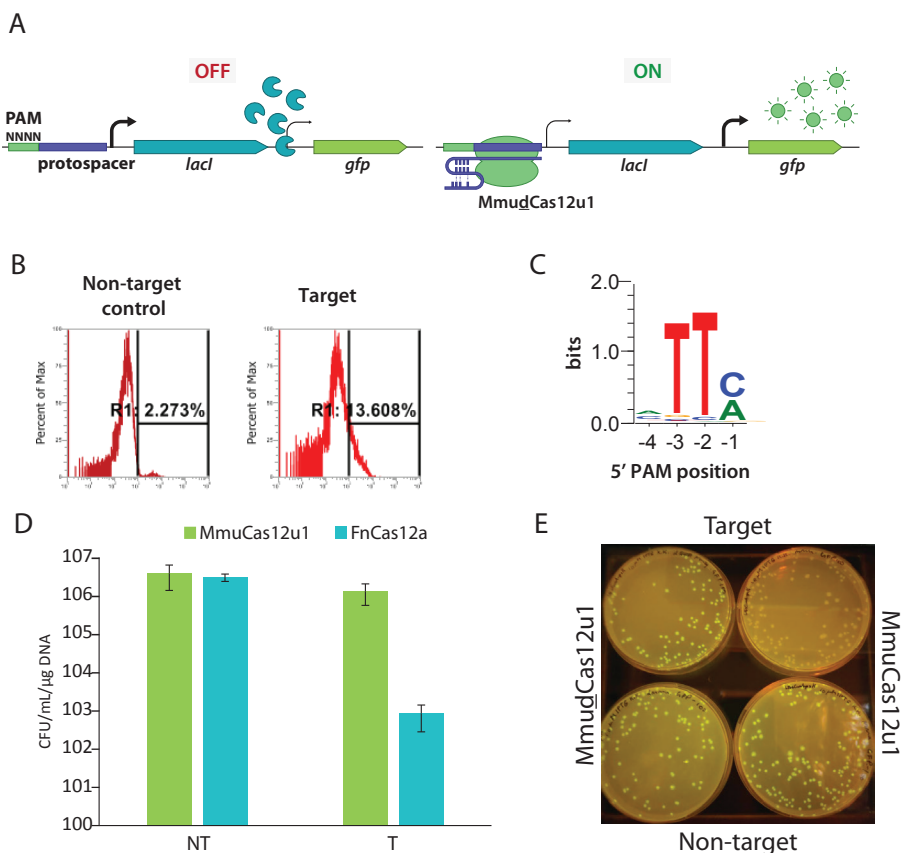


Figure 2 | The *Mycobacterium mucogenicum* CCH10-A2 Cas12u1 protein recognizes dsDNA targets flanked by a 5'-TTM PAM and does not cleave dsDNA. (A) Schematic illustrating the *in vivo* PAM screen achieved by PAM-SCANR. It consists of a library of randomized 5' PAM sequences (4N) cloned upstream of the *lacI* promoter. Immediately downstream of *lacI* is the *LacI*-dependent *lacZ* promoter controlling expression of GFP. A catalytically dead MmuCas12u1 (MmuCas12u1) protein is targeted to a protospacer within the *lacI* promoter, resulting in GFP fluorescence only in the presence of a functional PAM. **(B)** Cells harboring a targeting or non-targeting spacer against the pTarget-PS plasmid that led to a GFP fluorescence were isolated by fluorescence-activated cell sorting (FACS). The Y-axis represents the percentage of 10,000 cells and the X-axis represents GFP fluorescence. **(C)** Plasmids from the FACS-sorted cells were extracted and sequenced to determine functional PAM sequences. Sequence logo for the MmuCas12u1 PAM as determined by NGS sequencing of plasmids from sorted fluorescent cells. **(D)** Results of the *in vivo* dsDNA targeting experiment showing OD₆₀₀ measurements from cultures of *E. coli* harboring the pTarget-PS plasmid transformed with pCas-MmuCas12u1 and pCRISPR-PS compared to cells transformed with pCas-FnCas12a and pCRISPR-Cas12a-PS plasmid. **(E)** Qualitative comparison of GFP fluorescence in the cells harboring pTarget-GFP transformed with pCas-MmuCas12u1 with either pCRISPR-GFP (Target) or pCRISPR-NT (Non-target) versus the cells harboring pTarget-GFP transformed with pCas-MmuCas12u1 with either pCRISPR-GFP (Target) or pCRISPR-NT (Non-target).

To further investigate the dsDNA target-dependent interference activity by MmuCas12u1, we cloned a target plasmid (pTarget-Operon) containing a bi-cistronic operon with two fluorescence reporter genes, *rfp* and *gfp* (Fig. 3A). *E. coli* cells harboring either the pCas-MmuCas12u1 or pCas-MmuCas12u1 and the pTarget-operon were transformed with a set of different CRISPR array plasmids (pCRISPR-A1_F2) containing spacers targeting either the coding or the non-coding strand at different locations throughout the entire operon (Fig. 3A). As expected, crRNA guides that target dsDNA sequences in the proximity of the promoter region, low GFP and RFP fluorescent signals were attained, indicating high transcriptional silencing of both the genes (Fig. 3B and C, crRNAs A1/A2). Strikingly, although the transcriptional silencing of the fluorescent reporter genes by the MmuCas12u1 protein was weak for crRNA guides that target dsDNA towards the end of the operon (crRNAs D2/E1/E2), relatively strong repression of both the red and green fluorescence signal was observed for the cells expressing the wild-type MmuCas12u1 (Figure 2B and C, crRNAs D2/E1). The loss of red as well as the green fluorescence upon binding to the downstream *gfp* gene indicates that transcription and/or translation of the whole mRNA is being affected by MmuCas12u1 (Fig. 3B and C, crRNAs D1/D2). Moreover, the crRNA guides that target dsDNA sequences downstream the terminator (crRNAs E2/F1/F2) resulted in undetectable loss of fluorescence, suggesting a transcription-associated *trans* cleavage of nascent mRNA by MmuCas12u1. In addition to the RuvC nuclease domain, the zinc finger domain was also mutated in both MmuCas12u1 and MmuCas12u1 to generate double mutants (H549A & C552A), MmuCas12u1-ZF and MmuCas12u1-ZF, respectively. MmuCas12u1-ZF and MmuCas12u1-ZF silenced pTarget-operon using crRNA A1, D2 and E1. All four Mmu variants, MmuCas12u1, MmuCas12u1, MmuCas12u1-ZF and MmuCas12u1-ZF silenced RFP and GFP equally well with guides targeting the promoter, indicating similar dsDNA binding properties (Fig. 3D and E, crRNA A1). Likewise, no difference in silencing was detected between MmuCas12u1 and MmuCas12u1-ZF for all three crRNA's tested. Interestingly, however, MmuCas12u1-ZF did result in a reduced silencing effect when compared to MmuCas12u1, of which silencing was similar to that of MmuCas12u1. This strongly suggests that the zinc finger domain is involved in the activity of the RuvC, possibly through strengthening the binding of the mRNA target.

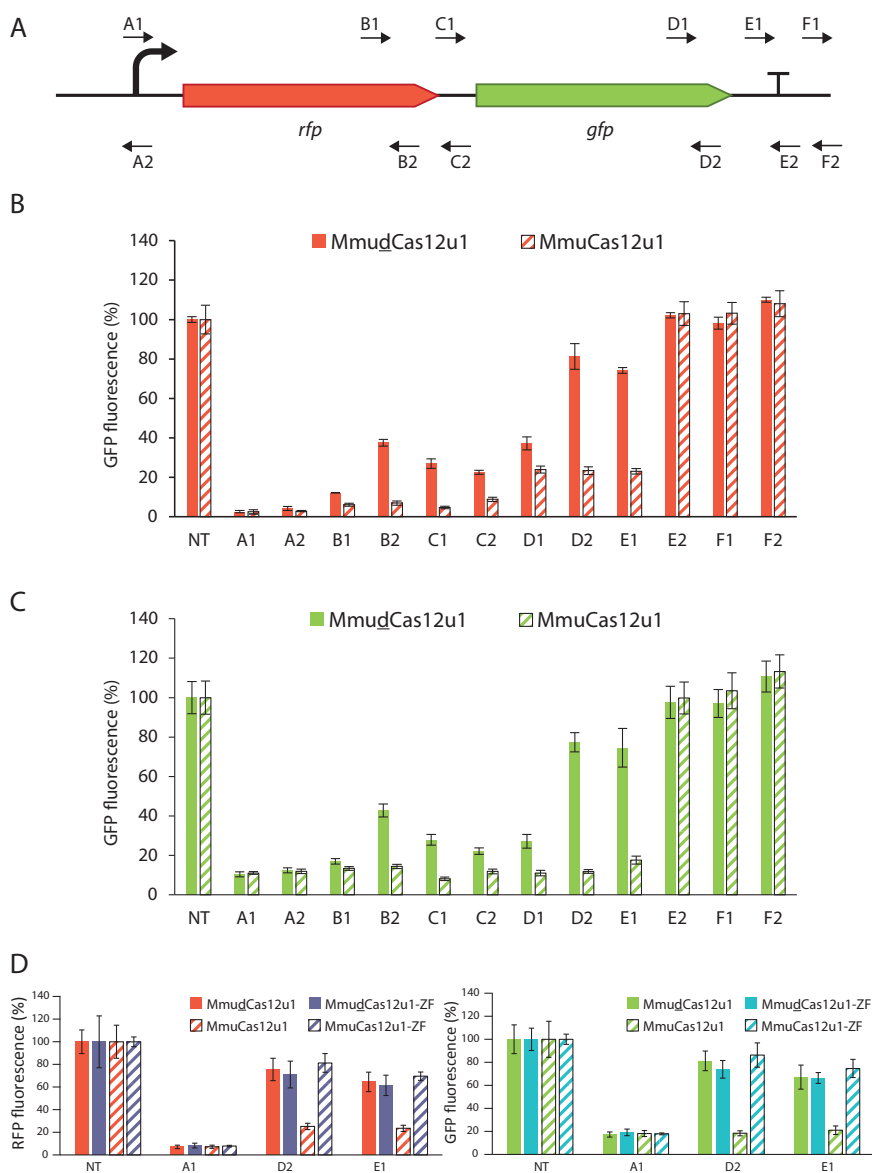


Figure 3 | Comparison of transcriptional silencing by MmuCas12u1 and MmudCas12u1. (A) Schematic of the pTarget-operon, including the bi-cistronic operon encoding the *rfp* and *gfp* genes. The arrows indicate the crRNAs used for targeting by MmuCas12u1 and MmudCas12u1 proteins (A1 to F2). **(B)** RFP fluorescence detected in the cells upon MmudCas12u1 and MmuCas12u1 targeting using the individual spacers ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer. **(C)** GFP fluorescence detected in the cells upon MmudCas12u1 and MmuCas12u1 targeting using the individual spacers ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer. **(D)** RFP (left) and GFP (right) fluorescence detected in the cells upon MmudCas12u1, MmuCas12u1, MmudCas12u1_ZF and MmuCas12u1_ZF targeting using the individual spacers ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer.

To test the reach of the mRNA targeting, we cloned a target plasmid (pTarget-divergent) with two fluorescence reporter genes, *rfp* and *gfp*, under the transcriptional control of two divergent constitutive promoters, P_{taq} and $P_{lacIq'}$ respectively (Fig. 4A). *E. coli* cells harboring either the pCas-MmudCas12u1 or pCas-MmuCas12u1 and the pTarget-divergent were transformed with a set of different CRISPR array plasmids (pCRISPR-a_f) containing spacers targeting different locations on the promoters and on the coding strand of either *rfp* or *gfp* (Fig. 4A). Specific repression of only the targeted reporter gene was observed, indicating only local interference. And yet again, the wild-type MmuCas12u1 generally performed better than the MmudCas12u1 in silencing the expression of the reporter gene (Fig. 4B and C). In addition to fluorescence measurements, mRNA transcripts were also measured by quantitative reverse transcription PCR (RT-qPCR) and found similar trends to that of the fluorescent signal (Fig. 4D and E). The increase in GFP fluorescence upon repression of the *rfp* gene, is most likely due to the relief of the burden on the transcription and translation machinery to produce both GFP and RFP as *gfp* transcripts also increase when targeting *rfp* (Fig. 4D,E). Interestingly, inefficient *gfp* repression is found for MmudCas12u1 when guided by spacer-e, but not for MmuCas12u1 (Fig. 4D and E, crRNA e). The lack in spacer efficiency in spacer-e suggests sequence- and context-dependent loss of RNA-directed nuclease activity, most likely due to hindering RNA secondary structures, similar to what has been observed for Cas12a (250). However, this lack of spacer efficiency does not affect silencing by MmuCas12u1. Collectively, these observations point toward a novel mechanism where crRNA-guided binding of MmuCas12u1 to a transcriptionally active dsDNA triggers it to cleave nascent mRNA. Cleavage of the mRNA appears to be confined to the transcript of the target DNA, rather than collateral cleavage activity that has been reported for some of the type V and type VI effectors (230, 246).

After determining local repression by MmuCas12u1, the next step was to silence both fluorescent proteins simultaneously, in other words *in vivo* multiplex gene silencing. MmuCas12u1 or MmudCas12u1 was guided by a single crRNA array resulting in two mature crRNA guides, one targeting *rfp* and the other *gfp*, on the pTarget-divergent plasmid (Fig. S8). Both RFP and GFP silencing was achieved, showing successful *in vivo* multiplex silencing by MmuCas12u1 and MmudCas12u1.

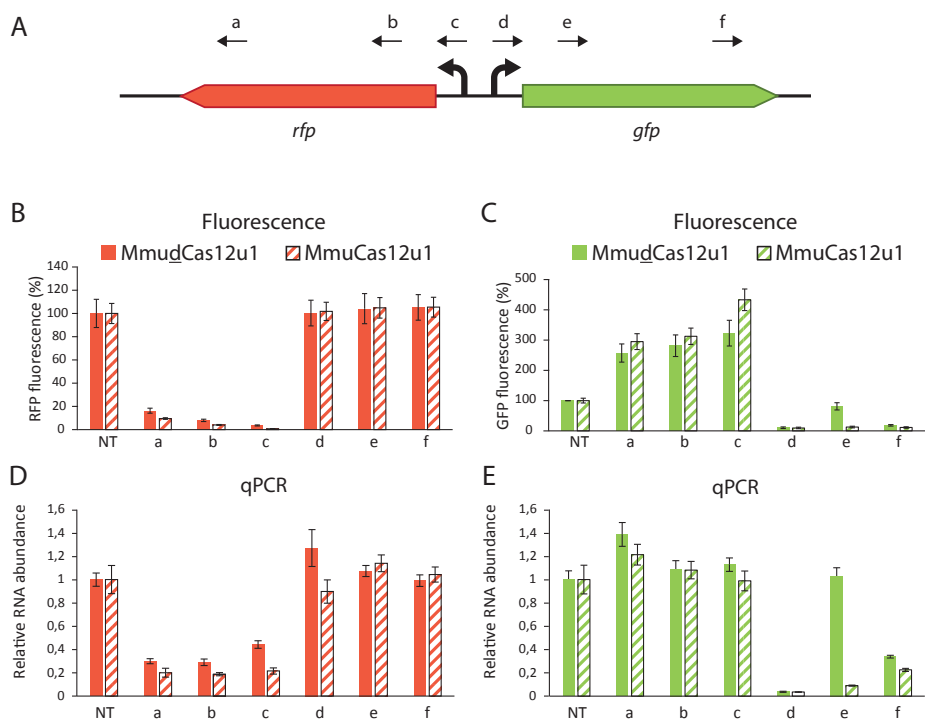


Figure 4 | MmuCas12u1 is activated by dsDNA binding to cleave nascent RNA transcripts. (A) Schematic of the pTarget-divergent, including the *rfp* and *gfp* genes under the transcriptional control of two different constitutive promoters, P_{taq} and P_{lacIq} . The arrows indicate crRNAs used for targeting by MmuCas12u1 and MmuCas12u1 proteins using the respective spacers (a to f). (B) RFP fluorescence detected in the cells upon MmuCas12u1 and MmuCas12u1 targeting using the individual spacers (n = 3; error bars represent mean \pm SD). NT refers to a non-targeting spacer. (C) GFP fluorescence detected in the cells upon MmuCas12u1 and MmuCas12u1 targeting using the individual spacers (n = 3; error bars represent mean \pm SD). NT refers to a non-targeting spacer; No PAM refers to a spacer targeting protospacer next to a GGCG PAM (non-functional PAM). (D) Relative *rfp* mRNA in the cells upon MmuCas12u1 and MmuCas12u1 targeting using the individual spacers (n = 3; error bars represent mean \pm SD) by RT-qPCR. NT refers to a non-targeting spacer. (E) Relative *gfp* mRNA in the cells upon MmuCas12u1 and MmuCas12u1 targeting using the individual spacers (n = 3; error bars represent mean \pm SD). NT refers to a non-targeting spacer.

To confirm the observations of the silencing activity of MmuCas12u1 in different *in vivo* experiments, an *in vitro* transcription and translation system was used, known as TXTL (251). The TXTL reaction consists of *E. coli* cell-free extract, salts, and buffers that provide amino acids and an ATP regeneration system. Using the TXTL system, an alternative *in vitro* approach is used to reveal mechanistic features of Cas nucleases (252, 253). pCas plasmids expressing the Cas nuclease and the pCRISPR plasmids expressing the guide, were initially pre-expressed in a TXTL reaction. This pre-expression was then subsequently used in a new TXTL reaction containing a deGFP plasmid (pdeGFP). deGFP fluorescence was measured over time to assess deGFP repression (Fig. 5B). MmuCas12u1 and MmudCas12u1 targeted pdeGFP at the promoter (crRNA 1), at the 3' end of *degfp* (crRNA 2) or at the vector backbone (crRNA 3) (Fig. 5A). deGFP repression was achieved in the TXTL using MmuCas12u1, MmudCas12u1 (Fig. 5C). FndCas12a was used as a control using a non-targeting spacer (NT) and a spacer targeting the promoter (crRNA 1). Similar as in our operon repression experiment (Fig. 3), MmuCas12u1 has higher silencing activity compared to MmudCas12u1 when targeting the transcribed region of *gfp* (Fig. 5C, crRNA 2). The same spacers were later tested *in vivo* and similar results were found (Fig. S9). Suggesting again for a dsDNA activated, mRNA interference activity by MmuCas12u1. However, endogenous cell nucleases are present in both *in vivo* and in the TXTL system, which can influence the silencing activity detected in the assay. To exclude endogenous cell nucleases, deGFP silencing is currently being tested in the PURE system, which contains only purified proteins involved in the transcription and translation machinery (251). Another explanation for the increased repression by MmuCas12u1 would be an enhanced binding affinity to dsDNA, as a result of which the RNA polymerase is unable to remove MmuCas12u1 during transcription. To investigate whether enhanced repression is due to stronger binding by MmuCas12u1, a surface plasmon resonance (SPR) analysis was done to study the kinetics of the protein/DNA interaction (254). Analysis by SPR indicated similar binding affinity between MmuCas12u1 and MmudCas12u1 for ssDNA with an association constant (K_{on}) of $6.26 \pm 0.18 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$ and $6.95 \pm 0.04 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$, respectively (Fig. S10). Furthermore, SPR experiments using dsDNA are currently ongoing.

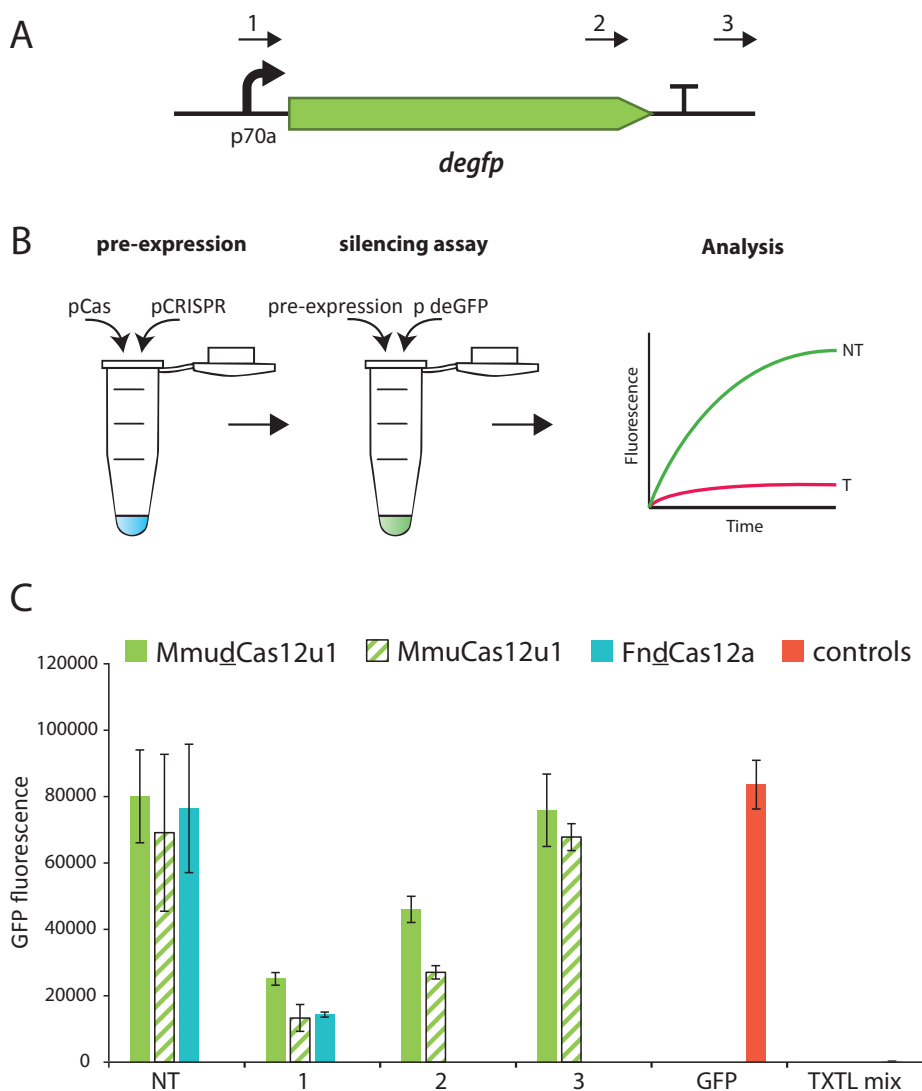


Figure 5 | MmuCas12u1 targeting deGFP in a TXTL system. (A) Workflow of deGFP silencing in TXTL. pCas and pCRISPR are pre-expressed in a TXTL reaction, which is then used in a subsequent TXTL reaction containing pdeGFP. The final TXTL reaction is incubated overnight and GFP fluorescence is measured over time to assess GFP silencing. **(B)** Schematic of the pdeGFP regulated by P70a promoter. The arrows indicate the crRNAs used for targeting by MmuCas12u1 and MmudCas12u1 proteins (crRNA 1, 2, and 3). **(C)** GFP fluorescence detected in TXTL upon MmudCas12u1 and MmuCas12u1 targeting using the individual spacers ($n = 4$; error bars represent mean \pm SD). NT refers to a non-targeting spacer. FndCas12a was used as a positive control with crRNA NT and crRNA 1. Other controls include only deGFP plasmid in TXTL mix and only TXTL mix.

Aside from investigating the RuvC domain activity, mismatch tolerance of MmuCas12u1 was also examined. We introduced single mismatches, tiled 2-nucleotide mismatches and tiled 4-nucleotide mismatches across the protospacer in the target *gfp* gene (Fig. S10A). MmuCas12u1 appeared to be relatively tolerant to most single mismatches (Fig. S10B), except for the mismatch at the PAM-proximal position 8th. This is in contrast to Cas12a which is highly sensitive to single or double substitution in most positions between 1 and 18 (255). Double and quadruple mismatches at PAM-proximal positions 1 to 11 severely impaired the MmuCas12u1 silencing activity (Fig. S10C and D), resembling a seed-like sequence (256). Notably, even though some mismatches impaired the activity of MmuCas12u1 in GFP repression, the effect of the mismatches on GFP silencing by the wild-type MmuCas12u1 was much less pronounced (Fig. S10B-D).

In conclusion, the characterization of the novel type V-U1, MmuCas12u1 protein, described here has revealed a unique mechanism. Akin to several other type V systems (Cas12a, h, i, j, k), MmuCas12u1 does not require a tracrRNA and is able to process its own pre-crRNA. In addition, specific crRNA-guide dependent binding has been demonstrated to dsDNA with 5' TTN-PAM. However, instead of DNA cleavage, MmuCas12u1 appears to target the nascent RNA during transcription of the targeted DNA. Although no direct evidence of RNA cleavage is currently available, RuvC-dependent ribonuclease activity seems most likely given the observed difference in silencing efficiency between the wild-type MmuCas12u1 and its catalytically inactivated variant, MmuCas12u1 (Fig. 3, 4, 5, S5, S6, S7). In addition, the zinc finger domain is speculated to also be involved in the silencing activity of MmuCas12u1. The implications of such a DNA binding RNA cleavage CRISPR-Cas system and the biological relevance of MmuCas12u1 is currently unclear and will be addressed in future research. In the absence of an MmuCas12u1 crystal structure, the molecular basis of the mismatch tolerance, crRNA binding/maturation and dsDNA binding mechanisms remain elusive. Moreover, the small size, multiplexing capability and potential activity of MmuCas12u1 in mammalian cells, which is currently being assessed, might facilitate delivery for applications in therapeutics and biomedical research (257, 258). The PAM-dependent DNA-targeting ability of Mmu-like Cas12 variants can be utilized to recruit transcriptional activators or repressors (259), as well as base editing enzymes (260). These are particularly interesting applications for understanding the molecular pathology of a range of human diseases as well as to develop novel therapeutic strategies to treat these diseases. The exciting finding that these miniature CRISPR-Cas effectors can accommodate crRNA and conduct targeted DNA binding and nascent mRNA cleavage underscores the rich natural functional diversity of CRISPR-Cas systems. We anticipate that the ongoing combination of biochemical and structural studies will reveal exciting insights into the molecular mechanisms of MmuCas12u1 in the near future.

Materials & Methods

Bacterial strains and growth conditions

Bacterial strains used for the cloning and propagation of plasmids in the current study are *E. coli* DH5 α and DH10 β . For protein expression, the *E. coli* Rosetta (DE3) (EMD Millipore) was used. The *E. coli* strains were routinely cultured at 37 °C and 220 rpm, unless specified, in either Luria Bertani medium (LB) [10 g L⁻¹ peptone (Oxoid), 5 g L⁻¹ yeast extract (BD), 10 g L⁻¹ NaCl (Acros)] or M9TG minimal medium [1xM9 salts (Sigma), 10 g L⁻¹ tryptone (Oxoid), 5g L⁻¹ glycerol (Acros)]. Plasmids were maintained with ampicillin (100 mg mL⁻¹), chloramphenicol (35 mg mL⁻¹), and/or kanamycin (50 mg mL⁻¹) as needed. Liquid media was supplemented with IPTG as specified. All fluorescence loss experiments were carried out in the derivative of *E. coli* BW25113 strain lacking the *lacI*, *lacZ* genes and the type I-E CRISPR-Cas system.

Plasmid construction

The plasmids constructed and the oligonucleotides (IDT) used for cloning and sequencing are listed in Table S1.

E. coli codon-harmonized *mmuCas12u1* gene was inserted into the plasmid pML-1B backbone (obtained from the UC Berkeley MacroLab, Addgene #29653) by ligation-independent cloning using oligonucleotides to generate a protein expression construct encoding the MmuCas12u1 polypeptide sequence (residues 1–596) fused with an N-terminal tag comprising a hexahistidine sequence and a Tobacco Etch Virus (TEV) protease cleavage site.

The three plasmids used for the PAM-SCNR screening platform were based on the previously published protocol (242). The *mmuCas12u1* and *mmuCas12u1* genes were inserted into the pBAD33 vector backbone under the control of the constitutive J23108 promoter to generate the pCas-MmuCas12u1 and pCas-MmuCas12u1 plasmids, respectively. The pCRISPR guideRNA plasmid series were generated by inserting a CRISPR array downstream the constitutive J23119 promoter in pBAD18 backbone. The pTarget-PS plasmid is comprised of the PAM-SCANR NOT gate-based circuit in a pAU66 plasmid backbone.

The pCas-MmuCas12u1 and pCas-MmuCas12u1 plasmids were constructed using NEBuilder® HiFi DNA Assembly (NEB). The fragments for assembling the plasmids were amplified by PCR using Q5® High-Fidelity 2X Master Mix (NEB). The catalytically inactive MmuCas12u1 (MmuCas12u1) gene fragment was created by site-directed mutagenesis of the aspartic acid in the RuvC domain to an alanine (D485A). Zinc finger mutants pCas-MmuCas12u1-ZF and pCas-MmuCas12u1-ZF were constructed by NEBuilder® HiFi DNA Assembly (NEB). Backbone

pCasMmuCas12u1 was digested with restriction enzymes *AvrII* and *HindIII* and MmuCas12u1 fragments were amplified by PCR.

The pCRISPR plasmids for MmuCas12u1 were constructed by restriction-digestion and ligation. By PCR amplification, a *BbsI* restriction site and a CRISPR repeat was added as an overhang to the vector fragment. The amplified fragment was digested (using *KpnI* and *BbsI* enzymes) and ligated to a spacer-repeat sequence generated by annealing two oligonucleotides containing complementary overhangs. Using the same method, a pCRISPR_NT plasmid was created, containing a spacer flanked by *BbsI* sites. Other CRISPR plasmids containing the different targeting spacers were created using pCRISPR-NT by digestion and ligation. Longer CRISPR arrays such as the four-spacer CRISPR array were created by annealing two oligonucleotides to create spacer-repeat fragments. Fragments were design to contain compatible overhangs to other spacer-repeat fragments. Spacer-repeat fragments are ligated together and PCR amplified to yield spacer-repeat-spacer-repeat-spacer flanked by *BbsI* restriction sites. The amplified linear fragment is then cloned into pCRISPR by digestion and ligation.

The pTarget-GFP plasmid was constructed using *BamHI* restriction and ligation of a linear P_{lacIq} and GFP gene fragment amplified from the pTarget-PS plasmid. pTarget-GFP containing different PAMs were constructed by site directed mutagenesis. The pTarget-operon plasmid was constructed by digesting the pTarget-GFP plasmid with *BamHI* enzyme to generate a linear vector which was assembled with an mRFP fragment containing compatible overhangs using the NEBuilder® HiFi DNA Assembly. The pTarget-divergent plasmid was constructed using a fragment of pTarget-GFP digested with the restriction enzymes, *AatII* and *BamHI* and subsequent ligated with a mRFP fragment under the control of a *Taq* promoter.

For testing the mismatch tolerance, targets were ordered as an oligonucleotide pair, which was phosphorylated with T4 PNK and annealed. The backbone pTarget-MM-BsmBI-entry was linearized with *BsmBI* and ligated to the target adaptors to create series pTarget-MM-[x], where x is the position from 1 to 20 on the protospacer where the mismatch is introduced. A frameshift (pTarget-MM-[FS]) was made in the *gfp* by digesting pTarget-MM-[WT] with *BstBI*, filling in the overhang with Klenow fragment and re-circularizing the plasmid. The CRISPR array plasmids pCRISPR-MM-[WT] were created using the same method described above.

pCRISPR plasmids for Cas12a were constructed by restriction digestion of pCas12a-pCRISPR-RFP with restriction enzyme *BbsI*. *BbsI* digestion removes a *rfp* gene flanked by two Cas12a repeats. Spacers are created by annealing two oligonucleotides containing complementary overhangs and subsequently ligated to pCRISPR-Cas12a.

MmuCas12u1 protein expression and purification

The purification protocol was adapted from established Cas12a purification methods previously (219). Briefly, the *mmuCas12u1* gene was heterologously expressed in *E. coli* and purified using a combination of Ni²⁺ affinity, cation exchange and gel filtration chromatography steps. Three liters of LB growth medium with 100 µg mL⁻¹ ampicillin was inoculated with 30 mL overnight culture of Rosetta (DE3) (EMD Millipore) cells containing the expression construct. Cultures were grown to an OD_{600nm} of 0.5 - 0.6; expression was induced by the addition of IPTG to a final concentration of 0.2 mM and incubation was continued at 18 °C overnight. Cells were harvested by centrifugation and the cell pellet was resuspended in 50 mL lysis buffer (20 mM Tris-HCl pH 8, 500 mM NaCl, 5mM imidazole, supplemented with protease inhibitors (Roche) Cells were lysed by sonication and the lysates were centrifuged for 45 min at 4 °C at 30,000x g to remove insoluble material. The clarified lysate was applied to a 5 mL HisTrap HP column (GE Healthcare). The column was washed with 10 column volumes of wash buffer (20 mM Tris/HCl pH 8, 250 mM NaCl, 20 mM Imidazole) and bound protein was eluted in elution buffer (20 mM Tris/HCl pH 8, 250 mM NaCl, 250 mM Imidazole). Fractions containing pure proteins were pooled and TEV protease was added in a 1:100 (w/w) ratio. The sample was dialyzed against Dialysis buffer (20 mM HEPES-KOH pH 7.5, 250 mM KCl) at 4 °C overnight. For further purification the protein was diluted 1:1 with 10 mM HEPES KOH (pH 7.5) and loaded on a HisTrap Heparin HP column (GE Healthcare). The column was washed with IEX Buffer A (20 mM HEPES-KOH pH 7.5, 150 mM KCl) and eluted with IEX Buffer B (20 mM HEPES-KOH pH 7.5, 2 M KCl) by applying a gradient from 0% to 50% over a total volume of 60 ml. Peak fractions were analysed by SDS-PAGE and fractions containing the Cas12u1 protein were combined, and DTT (Sigma-Aldrich) was added to a final concentration of 1 mM. The protein was fractionated on a HiLoad 16/600 Superdex 200 gel filtration column (GE Healthcare) and eluted with SEC buffer (20mM HEPES-KOH pH 7.5, 500mM KCl, 1mM DTT). Peak fractions were combined, concentrated to 10 mg mL⁻¹, flash frozen in liquid nitrogen and either used directly for biochemical assays or frozen at -80°C for storage.

Pre-crRNA processing

The pre-crRNA processing assay was conducted with ~varying amounts of MmuCas12u1 nuclease and ~100nM pre-crRNA. The assay was conducted in Cas9 Nuclease Reaction Buffer (NEB), in a total volume of 15 µl, at 37°C for an hour and quenched with 2 µL proteinase K (NEB) at 30 °C for 30 minutes. Subsequently, the samples were analyzed on a 10% urea-PAGE gel stained with SYBRTM Gold Nucleic Acid Stain (Invitrogen).

PAM-SCNR assay

A day prior to sorting, *E. coli* cells harboring the pCas and pCRISPR plasmids were made chemically competent and were transformed with the pTarget-PS plasmid containing the randomized 4N PAM library. After recovery, the transformation mix

was used to inoculate 10 mL LB medium (1:100) and grown overnight. The next day, the culture was used to inoculate 10 mL LB medium (1:100) and supplemented with different concentrations (0, 10, 1000 μ M) of IPTG and cultured to an OD₆₀₀ of ~0.5. Subsequently, the cultures were diluted 1:100 in phosphate buffer saline (PBS) and GFP-positive cells were sorted using a Sony SH800S Cell Sorter. GFP was excited using a blue laser (485 nm) and detected using a 525/50 filter. Pure cultures of either GFP expressing fluorescent or non-fluorescent cells were used as controls to set the gating and the sensitivity for the forward scatter, side scatter and photomultiplier tubes (PMT). A minimum of 100,000 single cell events were sorted and collected in 5 mL LB medium and grown overnight at 37 °C. The following day, the culture was used to inoculate (1:100) 10 mL fresh LB medium and grown for 3 hours. The cultures were diluted 1:100 in PBS and sorted for GFP positive cells. 500,000 single cell events were collected in 1 mL PBS, which was then immediately re-sorted to collect 50,000 single cell events in 5 mL LB medium and grown overnight. The next day, the culture was used to inoculate (1:100) 10 mL LB medium and grown overnight. The next day, plasmids were extracted and sent for deep sequencing.

Fluorescence repression assays

For the silencing assays, *E. coli* cells harboring either the pCas-MmudCas12u1 or the pMmuCas12u1 and the corresponding target plasmids were made chemically competent and transformed with the pCRISPR library. For the 5'-NTTN PAM determination assays, cells harboring either the pCas-MmudCas12u1 or the pMmuCas12u1 and the pCRISPR plasmid were made competent and then transformed with the target plasmid. For the mismatch tolerance assays, chemically competent *E. coli* BW225 cells harboring either targeting plasmid pCRISPR-MM-[WT] or non-targeting plasmid pCRISPR-BbsI, and either pCas-MmuCas12u1 or pCas-MmudCas12u1 were transformed with pTarget-MM-[x].

After recovery, the transformation mix was diluted 2 μ L:200 μ L M9TG medium in a 96 well 2 mL master block (Greiner) and sealed using a gas-permeable membrane (Sigma, AeraSeal™) and grown overnight at 37 °C at 900 rpm overnight. The next day, the cells were diluted 1:10000 in triplicate in fresh M9TG medium in a 96-wells masterblock and grown overnight at 37°C. Overnight cultures were then used for fluorescence measurements.

Plate reader measurements

Overnight cultures were diluted 1:10 in 200 μ L PBS for the mismatch tolerance assays and measured on a Biotek Synergy MX microplate reader a Synergy MX microplate reader. GFP and RFP fluorescence were measured with an excitation of 485 nm and 555 nm, respectively and an emission at 585 nm. GFP and RFP were measured with gain of 75 and 100, respectively.

Fluorescence was calculated as

$$\frac{\text{average} \left(\frac{Fl_{x_{\text{targeting}}} - Fl_{\text{Blank}}}{OD600_{x_{\text{targeting}}} - OD600_{\text{Blank}}} \right) - \text{average} \left(\frac{Fl_{FS} - Fl_{\text{Blank}}}{OD600_{FS} - OD600_{\text{Blank}}} \right)}{\text{average} \left(\frac{Fl_{x_{\text{non-targeting}}} - Fl_{\text{Blank}}}{OD600_{x_{\text{non-targeting}}} - OD600_{\text{Blank}}} \right) - \text{average} \left(\frac{Fl_{FS} - Fl_{\text{Blank}}}{OD600_{FS} - OD600_{\text{Blank}}} \right)}$$

RT-qPCR analysis

10 mL LB with 50 mg mL⁻¹ kanamycin, 34 mg mL⁻¹ chloramphenicol and 100 mg mL⁻¹ ampicillin was inoculated 1:1000 from a preculture. Cells were grown to an OD600 of 0.6 and cooled down on ice-water. Cells were pelleted and resuspended in 250 µL of 50 mM Tris-HCl pH8, 10 mM EDTA and 10 mM DTT. Cells were then lysed with 250 µL of [0.2 M NaOH and 1% SDS]. Protein, genomic DNA and SDS were precipitated by adding 250 µL [1.8 M potassium acetate and 1.2 M acetic acid]. Debris was pelleted in a microcentrifuge tube and 650 µL was transferred to a new Eppendorf tube. RNA was precipitated by adding 650 µL isopropanol and centrifuging for 5 minutes at maximum speed. RNA pellets were washed with 500 µL of [10 mM Tris-HCl pH8 and 70% ethanol] and dried in a laminar flow cabinet. Pellets were dissolved in 100 µL DNaseI buffer (NEB) with 0.25 µL DNase I (NEB) and incubated at 37 °C for 30 minutes. First, 300 µL of DNaseI buffer was added and then 200 µL of Roti aqua phenol (Roth). The phases were separated by centrifugation and 300 µL of the aqueous phase was transferred to a new Eppendorf tube. 300 µL of isopropanol was added to the aqueous phase and the mixture was loaded on a silica column (Thermo K0702). The RNA was washed twice with 400 µL [10 mM Tris-HCl pH8, 70% ethanol and 100 mM NaCl]. Finally, the RNA was eluted into 50 µL of [1 mM Tris-HCl pH8, 0.1 mM EDTA]. The RNA was diluted to 1 g/L in water and cDNA was generated with the Maxima H minus (Thermo) reverse transcriptase. RT-qPCR was performed with the SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad) using cDNA derived from 10 ng of total RNA in a 10 µL reaction.

In vitro TXTL assay

TXTL experiments were conducted in the laboratory of Chase Beisal at the Helmholtz centre for infection research in Würzburg, Germany. The TXTL reaction consisted out of myTXTL® master mix, pCas, pCRISPR and p70a-deGFP. The myTXTL® Sigma 70 Master Mix and p70a-deGFP was purchased from Arbor Biosciences. pCas and pCRISPR were plasmids used for *in vivo* silencing and were prepared by midiprep using the ZymoPURE™ II Plasmid Midiprep Kit (Zymo Research), followed by PCR purification using the DNA Clean & Concentrator-5 (Zymo Research). TXTL reactions were prepared according to (253). pCas (4nM) and pCRISPR (4nM) were first pre-

incubated in a TXTL reaction for 16 hours in 29 °C. 1 µL of the pre-incubated mix was added together with pTarget-eGFP (1mM) in a new TXTL mix with an end volume of 12 µL. The final reaction was pipetted into a 96-well plate using a Labcyte Echo 525 acoustic liquid dispensing system. Each well contained a 3 µL reaction with four replicates per sample. The 96-well plate was then incubated for 16 hours at 29°C in a Synergy Neo2 (Biotek) plate reader. deGFP fluorescence was measured every 3 min with an excitation and emission of 485 nm and 528 nm, respectively. Also, bandwidth and again were set to 13 nm and 60, respectively.

For pre-crRNA processing, the pCas and pCRISPR mixture was incubated at 29 °C for five hours in a thermocycler, and total RNA was extracted using Direct-zol RNA MiniPrep kit following the manufacturer's instructions (Zymo Research).

Northern blot

For Northern blotting analysis, 5 µg of each RNA sample obtained from TXTL was put on an 8% polyacrylamide gel (7 M urea) at 300V for 140 min. RNA was transferred onto Hybond-XL membranes (Amersham Hybond-XL, GE Healthcare) using an Electrobloetter using 50V for 1 h at 4 °C (Tank-Elektrobloetter Web M, PerfectBlue) and crosslinked with UV-light for a total of 0.12 Joules (UV-lamp T8C; 254 nm, 8 W). Hybridization occurred overnight in 17 mL Roti-Hybri-Quick buffer with 5 µL γ-32P-ATP end-labeled oligodeoxyribonucleotides at 42 °C. The membrane was visualized using a Phosphorimager (Typhoon FLA 7000, GE Healthcare).

Surface plasmon resonance

A 50 nt biotinylated oligo (Table S1) containing the MmuCas12u1 target site was obtained from IDT (IDT, Leuven, Belgium) and solubilized in 25 mM Hepes, 150 mM KCl pH 7.5. Surface plasmon resonance (SPR) spectrometry was performed on a Biacore T100 (GE Healthcare) at 25°C. A CM5 sensor chip surface was derivatized with 2500 response units (RU) of streptavidin (Invitrogen) using the amine coupling kit (GE Healthcare). Subsequently 9 RU of ssDNA oligo was immobilized on flow cell 2. The MmuCas12u1-RNA and MmuCas12u1-RNA complex were formed by diluting the protein into SPR running buffer (20 mM Hepes, 150mM KCl, and 0.05% Tween 20, pH 7.9) containing a 1.4-fold excess of RNA to a final concentration of 500 nM. The ribonucleoprotein complex was injected across the chip at 50 µL/min. The injection phase was performed for 1 minute, dissociation was followed for 100 minutes after injection. A model describing a 1:1 binding mode was fitted to the data using the BioEvaluation Software (GE Healthcare) to obtain approximate rate constants for binding and dissociation (only approximation as binding is close to irreversible). Plots were created using GraphPad Prism version 8.2.3.

Acknowledgements

We would like to thank Sanne Klompe, Jasper Groen, Yuxin Zhang, Patrick Barendse for their technical assistance and Christian Sudfeld for his assistance in the cell sorting experiments. J.v.d.O. is supported by the NWO/TOP grant 714.015.001. W.X.Y and D.A.S are employees and shareholders of Arbor Biotechnologies, Inc. K.S.M. and E.V.K. are supported by the intramural program of the U.S. Department of Health and Human Services (to the National Library of Medicine).

Author contributions

W.W., P.M., and J.v.d.O., conceived and designed the study. W.W., P.M., S.C.A.C., K.K., T.S.K., S.P. conducted all the experimental work and analyzed the data. K.S.M. and E.V.K. provided input on the computational and phylogenetic analysis. W.X.Y and D.A.S performed the NGS sequencing experiments. C.L. and C.L.B. performed the northern blot experiments. C.L. and J.H.G.K. performed the SPR experiments. W.W., P.M., and J.v.d.O. wrote the manuscript with input from all authors.

Competing interests

A patent application has been filed related to this work.

Corresponding author

Correspondence and requests for materials should be addressed to J.v.d.O. (john.vanderoost@wur.nl).


```

WP_116532935.1 hypot -----METLIY--EYGC-----RLD 1
SSPE20750.1 transposa -----MTRSVTITNPAQSAATVDASRSAISIPKYCDASTASY--EYGA-----RLD 4
WP_105479500.1 hypot -----MDASRSAISIPKYCDASTASY--EYGA-----RLD 28
JW42488.1 hypothehti -----MKITPASLPQGDVRIY--EFGA-----RLD 23
WP_018991635.1 hypot -----MIY--EFGV-----RID 10
WP_018079340.1 hypot -----MSIKPSLLPQGNVLIY--EYGA-----RLD 24
WP_081130164.1 hypot -----MKLSPALPTGDVLIY--EYGA-----RVD 23
WP_064217851.1 hypot -----MSTIY--EYGV-----RLE 13
WP_051690567.1 trans -----MSQIKIVPQINGSOLVY--KYGV--RNN 24
WP_051690567.1 trans -----MQGQHVY--EYGA-----RID 15
OPC35369.1 hypothehti -----MKPPTPTLRAPQIQGQHVY--EYGA-----RID 28
WP_077272831.1 trans -----MMKTY--VFGLLPP-----1
WP_106353755.1 trans -----MLKTY--VFGLLPP-----12
WP_045707069.1 trans -----MKRQEDTEALVY--AYGA-----RIP 20
WP_018234394.1 trans -----MAFSGV--TISVHY--TW 14
WP_064888210.1 hypot -----MTSIPGTAVTVH--TFGVHY--RW 20
WP_063045032.1 hypot -----MMAVTY--IIGIPYPSGW--18
MKWV19589.1 hypothehti -----MATVH--TAGVHY--RWT 15
WP_061006603.1 hypot -----MTTWTVH--TMGVHY--KW 15
Upred_sec.str. -----EEEE--E-DEE-----E-
WP_095663130.1 hypot -----MAPRDEPAPGPSELEGATVH--TMGVHY--RW 30
WP_073879989.1 hypot -----NASDDEPVQPGMTPEGATVH--TMGVHY--RW 30
WP_064893148.1 hypot -----MAPDDEPQPGMTPEGATVH--TMGVHY--RW 30
KBF95043.1 hypothehti -----MTWASDDEPVQPGVTPPEGATVH--TMGVHY--RW 32
WP_036456351.1 hypot -----MGVHY--RW 7
CDO91315.1 hypothehti -----MTWASDDEPVQPGVTPPEGATVH--TMGVHY--RW 32
WP_036473531.1 hypot -----MASDDEPVQPGVTPPEGATVH--TMGVHY--RW 30
WP_064942980.1 hypot -----MGVHY--RW 7
OOK65169.1 hypothehti -----MAVEQARVARPASNIAVH--TMGVHY--RW 27
WP_047323888.1 trans -----MWVASDGELEAAERAAVVEGDSRITVH--TMGVHY--RW 37
WP_101953221.1 hypot -----MVQPOWWLSERLWLCNDMWASKELQADERSVAGEVGDPTQITVH--TMGVHY--RW 54
GAGAB36148.1 hypothehti -----MTRVTVQ--TAGVHY--KW 15
WP_039994403.1 hypot -----MHY--KW 5
KBF41925.1 hypothehti -----MAVTVQ--TMGVHY--RW 14
WP_036444762.1 hypot -----MGVHY--RW 7
PPZ20932.1 hypothehti -----MIRIY--GYTLLPPTLNA-----16
WP_013159911.1 trans -----MPFGKKARHVKAY--QFGA-----17
PPZM90038.1 transposa -----MPRTDRARIMRAY--AYGADAPVSGW--24
WP_092118774.1 trans -----MPRIY--KYGJGKNEGPD 18
WP_052217029.1 hypot -----MFGHESKPCRVY--EYGLTPTAG--22
WP_081908191.1 trans -----MSRLAEARTRYIQAGOKRLGKIKRGRFEMETAATKNYLAUSFGCLSPTRG--49
AG088270.1 Transposa -----MTVTTSTTPFGGIKTEVIR--KYGLLOPT--NWA 30
WP_011733919.1 trans MKRVTTIDGEOQTKGIVGIIAANHHTAEWLLTASVSAKVRPDEEAVETSSLVMTAPTRTEKYLYLVPEQOVPTIVR--KYGLLSPL--DW 95
WP_096876841.1 trans -----MITY--KYSL--KAP 12

```

131

133

WP_116532935.1	hypot	VRV-VRRKICP-DVRWALQFQAVE-----EBHVFKINGIPKKRPLAALHFGWSMS-DGRLRLAATCDS-----GDAAAQFHD	321
SFE20750.1	transposase	ARV-VRRRIGF-DAGWTIQLIVKR-----PRATMVVPG--ARKPLAAVFWATD-TSG-RKVAGIATG-----ADPGCARLWQ	349
WP_105479500.1	hypot	ARV-VRRRIGF-DAGWTIQLIVKR-----PRATMVVPG--ARKPLAAVFWATD-TSG-RKVAGIATG-----ADPGCARLWQ	332
OJW42488.1	hypotheteti	VRV-VRRRIGF-DAGWTIQLIVKR-----PRATMVVPG--ARKPLAAVFWATD-TSG-RKVAGIATG-----ADPGCARLWQ	327
WP_018991635.1	hypot	ARV-VRRKIGK-DYKWAQLMKP-----PPFVATY--ARKPLVSHFGMAAD-VG-RVVAIADS-----ADPHAARLWY	314
WP_018079340.1	hypot	CRV-VRRYCK-DYKWAQLMKP-----PIEQALAG--RKPLVAVFWAAN-DEG-RCVAGITDG-----ADPGQAVYVK	328
WP_081130164.1	hypot	ARV-VRRYCK-DEKWAQLMKP-----ATEPAMGHE--RKPLVAVFWAAN-DEG-RCVAGITDG-----ADPGQAVYVK	328
WP_064217851.1	hypot	ARV-VRRKTCG-RMKYIMQVINTA-----QIRQSDHGA--RKALLAVGMSAD-ISG-RVRCGITDA-----ADPELAQIIQ	318
WP_051609567.1	trans	AHT-VRRKAGR-KYQVELQIATLAE-----PINLLPDHR--RKPLVAVFWGSGD-EEG-RRLAGIADN-----ADPELARIIT	330
ORC35369.1	hypotheteti	VRI-VRRKGP-RYRYLQFINLAD-----PRLEVANR--RKPLVAVFWGSGD-EEG-RRLAGIADN-----ADPELARIIT	321
WP_077272831.1	trans	VRI-VRRKGP-RYRYLQFINLAD-----PRLEVANR--RKPLVAVFWGSGD-EEG-RRLAGIADN-----ADPELARIIT	330
WP_106353755.1	trans	ARV-VTRVGV-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_045707069.1	trans	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_102857306.1	hypot	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_018234394.1	trans	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_061559521.1	hypot	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_06488210.1	hypot	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_063045032.1	hypot	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
KMV19589.1	hypotheteti	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
WP_061006603.1	hypot	ARV-VTRVGA-SIKVEIHLVDM-----GVTLLPKQG--ELTAAVMGRIIT-ETG-LRVAALR-----FSDGSEEVIE	302
Upred_sec.str.		-----EEEEEE-----	
WP_095663130.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_073879989.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_064893148.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
KEF95043.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	380
WP_036456351.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	355
CD091315.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	380
WP_036473531.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_064942980.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
OOK65169.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_047323888.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	355
WP_101953221.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
GAB36148.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_039994403.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
KEF41925.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_036444762.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
PZN20932.1	hypotheteti	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_013159911.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
PZM90038.1	transposase	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_092118774.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_052217029.1	hypot	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_081908191.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
AG08270.1	transposase	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_011733919.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378
WP_096876841.1	trans	AKT-VTRVGA-DIRARLSITARL-----PRAVPRRG--DLPTLALHGWHAH-ETG-IYVAHWCRDPVVPPELADVIES-TVEGVSGRV	378

WP_1165323935.1	hypot	LPYDQEDFARAK-A-QOS-RQDT-----LRDEFVNLA-----LKGDNIAQGMWEDICIRLHMLRLP-----VQHIAPSP[LCH]-N	390
SPE20750.1	transposa	LPSPVEDIQRAS-ALQA-ARDA-----ARQDVVR-----KDTGCAVPVAQAEFTALVALPAQO-----VSQRRUAAF-C	415
WP_105479500.1	hypot	LPSPVEDIQRAS-ALQA-ARDA-----ARQDVVR-----KDTGCAVPVAQAEFTALVALPAQO-----VSQRRUAAF-C	398
OW42488.1	hypotheti	LPPEEAGIQHSG-EVES-RSV-----ARDNVATKAWHPDOLI-----DRAEQTEDEATPATR-SQAADLLVIRRLPATHVATRLHRL-A	411
WP_018901635.1	hypot	LPPEHEATIERAA-TIQS-ERDT-----SRDKMPL-----KAIENENWSEGLVLEALRLP-----AQHVATRLHRL-C	361
WP_018079340.1	hypot	LPPEVQSIIVRSS-ALQS-BDS-----ARDKMPV-----KEIE--VPDMDIESVESLPDPSPEVRLARADELKAIHRLP-----ANHVAITRLHRL-C	410
WP_081130164.1	hypot	LPPEVEDGIRRAA-EFOS-ARDE-----ARDVMTTKNIANG-----DAVALGESQMHGSEPMULRAKSEELSTIRLPQAQHVAPRLHRL-C	409
WP_064217851.1	hypot	LPPEIERNIQRAA-NIQG-RDQ-----ARDEAPK-----RADGSLPPBWDES---TQDYSHWK-----VLPANHMAASRI	383
WP_051609567.1	trans	LPPEDEDDIREAS-ALQA-RQDT-----YRDEFILR-----KEENTLIPKGE---TIDSEHNKIRKLPQAQHVANSRMHHL--	397
ORF33369.1	hypotheti	LPSPSEADITRAA-TIRS-QDII-----ELAAVQQR-KTEWK-----LPSLEG-DALRSOWEEFCGMPAHRISHRUHAL-A	390
WP_077272831.1	trans	LPSPSEADITRAA-TIRS-QDII-----ELAAVQQR-KTEWK-----LPSLEG-DALRSOWEEFCGMPAHRISHRUHAL-A	403
WP_106353755.1	trans	LPPEWMLRFDIME-DIEAGVRA-----AAADMPLI-----PELQAPAKLRH--SVERTIARPG-----SALGVRAI-G	364
WP_045077069.1	trans	LPPEWMLRFDIME-DIEAGVRA-----AAADMPLI-----PELQAPAKLRH--SVERTIARPG-----SALGVRAI-G	364
WP_102857306.1	hypot	LPPEWMLRFDIME-DIEAGVRA-----AAADMPLI-----PELQAPAKLRH--SVERTIARPG-----SALGVRAI-G	364
WP_018234394.1	trans	LPDWMRGMDQVE-RUSS-H-----LDENAMEV-----AAWV-----HAH-RDELPEKLTQ-----PAANWS-----PKGSKWLRD	392
WP_061559521.1	hypot	VPATHIDGFARAD-QTRA-QRGO-----ATRAQOLS-----VTWL-----AEHGPTDDPTPGG-VLDAATVEHWR-----GVAQFHAL-A	423
WP_064888210.1	hypot	VPATHIDGFARAD-QTRA-QRGO-----ATRAQOLS-----VTWL-----AEHGPTDDPTPGG-VLDAATVEHWR-----GVAQFHAL-A	423
WP_063045032.1	hypot	FNSAIFARLDRTD-QIGS-MRAE-----SLPHRDAL-AAWL-----HAHGVPQSGR-----ALITADVVRQWR-----SQAQFAAV-A	430
KMW19589.1	hypotheti	VPFAVPRVRAH-LTAS-IRDN-----RMNEIRAR-----VDYL-----AETGPRPHPSRGE-ELGAGNVRMWK-----SPNREAVL-A	428
WP_061006603.1	hypot	VPATHERRTRTE-NLAS-SGL-----ALDARDKV-----VGWL-----SDNDA---PYRDA-PIEAAIVKQWK-----SQRFASL-A	425
Jpred_sec.str.	EE	HHHH-----HHHHHHHH-----HHHHHHHH-----H-HHHHHH-----HHHHHH-H	
WP_073879989.1	hypot	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	446
WP_064893148.1	hypot	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	446
KEF95043.1	hypotheti	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	448
WP_036456351.1	hypot	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	423
CD091315.1	hypotheti	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	448
WP_036473531.1	hypot	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	446
WP_064942980.1	hypot	APSGVTRLERYA-ETAS-ARDA-----ALNDRDR-----STWL-----ADHGPRPHIRPDE-QITAAADAARWR-----SPARFAAL-A	423
OOK65169.1	hypotheti	APASIAARLIERHA-ATAS-VRHL-----ALKGIRDEV-----VGWL-----REGLPHIPARPOB-QITAAAVAQWR-----SGAQFAAL-A	442
WP_0473232388.1	trans	APASIAARLIERHA-ATAS-VRHL-----ALKGIRDEV-----VGWL-----REGLPHIPARPOB-QITAAAVAQWR-----SGAQFAAL-A	442
WP_101953221.1	hypot	APALATRIQVHLS-HTAS-TRDL-----ALDSRSKL-----TDWL-----DEHGPIPHSRPDE-QITAAAVAQWR-----SPARFAAL-A	454
GAB36148.1	hypotheti	APALATRIQVHLS-HTAS-TRDL-----ALDSRSKL-----TDWL-----DEHGPIPHSRPDE-QITAAAVAQWR-----SPARFAAL-A	471
WP_039994403.1	hypot	VEHLEQRVHHA-TTAS-TRDL-----AVDSRDTL-----VAVL-----TEHGPOPHVDGDP--ITAAASVQRWK-----APRFAWL-A	423
KEP41925.1	hypotheti	VPRIHTERTHSD-EIURS-QDII-----ALDAIRAK-----AAWL-----AEHGVPVHPRPEA-TVESGVDARWR-----SPARFAAL-A	426
WP_036444762.1	hypot	VPRIHTERTHSD-EIURS-QDII-----ALDAIRAK-----AAWL-----AEHGVPVHPRPEA-TVESGVDARWR-----SPARFAAL-A	419
PZN20932.1	hypotheti	VPESIVDRIEKSE-SIRS-IRDR-----HLNEMRPR-----IESI-----GAI--AGPPDEIVB--RCEHMHAWR-----SPARFAAL-A	380
WP_013159911.1	trans	LEGEDLQFSKVE-DURS-IRDQ-----HLNLEKEAL-----AAWL-----EAPPALPQWLAB--ETKTLPOWR-----SPARFAAL-F	365
PZM90038.1	transposa	LPPEWVAAMAQCD-RIRG-YEDSDDLTPEGGLEQARAR-----SAWV-----DAODPTLPEWIRY--ARREWGRW-----SHGRFAAL-A	359
WP_092118774.1	trans	LDVGYATQDRLD-GIKS-VLDK-----NRDQTAQ-----LDWG-----KQHEHLPHWLDLQKSVVKSQY-----FLHL-A	419
WP_052217029.1	hypot	LPBWRVVSQFYKLD-DURS-IRDK-----HFNEKEET-----SKFL-----DMFPWLKE--ELATIDKWR-----SOARLYSV-Y	388
WP_081908191.1	trans	LDNDLHEFNKIK-DLOS-IRDN-----LFNETKAK-----MELL-----KTLELPEAKE--RTSHMANWR-----SOQKMLRL-H	422
AG08270.1	transposa	LPASMLVFSFLID-EIURS-RLQ-----ASNRGIDM-----LQR-----ADIGSAQPFQDLRDFQSISEKRP-----DLAKFCESDF	380
WP_011733919.1	trans	LPDATTDLGLDYVNGDLOG-----RIDSANEN-----HAWL-----LEQWGGDELPSIQLRLSMLRSKRPH-----PAKFAKA-V	500
WP_096876841.1	trans	VPRELIVSKFEAAE-TIQK-----AADPARNEM-----LSWL-----RTFYQNRDEAPQWRESIQGLLRNR-----PSVDAANHLM	384

WP_1165323935.1 hypot LYE[KOAL]DW[EAQ]AGTA---LLM[KGFEHTKPCA-LCGAPV-----ERVTD[L]CCH[AGASAD]NKANSAANLFRDWLGRYAQAABEAKAIIKAKQD 570
 SPE20750.1 transposon VSE[ESAVRWACAK]AGSV---VLD[IAIP-TASTCS-ICGALSDETRDPQSAVQT]ACPH[GARIDRKCN]GAVAWQ[V]WSE[RDAMIERYHLEAAQAMAS 604
 WP_105479500.1 hypot VSE[ESAVRWACAK]AGSV---VLD[IAIP-TASTCS-ICGALSDETRDPQSAVQT]ACPH[GARIDRKCN]GAVAWQ[V]WSE[RDAMIERYHLEAAQAMAS 587
 OJW42488.1 hypotheti VSE[EGAI]RW[AAK]CGTA---VLE[DTGE-TAGHCA-YCGAV-----KPVDDSQRLPACTCGADIDRKCN]GAA[LAQWATESESLPTLVEDFWETLAARDG 594
 WP_018991635.1 hypot VSE[ESAI]RWSVTKQTA---VLE[IVGK-TAGCCA-LCGEKV-----LADVEDSQRLPACTCGADIDRKCN]GAA[LAQWATESESLPTLVEDFWETLAARDG 564
 WP_018079340.1 hypot VSE[ESAI]RW[AAK]AGSA---MFE[DTGETASR-CS-ICGGDV-----LPETNGQI[LHCTEGGADLRKCN]GAAMAQI[LNDLLESLVAFWTFETFAARRS 593
 WP_081130164.1 hypot VSE[ESAI]RW[AAK]AGTA---LDS[GABETARC-GICGAS-----QSDENSQV[LHCTEGGADLRKCN]GAA[LAQWATESESLPTLVEDFWETLAARDG 593
 WP_064217851.1 hypot LXT[LDSAI]RWACQNGTA---ILDNG[EGKTAATCA-MCSEA-----IRATEDQGV[LHCTEGGADLRKCN]GAA[LAQWATESESLPTLVEDFWETLAARDG 570
 WP_051690567.1 trans LVU[ESAI]QW[AC]HGS---VLK[EKGKETSVCFA-FCGDDH---LEEKEDHSQ[V]CP[C]GSTVD[SKL]CAANAKW[FAAS]DLESLVTEYWEETREKQMG 585
 OFC33369.1 hypotheti LHS[LEAL]HW[AC]CGSA---VLH[USGETVTILCS-HCGSTT---ISPTPLNNQMC]CSCGSTIDRKCN]GAANAKW[FAAS]DLESLVTEYWEETREKQMG 575
 WP_077272831.1 trans LST[LRAFL]SESERAGFA---VH[SIPYLSQECH-ICGTRN-----AVASPLM[V]TSCCAQWQ[C]FNNAANL[RAL]QONR[TA]-----FNNAAANL[RAL]QONR[TA]----- 588
 WP_106353755.1 trans LST[LRAFL]SESERAGFA---VH[SIPYLSQECH-ICGTRN-----AVASPLM[V]TSCCAQWQ[C]FNNAANL[RAL]QONR[TA]-----FNNAAANL[RAL]QONR[TA]----- 529
 WP_045077069.1 trans LST[LRAFL]SESERAGFA---VH[SIPYLSQECH-ICGTRN-----AVASPLM[V]TSCCAQWQ[C]FNNAANL[RAL]QONR[TA]-----FNNAAANL[RAL]QONR[TA]----- 529
 WP_102857306.1 hypot LST[LRAFL]SESERAGFA---VH[SIPYLSQECH-ICGTRN-----AVASPLM[V]TSCCAQWQ[C]FNNAANL[RAL]QONR[TA]-----FNNAAANL[RAL]QONR[TA]----- 470
 WP_018234394.1 trans VIAL[RHEI]HOAM[HGAQ---LVH[SGK-TTTTCR-ACGAAT---GQKDRASLIWTEHC]GAVMDQ[L]MAGNILDSEMGASAFAAITLAKAKSRRYDL 580
 WP_061559521.1 hypot PGR[LRVT]TALREGCA---VREGR[GLSR]IHGDCGYEN---PADRYAAAT[QDCCGDYQD]HAT[ALMR]RAGAI[RHGARASTSVGP]----- 607
 WP_064888210.1 hypot PGR[REAV]EAAARAGLR---CEA[SPKGIARIHA-ACGYQN---PGDGRFASLI]TCECG[QYEV]AAS[TLM]RGAGVLS----- 600
 WP_063045032.1 hypot PRT[LRTI]QACV[RGVT---VTV]PATGLSR[THA-RCGHQN---PADRYKAPPV]RCECK[QYD]PSS[TVIMLR]RGR----- 602
 KMW19589.1 hypotheti PGE[LRQTL]VAAD[DAVP---VDT]SH[TVGSV]VHA-KCGHEN---PSDGRFMSV[V]ACDCG[QYD]QES[LTHMLT]RAVOSAA----- 603
 WP_061006603.1 hypot PGG[LRASV]AMT[DGVP---VTV]AAADFT[THS-RCGHVN---PADRYLSNP]RDCG[QYD]QES[LTHMLT]RAVOSAA----- 596
 Jpred_sec_str. HHHHHHHHHHHH--E-EEEE-----EE-----HHHHHHHHHH-----
 WP_095663130.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---EVESRRRRRT[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 618
 WP_073879989.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 618
 WP_064893148.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 618
 KEF95043.1 hypotheti PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 620
 WP_036456351.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 595
 CD091315.1 hypotheti PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 620
 WP_036473531.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 618
 WP_064942980.1 hypot PGL[RSMSV]AAT[RGVP---VTV]P[SAGLS]RIHA-CGCVEN---RVESRRRRRK[V]CAGC[RTYD]PLS[ETVIML]AR[ARPSENP]----- 595
 OOK65169.1 hypotheti PGL[LRQI]IAAAT[DAVP---VTV]P[SAGLS]RIHA-CGCVEN---PAETQK[NGV]T[RA]C[RTYD]PLS[ETVIML]AR[ARPSENP]----- 615
 WP_047323888.1 trans PGL[LRQI]IAAAT[DAVP---VTV]P[SAGLS]RIHA-CGCVEN---PAETQK[NGV]T[RA]C[RTYD]PLS[ETVIML]AR[ARPSENP]----- 620
 WP_101953221.1 hypot PGL[LRQI]IAAAT[DAVP---VTV]P[SAGLS]RIHA-CGCVEN---PAELQ[PRKGV]T[CR]C[RTYD]PLS[ETVIML]AR[ARPSENP]----- 637
 GAB36148.1 hypotheti PGM[LRAL]VAAT[DEVP---TTV]SHTGLSR[VHA-ACGHEN---PADRYLMQ[P]LDCG[RTYD]PLS[ETVIML]AR[ARPSENP]----- 607
 WP_039994403.1 hypot PGM[LRAL]VAAT[DEVP---TTV]SHTGLSR[VHA-ACGHEN---PADRYLMQ[P]LDCG[RTYD]PLS[ETVIML]AR[ARPSENP]----- 597
 KEP41925.1 hypotheti PAN[LRAL]TSAAT[REGVP---VSV]P[AAGLT]RIHA-HCGYQN---PADGRH[AR]P[V]LDCG[SSYD]P[AS]T[LMQV]NAY[PATRTK]----- 603
 WP_036444762.1 hypot PAN[LRAL]TSAAT[REGVP---VSV]P[AAGLT]RIHA-HCGYQN---PADGRH[AR]P[V]LDCG[SSYD]P[AS]T[LMQV]NAY[PATRTK]----- 596
 PZN20932.1 hypotheti PAN[LRAL]TSAAT[REGVP---VSV]P[AAGLT]RIHA-HCGYQN---PADGRH[AR]P[V]LDCG[SSYD]P[AS]T[LMQV]NAY[PATRTK]----- 566
 WP_013159911.1 trans PST[LRAL]VNAFA[ORQKP---VRK]NP[HA]TTT[DCH-ACGAL---VGDPAKEL[V]T]CP[C]EAFYQ[Q]EN[ARNL]RE[QEV]QAQV----- 536
 PZW90038.1 transposon PHV[LRAL]VNAFA[ORQKP---VRK]NP[HA]TTT[DCH-ACGAL---VGDPAKEL[V]T]CP[C]EAFYQ[Q]EN[ARNL]RE[QEV]QAQV----- 531
 WP_092118774.1 trans PAF[ROEL]QNFCK[NTGSL---IF]EKG[SKASTCP-ECKNKI---AKDMA[RLIM]T]CP[C]EAFYQ[Q]EN[ARNL]RE[QEV]QAQV----- 600
 VST_YLAI[EN]CK[FGRT---FANT]PAS----- 508
 WP_0821908191.1 trans ISE[FRNL]AN[ACRN]HVE---FTY[PAEN]TTT[ITCH-KCGHKE-----KFDAA[QI]IHT[CTC]GELMDQ[C]YMA[KNL]AF[SQ]GGVK----- 594
 AGO88270.1 Transposon PSE[LRAL]KLOAE[ERKTA---FNK]EAE[SPVR-CP-TCGSLS-----RKTRADALPOV]CANCDSFDQ[VV]V[CESILS]PAPTARTTSRVKARGAAAT--- 559
 WP_011733919.1 trans ISEL[RCLSK]AAK[NGTQ---TEQ]STA-SSATCS-ACKGKM-----FQV[DGIM]R[OR]C[RA]LVQ[Q]IN[AA]NFREVL----- 664
 WP_096876841.1 trans IYS[KEW]IGKQAAK[GTST---VET]TGK-MTATCH-KCGYVA-----EKBLRG[SYQ]VTK[CSGSE]LE[EN]E[INCR]N[HA]SGAVLISDKPEKTGRFORAKM 571

WP_116532935.1	hypot	AAEAKKKRLALMQAKRAEVRAKAEKNEGESTRCK-----	604
SPE20750.1	transposa	REVNAVARKTKMAAARNAKRQALQEASIAAKETQAGEKAPTCRTGR	650
WP_105479500.1	hypot	REVNAVARKTKMAAARNAKRQALQEASIAAKETQAGEKAPTCRTGR	633
OJW42488.1	hypotheti	AAAKRKEKREKVAEARRASRVVE-----	617
WP_018991635.1	hypot	RSAKKADRLARMTDGRQARGANSSKAP-----	592
WP_018079340.1	hypot	AENEQAQKKQKMAEGRRKARTPIGGENTEVSRDSGNGANA-----	633
WP_081130164.1	hypot	HAECTREKKAKMAEGRRRLARTLSAGVSAVGSRRNV-----	627
WP_064217851.1	hypot	AAEAKASRLEKMQAARRAKREPALAD-----	596
WP_051690567.1	trans	KAETKRLKSEKMAEARRLKRQAASQASAGA-----	615
OFC35369.1	hypotheti	MATLRAQKASGRAQARRASAAAKEKNRAARIAALDAKSEP-----	615
WP_077272831.1	trans	MATLRAQKASGRAQARRASAAAKEKNRAARIAALDAKSEP-----	628
WP_106353755.1	trans	-----	529
WP_045707069.1	trans	-----	529
WP_102857306.1	hypot	-----	470
WP_018234394.1	trans	TQPNFRERSKTGSRASARA-----	599
WP_061559521.1	hypot	-----	607
WP_064888210.1	hypot	-----	600
WP_063045032.1	hypot	-----	602
KMV19589.1	hypotheti	-----	603
WP_061006603.1	hypot	-----	596
Jpred.sec.str.		-----	
WP_095663130.1	hypot	-----	618
WP_073879989.1	hypot	-----	618
WP_064893148.1	hypot	-----	618
KEF95043.1	hypotheti	-----	620
WP_036456351.1	hypot	-----	595
CD091315.1	hypotheti	-----	620
WP_036473531.1	hypot	-----	618
WP_064942980.1	hypot	-----	595
OOK65169.1	hypotheti	-----	615
WP_047323888.1	trans	-----	620
WP_101953221.1	hypot	-----	637
GAB36148.1	hypotheti	-----	607
WP_039994403.1	hypot	-----	597
KEP41925.1	hypotheti	-----	603
WP_036444762.1	hypot	-----	596
PZN20932.1	hypotheti	RSERLRRGRRKAA-----	579
WP_013159911.1	trans	-----	536
PZM90038.1	transposa	-----	531
WP_092118774.1	trans	-----	600
WP_052217029.1	hypot	-----	508
WP_081908191.1	trans	-----	594
AGO88270.1	Transposa	-----	559
WP_011733919.1	trans	-----	664
WP_096876841.1	trans	AENDFARKIGDNASPLVT-----	589

Figure S2 | Multiple sequence alignment of type V-U1 orthologues. MmuCas12u1 is indicated as WP_061006603.1.

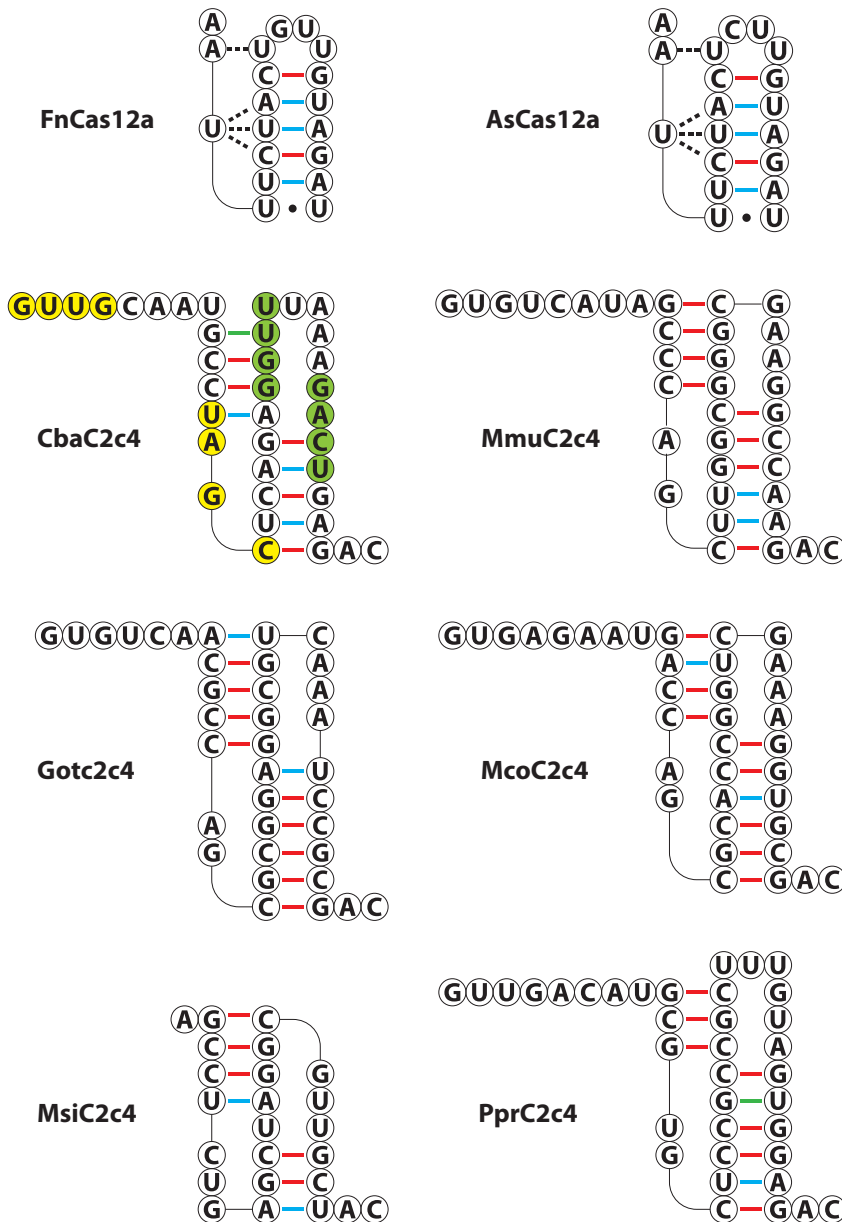


Figure S3 | Type V-U1 repeats from different bacteria. FnCas12a and AsCas12a crRNA structures are based on Xtal structures. Putative Cas12u1 pseudoknot structures in the CRISPR RNA repeat regions as predicted by Vsfold (261) except in case of CbaCas12u1 the predicted base pairing deviates from the structure shown here. FnCas12a: *Francisella tularensis* subsp. *novicida* U112 Cas12a; AsCas12a: *Acidaminococcus* sp. BV3L6 Cas12a; CbaCas12u1: *Clostridiales bacterium* DRI 13 Cas12u1; MmuCas12u1: *Mycobacterium mucogenicum* CCH10-A2 Cas12u1; Mcocas12u1: *Mycobacterium conceptionense* MLE Cas12u1; MsiCas12u1: *Meiothermus silvanus* DSM 9946 Cas12u1.

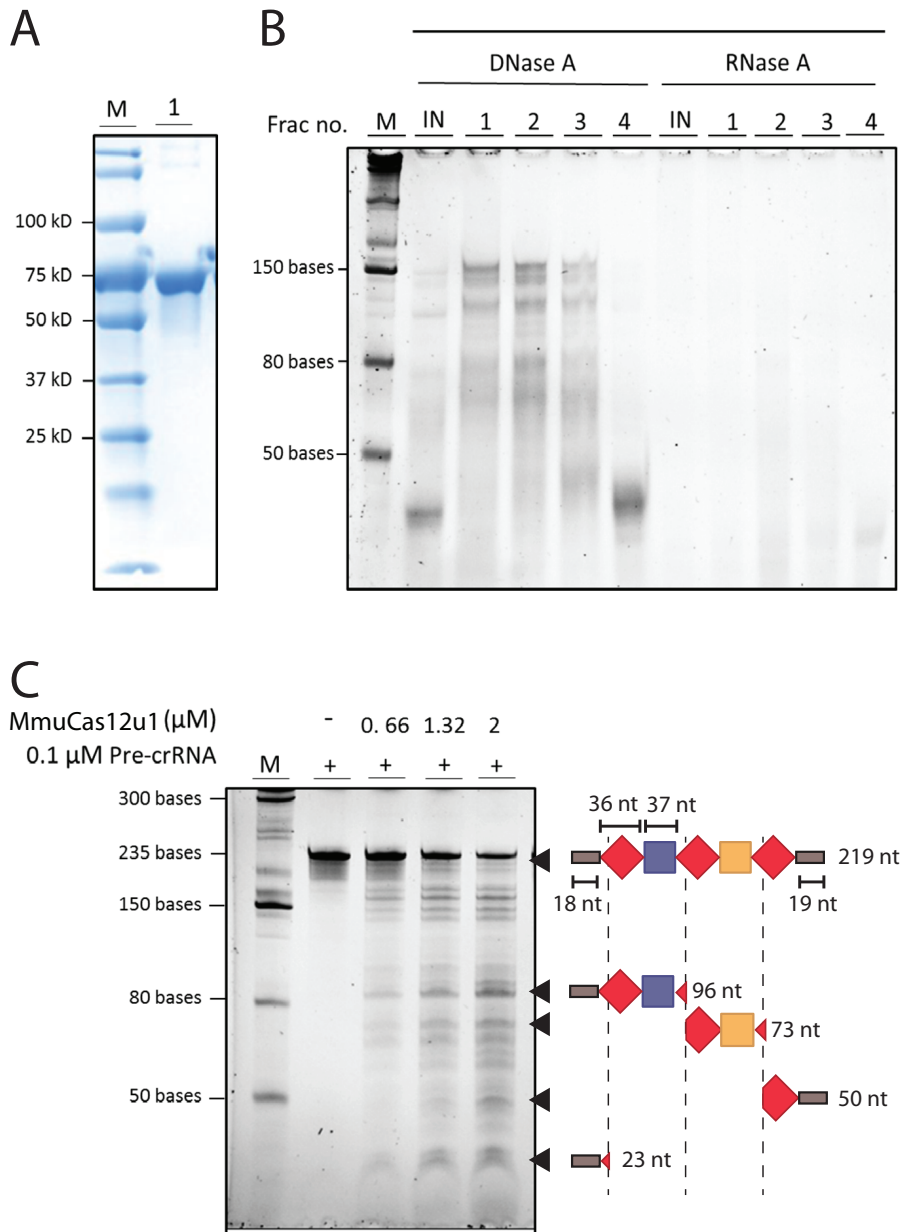


Figure S4 | Co-purified nucleic acids and pre-crRNA processing by MmuCas12u1. (A) Coomassie blue stained SDS-PAGE gel in which the purified MmuCas12u1 protein (66.2 kD) is visualized. (B) Co-purified nucleic acids from MmuCas12u1 treated with enzymes as indicated. M: low range ssRNA ladder (NEB), IN: input fraction for Size Exclusion Chromatography (SEC) 1-4: different fractions from the SEC purification. (C) 10% Urea-PAGE gel on which the processed pre-crRNA transcripts were resolved. RNA was visualized after staining with SYBR-gold. M: low range ssRNA ladder (NEB).

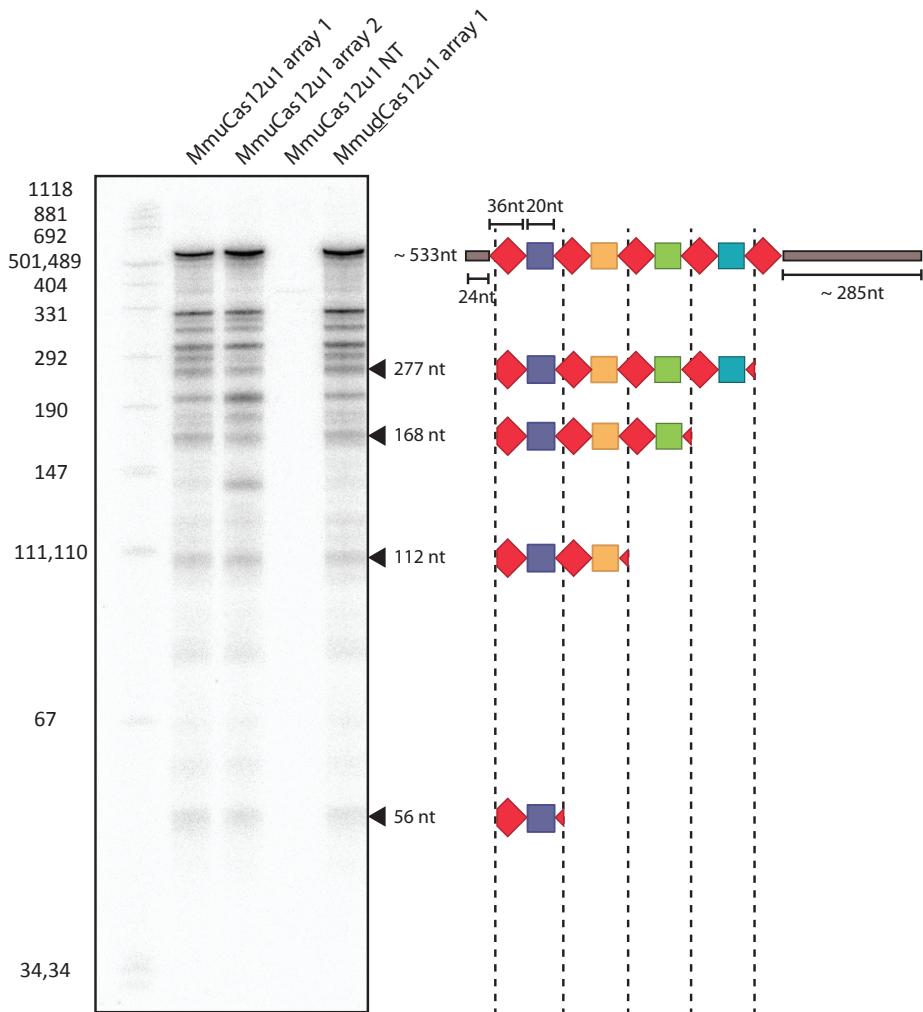


Figure S5 | Processing by MmuCas12u1 and MmudCas12u1 by northern blot analysis in TXTL
 A plasmid expressing the MmuCas12u1 CRISPR array containing four spacers were incubated with a plasmid expressing MmuCas12u1 or MmudCas12u1 in TXTL. RNA was visualized by northern blot, using a probe that binds to the first spacer of array A (purple). Array B is similar to that of array A, only the order of spacers is shifted (yellow-green-blue-purple).

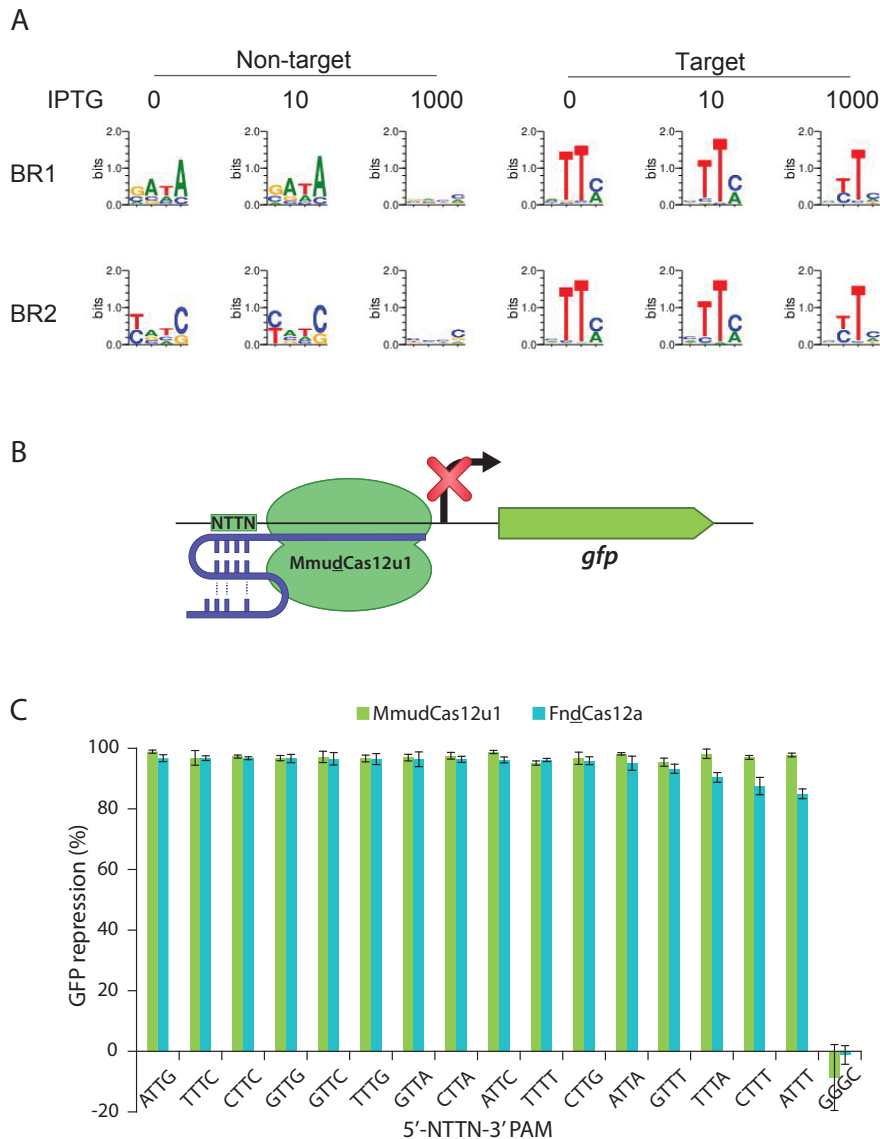


Figure S6 | MmuCas12u1 PAM determination. (A) Deep sequencing analysis of PAM-SCNR after FACS sorting. (A) Plasmids from the FACS-sorted cells were extracted and sequenced to determine functional PAM sequences. Sequence logo for the MmuCas12u1 PAM at different IPTG concentrations (0, 10, 1000 μ M) as determined by NGS sequencing. NT: non-targeting, T: targeting, BR1 and BR2 are two independent biological replicates. Letter height at each position is measured by information content. **(B)** Schematic of the pTarget-GFP encoding the *gfp* gene. The protospacer flanked by 5'-NTTN-3' PAM upstream of the promoter is targeted by the MmuCas12u1 and FndCas12a proteins using the respective crRNAs. **(C)** GFP repression detected in the cells upon MmuCas12u1 and FndCas12a targeting is shown on the Y-axis and the different PAM sequences used are shown in the X-axis ($n = 3$; error bars represent mean \pm SD).

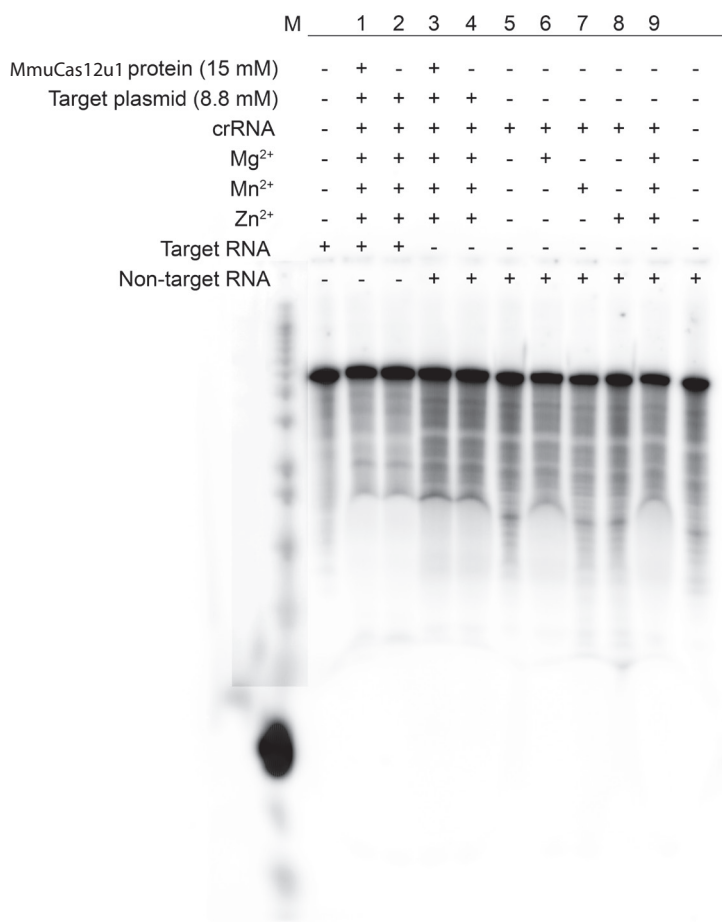


Figure S7 | In *vitro* dsDNA activated RNA cleavage by MmuCas12u1. Urea-PAGE assessing the ability of MmuCas12u1 protein incubated with a crRNA and an activator target DNA to cleave a [γ -³²P] ATP labelled target or a non-target substrate RNA.

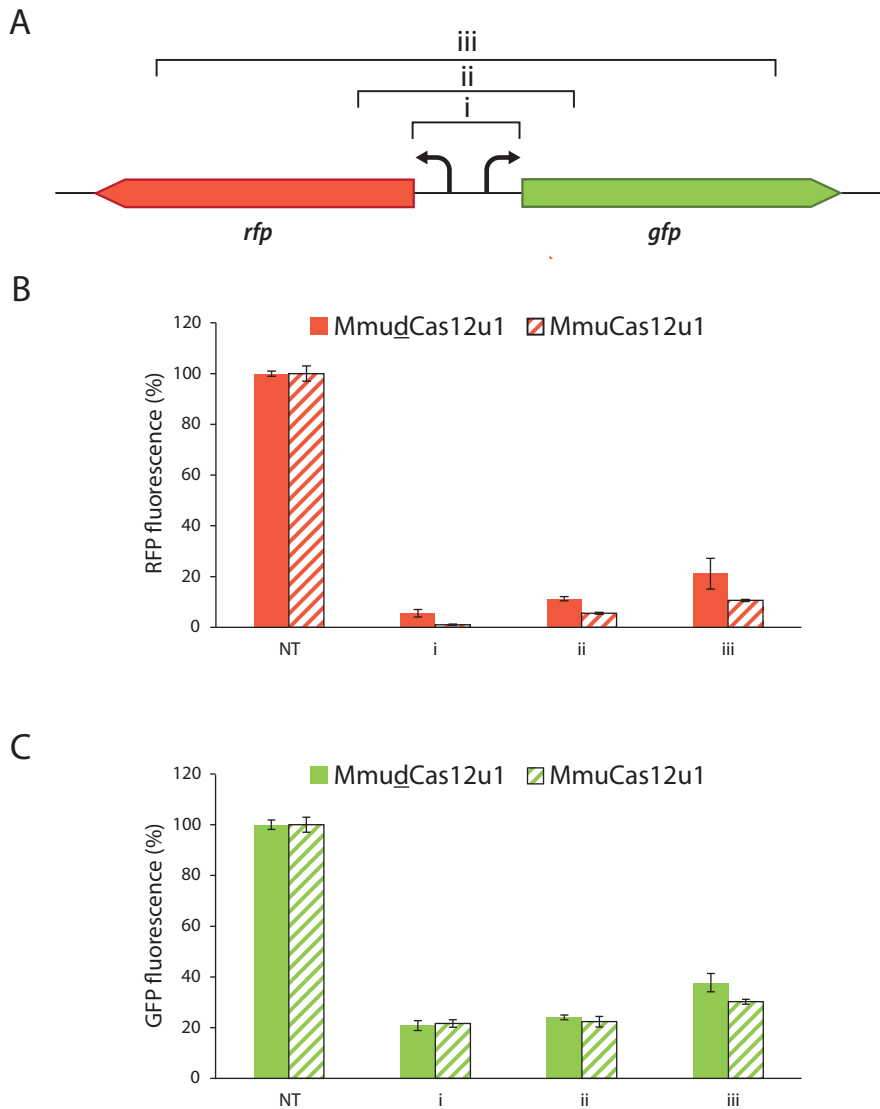


Figure S8 | MmuCas12u1 can be used for multiplex transcriptional silencing. (A) Schematic of the pTarget-divergent including the *rfp* and *gfp* genes under the transcriptional control of two different constitutive promoters, P_{tag} and P_{lacIq} . i to iii indicate the crRNA spacer pairs used in the pCRISPR array plasmid to target the *gfp* and *rfp* using the MmuCas12u1 and MmudCas12u1 proteins. (B) RFP fluorescence detected in the cells upon MmudCas12u1 or MmuCas12u1 targeting using the crRNA spacer pairs is shown on the Y-axis and the different mismatches are shown on the X-axis ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer (C) GFP fluorescence detected in the cells upon MmudCas12u1 or MmuCas12u1 targeting using the crRNA spacer pairs is shown on the Y-axis and the different mismatches are shown on the X-axis ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer

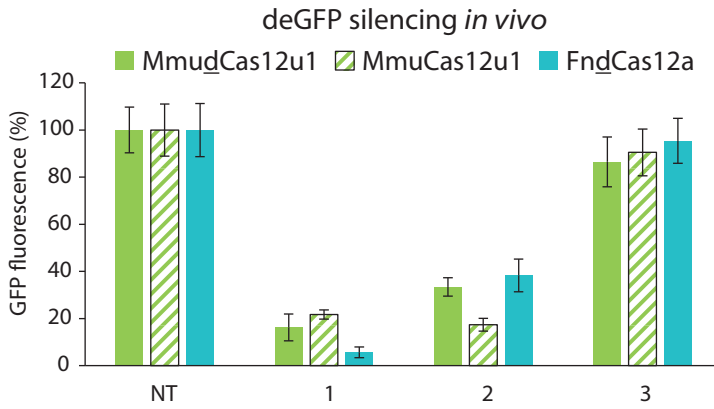


Figure S9 | deGFP silencing by MmuCas12u1, MmudCas12u1 and FndCas12a. GFP fluorescence detected in cells upon MmudCas12u1 and MmuCas12u1 targeting using the individual spacers ($n = 3$; error bars represent mean \pm SD). NT refers to a non-targeting spacer.

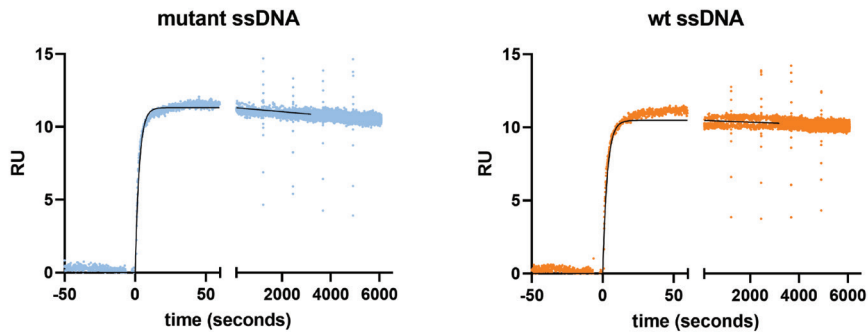


Figure S10 | SPR-based Sensorgram of binding of 500 nM RuvC-mutant (MmudCas12u1) and wild type MmuCas12u1-RNA complex to ssDNA.

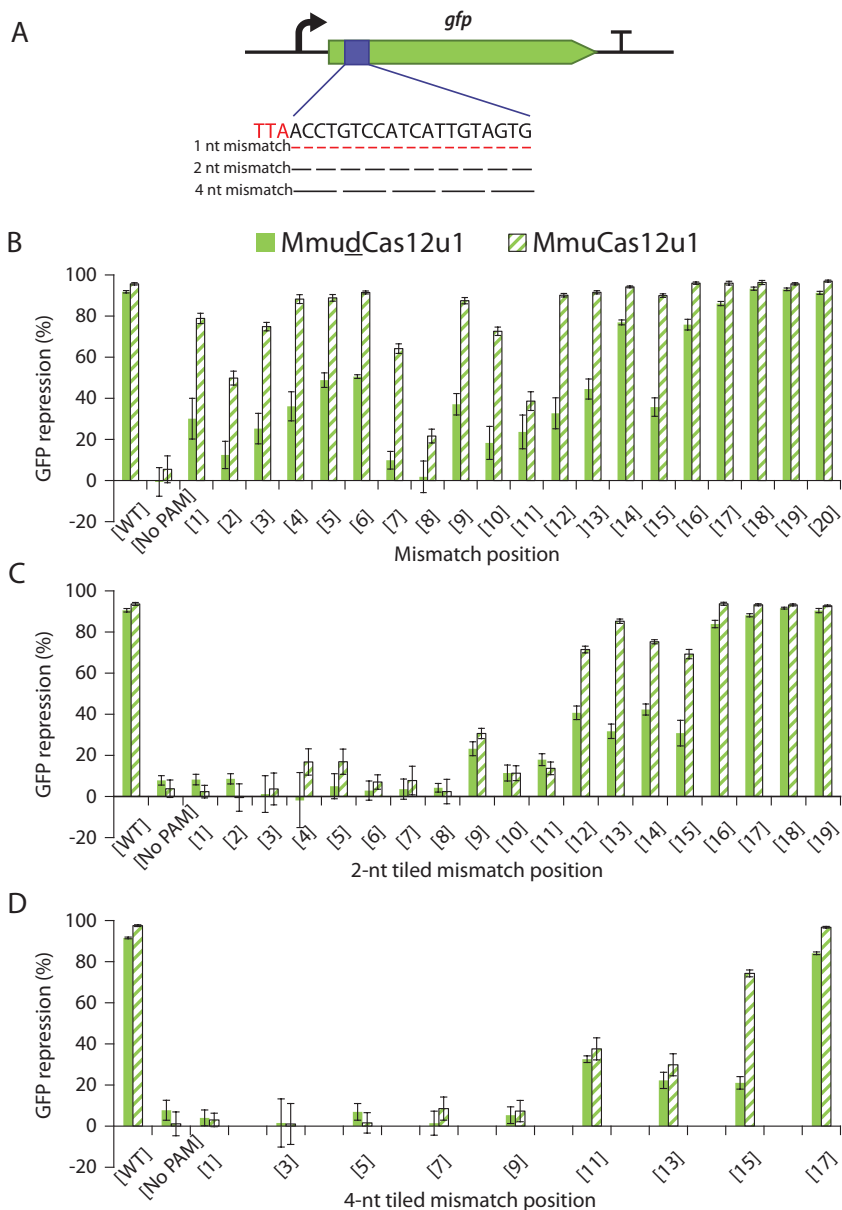


Figure S11 | Tolerance of MmuCas12u1 to mismatched crRNAs. (A) Schematic *gfp* silencing to assess mismatch tolerance. Mismatches are tiled through the protospacer in one, two or four nucleotides. **(B)** Comparison of the mismatch tolerance of MmuCas12u1 with MmuCas12u1 for single mismatches across the protospacer sequence. **(C)** Comparison of the mismatch tolerance of MmuCas12u1 with MmuCas12u1 for 2-nucleotide mismatches tiled across in the target sequence. **(D)** Comparison of the mismatch tolerance of MmuCas12u1 with MmuCas12u1 for 4-nucleotide mismatches tiled across in the target sequence. GFP repression detected in the cells upon MmuCas12u1 or MmuCas12u1 targeting is shown on the Y-axis and the different mismatches are shown on the X-axis ($n = 3$; error bars represent mean \pm SD). No PAM refers to a spacer targeting protospacer next to GGCG motif (non-functional PAM).

Table S1 | oligonucleotides used in this study

oligo ID	sequence (5'-3')	description
Construction of the pMmuCas12u1 plasmids		
BG14064	GATGTCCTCCTGAGCTCGC	FW for amplification of the plasmid backbone for construction of the pMmuCas12u1 and pMmudCas12u1
BG14065	AAGCTTGGCTGTGTTTGGCG	FW for amplification of the plasmid backbone for construction of the pMmuCas12u1 and pMmudCas12u1
BG14070	GGAGCTCAGGAGGACATCATGCAACAATGACATACATAATGG	FW for amplification of the mmuCas12u1 and mmudCas12u1 gene for construction of the pMmuCas12u1 and pMmudCas12u1
BG14073	CGCCAAAACAGCCAAGCTTCTAGGGGTTTCGAGGGGGC	RV for amplification of the mmuCas12u1 gene for construction of the pMmuCas12u1 and pMmudCas12u1
BG14402	GATAATTCTGCTACCGATGTATCTGCAACTAAAACATGTCTCTGCTGATC	RV for amplification of the mmudCas12u1 gene for construction of the pMmuCas12u1
BG14403	CAGGCAGGACATGTTTTAGTTGCAGATACATCGGTAGCAGAAATTATCGGC	FW for amplification of the mmudCas12u1 gene for construction of the pMmudCas12u1
BG20338	GATGCATCTGACAGTAGCTCAGTC	FW for amplification of Mmu(d)Cas12u1 H549A C552A fragment 1
BG20345	GTCCGGCGCGGAACAGCTCCGAGTAAAGTCTGCGGC	RV for amplification of Mmu(d)Cas12u1 H549A C552A fragment 1
BG20352	GACTGCTTCGCGCGCGGACATGTTAAATCCCGCCGATG	FW for amplification of Mmu(d)Cas12u1 H549A C552A fragment 2
BG20353	CTTCTGCGGTTCTGATTTTAAATCTGTATCAGG	RV for amplification of Mmu(d)Cas12u1 H549A C552A fragment 2
Construction of MmuCas12u1 pCRISPR plasmids		
BG14103	GGAACTCGAGGTGGTACCG	FW for amplification of the vector for the construction of the pCRISPR
BG14158	GATCGAAGACTAGTGTGCATAGCCAGCTTGGCGGGCGAAGCCAGAGCGTTTGGCGGA TCAGAGAAG	FW for amplification of the vector for the construction of the pCRISPR
BG14086	ACACTGCCATACCGCGAAAGGTTTTGCACTCGACGCTCTTGGCCTTCGCCCGCCCAAGCTG GGCTATGACACGGTAC	FW oligo for pCRISPR-PS
BG14087	CGTTTCATCTGGCCATCGCGGGCGGCTCGTAGTGGACGTGCGAAGCAAAACCTTTTCGC GGTATGGCA	RV oligo for pCRISPR-PS
BG15637	CGTGTACATAGCCACGTTTGGGGCGAAGGCCAAGACTGGTCTTCGCATCTTGGCCGTTA GAAGACAA	FW oligo for introducing BbsI sites to the pCRISPR, creating pCRISPR-NT plasmid
BG15638	ACACTTGTCTTCTAACGGCAAGATCGGAAGACCAGTCTTGGCCTTCGCCCGCCCAAGCTG GGCTATGACACGGTAC	RV oligo for introducing BbsI sites to the pCRISPR, creating pCRISPR-NT plasmid
BG15106	CGTGTACATAGCCACGTTTGGGGCGAAGGCCAAGACTCGAGTGCAGTCAAAACCTTTTCG	FW oligo for construction of pCRISPR-A1 and pCRISPR-d

BG15107	ACACGAAAGGTTTTCACATCGACGCTCTGGCCTTCGCCCGCCCAAGCTGGGCTATGACA CGGTAC	RV oligo for construction of pCRISPR-A1 and pCRISPR-d
BG17087	AGACGCTATCATGCCATACCGCGA	FW oligo for construction of pCRISPR-No PAM
BG17088	ACACTCGCGGTATGCATGATAGC	RV oligo for construction of pCRISPR-No PAM
BG16559	AGACACTCTCTTCCGGCGCTATC	FW oligo for construction of pCRISPR-A2
BG16560	ACACGATAGCGCCCGAAGAGAGT	FW oligo for construction of pCRISPR-A2
BG16299	AGACCAAAACCGACATCAAACTGG	FW oligo for construction of pCRISPR-B1 and pCRISPR-a
BG16300	ACACCCAGTTTGATGTCGGTTTGG	RV oligo for construction of pCRISPR-B1 and pCRISPR-a
BG16301	AGACGTTGTGGGAGGTGATGTCCA	FW oligo for construction of pCRISPR-B2
BG16302	ACACTGGACATCACCTCCCAAC	RV oligo for construction of pCRISPR-B2
BG16303	AGACACCTCTAGATTAAAGAAGGA	FW oligo for construction of pCRISPR-C1
BG16304	ACACTCCTTCTTAAATCTAGAGGT	RV oligo for construction of pCRISPR-C1
BG16305	AGACAATCTAGAGGTAAACAAAA	FW oligo for construction of pCRISPR-C2
BG16306	ACACTTTTGTTTAACTCTAGATT	RV oligo for construction of pCRISPR-C2
BG16858	AGACCTGTCCACACAATCTGCC	FW oligo for construction of pCRISPR-D1 and pCRISPR-f
BG16859	ACACGGGCAGATTGTGGACAGG	RV oligo for construction of pCRISPR-D1 and pCRISPR-f
BG16096	AGACGAAAGGGCAGATTGTGTGGA	FW oligo for construction of pCRISPR-D2
BG16097	ACACTCCACACAATCTGCCCTTTC	RV oligo for construction of pCRISPR-D2
BG16886	AGACGTTTTATCTGTGTTGTGTCG	FW oligo for construction of pCRISPR-E1
BG16887	ACACCGACAACAACAGATAAAAC	RV oligo for construction of pCRISPR-E1
BG16563	AGACTCCTTACTCAGGAGACGTTTC	FW oligo for construction of pCRISPR-E2
BG16564	ACACGAACGCTCTCCTGAGTAGGA	RV oligo for construction of pCRISPR-E2
BG16385	AGACTGGGTGTTGCTAGTTTGTAT	FW oligo for construction of pCRISPR-F1
BG16386	ACACATAACAACACTAGCAACACA	RV oligo for construction of pCRISPR-F1
BG16561	AGACTGATAACAACCTAGCAACAC	FW oligo for construction of pCRISPR-F2

BG16562	ACACGTGTTGCTAGTTTGTGTATCA	RV oligo for construction of pCRISPR-F2
BG16860	AGACGTATGGAAGTTCGTTAAAC	FW oligo for construction of pCRISPR-b
BG16861	ACACGTTAAACGGAACTTCCATAC	RV oligo for construction of pCRISPR-b
BG16888	AGACAAGTTGACAA'TTAATCATCG	FW oligo for construction of pCRISPR-c
BG16889	ACACCGATGATTAA'TTGTCAACTT	RV oligo for construction of pCRISPR-c
BG17641	AGACATGGGCACAAATTTTCTGTC	FW oligo for construction of pCRISPR-e
BG17642	ACACACAGAAAATTTTGCCCAT	RV oligo for construction of pCRISPR-e
BG16890	AGACAAGTTGACAA'TTAATCATCGGTGTCATAGCCAGCTTGGCGGGCGAAGGCCAAG ACGTGAGTGC AAAACCTTTTCG	FW oligo for construction of pCRISPR-i
BG16891	ACACCGAAAAGTTTTCGCACTCGACGCTCTTGGCCTTCGCCCGCCAAAGCTGGGCTATGAC ACCGATGATTAATTTGTCAACTT	RV oligo for construction of pCRISPR-i
BG18661	AGACGTATGGAAGTTCGTTAACTGTCTATAGCCAGCTTGGCGGGCGAAGGCCAAG ACTGTCACTGGAGAGGGTGAAG	FW oligo for construction of pCRISPR-ii
BG18662	ACACCTTCACTCCACTGCACAGTCTTGGCCTTCGCCCGCCAAAGCTGGGCTATGAC ACGTTAACGGAACTTCCATAC	RV oligo for construction of pCRISPR-ii
BG16894	AGACAAAACCGACATCAAACTGGGTGTCATAGCCAGCTTGGCGGGCGAAGGCCAAG ACCCGTGCCACAAATCTGCCC	FW oligo for construction of pCRISPR-iii
BG16895	ACACGGGCGAGTTGTGTGGACAGGGTCTTGGCCTTCGCCCGCCAAAGCTGGGCTATGAC ACCCAGTTTGATGTCGGTTTTCG	RV oligo for construction of pCRISPR-iii
BG16520	AGACACCACTCCATCATTTGTAGTG	FW for construction of pCRISPR-MM
BG16521	ACACCACTCAATGATGAGACTGGT	RV for construction of pCRISPR-MM
BG18829	AGACACAATTTTACCTCTGGGGGT	FW for construction of pCRISPR-1
BG18830	ACACACCCGACAGGTAATAATTGT	RV for construction of pCRISPR-1
BG18831	AGACGTGACCGCGCCGGGATCTA	FW for construction of pCRISPR-2
BG18832	ACACTAGATCCCGCGCGGTACAC	RV for construction of pCRISPR-2
BG18833	AGACAACCCAGTCACCTCCTCCG	FW for construction of pCRISPR-3
BG18834	ACACCGGAAGGAGCTGACTGGGTT	RV for construction of pCRISPR-3
BG20645	AGACAAGTTGACAA'TTAATCATCGGT	FW MmuCas12u1 array APiAd spacer 1

BG20646	TATGACACCGATGATTAATTTGTCAACTT	RV MmuCas12u1 array A P1acl spacer 1
BG20655	GTCA TAGCC CAGCTTGGCGGGCAAGGCCAAGCTCGAGTGC AAAACCTTTCGGTGT	FW MmuCas12u1 array A P1acl spacer 2
BG20656	GCTATGACACCCAGAAAGGTTTGCACCTCAGCTCTGGGCTTGGCCGCGCAAGCTGGGC	RV MmuCas12u1 array A P1acl spacer 2
BG20665	CATAGCCCAAGCTTGGCGGGCAAGGCCAAGACACAATTTTACCTCTGGCGGTGTCTCA	FW MmuCas12u1 array A P70A spacer 3
BG20666	GGGCTATGACACACCGCCAGAGGTAAATTTGTTGGCTTTCGCCCTTCGCCCGCAAGCTGG	RV MmuCas12u1 array A P70A spacer 3
BG20675	TAGCCCAAGCTTGGCGGGCAAGGCCAAGACCTTTACACTTTATAGTTTCCG	FW MmuCas12u1 array A NT spacer 4
BG20676	ACACCGGAAGACATAAAGTGTAAGGTCTTGGCTTTCGCCCGCGCAAGCT	RV MmuCas12u1 array A NT spacer 4
BG20663	AGACGTGAGTGCAAAACCTTTTCGGT	FW MmuCas12u1 array B P1acl spacer 1
BG20654	TATGACACCGAAAGGTTTTTGACCTCGAC	RV MmuCas12u1 array B P1acl spacer 1
BG20663	GTCA TAGCC CAGCTTGGCGGGCAAGGCCAAGACACACAATTTTACCTCTGGCGGTGTGT	FW MmuCas12u1 array B P70A spacer 2
BG20664	GCTATGACACACCGCCAGAGGTAAATTTGTGCTTGGGCTTGGCCGCGCAAGCTGGGC	RV MmuCas12u1 array B P70A spacer 2
BG20673	CATAGCCCAAGCTTGGCGGGCAAGGCCAAGACCTTTACACTTTATGCTTTCGGGTGTCA	FW MmuCas12u1 array B NT spacer 3
BG20674	GGGCTATGACACCGGAAGCATAAAGTGTAAAGGTCTTGGCCTTCGCCCGCGCAAGCTGG	RV MmuCas12u1 array B NT spacer 3
BG20651	TAGCCCAAGCTTGGCGGGCAAGGCCAAGACACAAGTTGACAAATTAATCATCG	FW MmuCas12u1 array B P1acl spacer 4
BG20652	ACACCGATGATTAA TTGTC AACTTGTCTTGGCTTTCGCCCGCGCAAGCT	RV MmuCas12u1 array B P1acl spacer 4
Construction of Cas12a pCRISPR plasmids		
BG19471	AGATACAATTTTACCTCTGGCGGT	FW for construction of pCRISPR-1
BG19472	AGACACCGCCAGAGTAAATGT	RV for construction of pCRISPR-1
BG19473	AGATGTGACCGCGCGCGGATCTA	FW for construction of pCRISPR-2
BG19474	AGACTAGATCCCGCGCGGTAC	RV for construction of pCRISPR-2
BG19475	AGATCAACCCAGTCACTCTTCGG	FW for construction of pCRISPR-3
BG19476	AGACCGGAAGGAGCTGACTGGGT	RV for construction of pCRISPR-3
Construction of pTarget plasmids		
BG15568	ATACTCGGATCCCCGTAATTGACTCTCTTC	FW for construction of pTarget-GFP
BG15569	GGGATCCCTCTAGATTAAAG	RV for construction of pTarget-GFP

BG17549	CTTC TT TAG TCGAG TCGCAA AACCTTTCGCG	FW for construction of pTarget-GFP with a TTTA PAM
BG16843	ACGAAAGGCCCTCGACGC	RV for construction of pTarget-GFP with different PAMs
BG17550	CTTCGGGCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a GGCC PAM
BG16844	CTTCGTTAGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a GTTA PAM
BG16845	CTTCATTAGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a ATTA PAM
BG16846	CTTCCTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a CTTC PAM
BG16847	CTTC TTTT CGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a TTTT PAM
BG16848	CTTCGTTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a GTTC PAM
BG16849	CTTCATTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a ATTC PAM
BG16850	CTTCCTTTGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a CTTT PAM
BG16851	CTTC TTTT TCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a TTTT PAM
BG16852	CTTCGTTTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a GTTT PAM
BG16853	CTTCATTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a ATTT PAM
BG16854	CTTCCTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a CTTG PAM
BG16855	CTTC TTTT GTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a TTTG PAM
BG16856	CTTCGTTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a GTTG PAM
BG16857	CTTCATTTCGTCGAGTGC AAAACCTTTCGCG	FW for construction of pTarget-GFP with a ATTG PAM
BG16134	AGAGTCAATTCAGGGGAGACCAACGGTTTCCC	FW for construction of the pTarget-operon
BG16135	TTCTTAAATCTAGAGGTTAAACAAAATTTATTTCTAGTTTAAGCACCGG	RV for construction of the pTarget-operon
Construction of pTarget plasmids for mismatch tolerance assays		
BG16430	TATGTTTAAACCAGTCCATCATGTAGTG	FW for construction of pTarget-MM-[W/T]
BG16431	TACTCACTACAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[W/T]
BG16432	TATGAAATACACAGTCCATCATGTAGTG	FW for construction of pTarget-MM-[No PAM]
BG16433	TACTCACTACAATGATGGACTGGTATTT	RV for construction of pTarget-MM-[No PAM]
BG16434	TATGTTTATCCAGTCCATCATGTAGTG	FW for construction of pTarget-MM-[T]

BG16435	TACTCACTACAAATGATGCACTGGATAAA	RV for construction of pTarget-MM-[1]
BG16436	TATGTTTAAAGCAGTCCATCATTTGTAGTG	FW for construction of pTarget-MM-[2]
BG16437	TACTCACTACAAATGATGCACTGCTTAAA	RV for construction of pTarget-MM-[2]
BG16438	TATGTTTAAACGAGTCCATCATTTGTAGTG	FW for construction of pTarget-MM-[3]
BG16439	TACTCACTACAAATGATGCACTGGTTAAA	RV for construction of pTarget-MM-[3]
BG16440	TATGTTTAAACCTGTCATCATTTGTAGTG	FW for construction of pTarget-MM-[4]
BG16441	TACTCACTACAAATGATGCACTGGTTAAA	RV for construction of pTarget-MM-[4]
BG16442	TATGTTTAAACCACTCCATCATTTGTAGTG	FW for construction of pTarget-MM-[5]
BG16443	TACTCACTACAAATGATGAGTGGTTAAA	RV for construction of pTarget-MM-[5]
BG16444	TATGTTTAAACCAAGACCATCATTTGTAGTG	FW for construction of pTarget-MM-[6]
BG16445	TACTCACTACAAATGATGCTGCTGGTTAAA	RV for construction of pTarget-MM-[6]
BG16446	TATGTTTAAACCACTGCATCATTTGTAGTG	FW for construction of pTarget-MM-[7]
BG16447	TACTCACTACAAATGATGCACTGGTTAAA	RV for construction of pTarget-MM-[7]
BG16448	TATGTTTAAACCACTGCATCATTTGTAGTG	FW for construction of pTarget-MM-[8]
BG16449	TACTCACTACAAATGATGCACTGGTTAAA	RV for construction of pTarget-MM-[8]
BG16450	TATGTTTAAACCACTCCTTCATTTGTAGTG	FW for construction of pTarget-MM-[9]
BG16451	TACTCACTACAAATGAAGCACTGGTTAAA	RV for construction of pTarget-MM-[9]
BG16452	TATGTTTAAACCACTCCAACATTTGTAGTG	FW for construction of pTarget-MM-[10]
BG16453	TACTCACTACAAATGATGCACTGGTTAAA	RV for construction of pTarget-MM-[10]
BG16454	TATGTTTAAACCACTCCATGATTTGTAGTG	FW for construction of pTarget-MM-[11]
BG16455	TACTCACTACAAATCATGCACTGGTTAAA	RV for construction of pTarget-MM-[11]
BG16456	TATGTTTAAACCACTCCATCATTTGTAGTG	FW for construction of pTarget-MM-[12]
BG16457	TACTCACTACAAAGATGCACTGGTTAAA	RV for construction of pTarget-MM-[12]
BG16458	TATGTTTAAACCACTCCATCATTTGTAGTG	FW for construction of pTarget-MM-[13]
BG16459	TACTCACTACATTTGATGCACTGGTTAAA	RV for construction of pTarget-MM-[13]

BG16460	TATGTTTAAACCACTCCATCATAGTAGTG	FW for construction of pTarget-MM-[14]
BG16461	TACTCACTACTATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[14]
BG16462	TATGTTTAAACCACTCCATCATCTAGTG	FW for construction of pTarget-MM-[15]
BG16463	TACTCACTAGAAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[15]
BG16464	TATGTTTAAACCACTCCATCATTTGAAGTG	FW for construction of pTarget-MM-[16]
BG16465	TACTCACTTCAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[16]
BG16466	TATGTTTAAACCACTCCATCATTTGTTGTG	FW for construction of pTarget-MM-[17]
BG16467	TACTCACAAACCAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[17]
BG16468	TATGTTTAAACCACTCCATCATTTGTACTG	FW for construction of pTarget-MM-[18]
BG16469	TACTCAGTACAAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[18]
BG16470	TATGTTTAAACCACTCCATCATTTGTAGAG	FW for construction of pTarget-MM-[19]
BG16471	TACTCTCTACAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[19]
BG16472	TATGTTTAAACCACTCCATCATTTGTAGTC	FW for construction of pTarget-MM-[20]
BG16473	TACTGACTACAATGATGGACTGGTTAAA	RV for construction of pTarget-MM-[20]
BG17030	TATGTTTATGCACTCCATCATTTGTAGTG	FW for construction of pTarget-2MM-[1]
BG17031	TACTCACTACAATGATGGACTGCATAAA	RV for construction of pTarget-2MM-[1]
BG17032	TATGTTTAAAGGAGTCCATCATTTGTAGTG	FW for construction of pTarget-2MM-[2]
BG17033	TACTCACTACAATGATGGACTCCCTTTAAA	RV for construction of pTarget-2MM-[2]
BG17034	TATGTTTAAACGTTGTCATCATTTGTAGTG	FW for construction of pTarget-2MM-[3]
BG17035	TACTCACTACAATGATGGACACAGTTAAA	RV for construction of pTarget-2MM-[3]
BG17036	TATGTTTAAACCTCTCCATCATTTGTAGTG	FW for construction of pTarget-2MM-[4]
BG17037	TACTCACTACAATGATGGAGAGGTTAAA	RV for construction of pTarget-2MM-[4]
BG17038	TATGTTTAAACCAACCATCATTTGTAGTG	FW for construction of pTarget-2MM-[5]
BG17039	TACTCACTACAATGATGTTGTTGTTAAA	RV for construction of pTarget-2MM-[5]
BG17040	TATGTTTAAACCAAGCATCATTTGTAGTG	FW for construction of pTarget-2MM-[6]

BG17041	TACTCACTACAAATGATGCTCTGGTTTAAA	RV for construction of pTarget-2MM-[6]
BG17042	TATGTTTAAACCAAGTGGATCATTTGTAGTG	FW for construction of pTarget-2MM-[7]
BG17043	TACTCACTACAAATGATCCACTGGTTAAA	RV for construction of pTarget-2MM-[7]
BG17044	TATGTTTAAACCAAGTGGTTCAATTTGTAGTG	FW for construction of pTarget-2MM-[8]
BG17045	TACTCACTACAAATGAACGACTGGTTAAA	RV for construction of pTarget-2MM-[8]
BG17046	TATGTTTAAACCAAGTCCCTACATTTGTAGTG	FW for construction of pTarget-2MM-[9]
BG17047	TACTCACTACAAATGATGAGACTGGTTAAA	RV for construction of pTarget-2MM-[9]
BG17048	TATGTTTAAACCAAGTCCAAGATTTGTAGTG	FW for construction of pTarget-2MM-[10]
BG17049	TACTCACTACAAATCTTGGACTGGTTAAA	RV for construction of pTarget-2MM-[10]
BG17050	TATGTTTAAACCAAGTCCATGTTTGTAGTG	FW for construction of pTarget-2MM-[11]
BG17051	TACTCACTACAAACATGGACTGGTTAAA	RV for construction of pTarget-2MM-[11]
BG17052	TATGTTTAAACCAAGTCCATCTATTTGTAGTG	FW for construction of pTarget-2MM-[12]
BG17053	TACTCACTACATAGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[12]
BG17054	TATGTTTAAACCAAGTCCATCAAAAGTAGTG	FW for construction of pTarget-2MM-[13]
BG17055	TACTCACTACTTTTGTATGGACTGGTTAAA	RV for construction of pTarget-2MM-[13]
BG17056	TATGTTTAAACCAAGTCCATCATCTACTAGTG	FW for construction of pTarget-2MM-[14]
BG17057	TACTCACTAGTATGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[14]
BG17058	TATGTTTAAACCAAGTCCATCATTCAGTG	FW for construction of pTarget-2MM-[15]
BG17059	TACTCACTTGAATGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[15]
BG17060	TATGTTTAAACCAAGTCCATCATTTGATGTG	FW for construction of pTarget-2MM-[16]
BG17061	TACTCACATCAATGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[16]
BG17062	TATGTTTAAACCAAGTCCATCATTTCTCTG	FW for construction of pTarget-2MM-[17]
BG17063	TACTCAGAACCAATGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[17]
BG17064	TATGTTTAAACCAAGTCCATCATTTGTACAG	FW for construction of pTarget-2MM-[18]
BG17065	TACTCTGTACAATGATGGACTGGTTAAA	RV for construction of pTarget-2MM-[18]

BG17066	TATGTTTAAACCAAGTCCATCATTTGTAGAC	FW for construction of pTarget-2MM-[19]
BG17067	TACTGTCTACAATGATGACTGGTTAAA	RV for construction of pTarget-2MM-[19]
BG17068	TATGTTTATGTGTCCATCATTTGTAGTG	RV for construction of pTarget-4MM-[1]
BG17069	TACTCACTACAATGATGGACACCAATAAA	FW for construction of pTarget-4MM-[1]
BG17070	TATGTTTAAAGTCACCATCATTTGTAGTG	RV for construction of pTarget-4MM-[3]
BG17071	TACTCACTACAATGATGGTGACGTTAAA	FW for construction of pTarget-4MM-[3]
BG17072	TATGTTTAAACCAACGATCATTTGTAGTG	RV for construction of pTarget-4MM-[5]
BG17073	TACTCACTACAATGATCCTGTGGTTAAA	FW for construction of pTarget-4MM-[5]
BG17074	TATGTTTAAACCAAGTGGTACATTTGTAGTG	RV for construction of pTarget-4MM-[7]
BG17075	TACTCACTACAATGTACCACACTGGTTAAA	FW for construction of pTarget-4MM-[7]
BG17076	TATGTTTAAACCAAGTCCGAGTTTGTAGTG	RV for construction of pTarget-4MM-[9]
BG17077	TACTCACTACAACACTCGGACTGGTTAAA	FW for construction of pTarget-4MM-[9]
BG17078	TATGTTTAAACCAAGTCCATGGAAGTAGTG	RV for construction of pTarget-4MM-[11]
BG17079	TACTCACTACTTCCATGGACTGGTTAAA	FW for construction of pTarget-4MM-[11]
BG17080	TATGTTTAAACCAAGTCCATCAACAAAGTG	FW for construction of pTarget-4MM-[13]
BG17081	TACTCACTTGTTTGATGACTGGTTAAA	RV for construction of pTarget-4MM-[13]
BG17082	TATGTTTAAACCAAGTCCATCATCTG	FW for construction of pTarget-4MM-[15]
BG17083	TACTCAGATGAATGATGGACTGGTTAAA	RV for construction of pTarget-4MM-[15]
BG17084	TATGTTTAAACCAAGTCCATCATTTTCAC	FW for construction of pTarget-4MM-[17]
BG17085	TACTGTGAACAATGATGACTGGTTAAA	RV for construction of pTarget-4MM-[17]
SPR		
	bio- CAGCTATAGTTCTCGAAAGGTTTGGCACTCGACTAAAGGACTCTATGACC	biotinylated PAM-SCNR oligo
	GUGUCAUAGCCAGCUUGGCGGGCGAAGGCCAA GACGUCGAGUGCAAAAACCUUUCG	MmuCas12u1 PAM-SCNR RNA

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
C	C	C	C	C	C	A	A
T	T	G	A	T	T	C	C
T	T	G	A	A	T	T	G

A T A A T T T C
T C H A P T E R
A A A T T G A A
C C C C C A A 7
A A A A C T T T

Small and mighty: MmuCas12u1 C-to-T base editors

Wen Y. Wu, Sjoerd C.A. Creutzburg, Belén Adiego-Pérez, Timon Lindeboom, Karlijn Keessen,
John van der Oost[†]

Manuscript in preparation

[†]To whom correspondence should be addressed:
J.V.D.O. (john.vanderoost@wur.nl)

Abstract

CRISPR-Cas Base editors have recently been developed for RNA-guided targeted nucleotide substitution. The current toolbox consists of a wide variety of Cas9 and Cas12a-based fusion proteins that act as RNA-guided deaminases. Here we describe the development of several C to T base editors using a small Cas protein, MmuCas12u1 from the Type V-U1 system, named Mmu base editors (MmuBE). Molecular characterization of the designed MmuBEs has been performed in *Escherichia coli*, revealing that most variants are relatively efficient, with a base editing window consisting of two regions, a PAM-proximal (2-5) and a PAM-distal (13-19) region. In addition, a small-scale pilot experiment also demonstrated on-target base-editing by MmuBE in *Saccharomyces cerevisiae*. MmuBEs are currently the smallest base editors (genes ~2.8 kb) known, further expanding the current toolbox for prokaryotic base editing, and with great promise for eukaryotic base editing.

Introduction

In the last decade, genome editing by CRISPR-Cas nucleases has taken the world by storm, offering an effective, precise, and efficient way of genome editing (15, 203, 262). On the one hand, gene disruption relies on generating a double strand DNA break in the gene of interest, after which an error-prone repair of the break occurs through the non-homologous end joining (NHEJ) system, which appears abundant in eukaryotes but rare in prokaryotes (263, 264). For precise genome editing, on the other hand, a repair template must be delivered to the cell, requiring a homology-directed repair (HDR) system, the availability of which can substantially differ from one cell type to the other (265). It is important to note, however, that not all genome editing applications require large modifications, e.g. repairing a single nucleotide polymorphism (SNP) can be accomplished by a specific single nucleotide substitution (266). In addition, apart from repairing SNPs, single nucleotide mutations can also introduce a premature stop codon for generating gene knockouts (159, 161).

To circumvent the need to deliver a repair template for each single nucleotide mutation, base editors were developed. Synthetic CRISPR-associated base editor allows for RNA-guided, targeted nucleotide substitutions (C to T) on the non-target strand. The first base editor that was developed consisted of a chimeric construct of a Cas9, a cytidine deaminase and an uracil glycosylase inhibitor (UGI) (159, 160, 267). After crRNA-guided recognition, the catalytically inactive variant of Cas9 (D10A and H840A) also known as dead Cas9 (dCas9), which is unable to cleave dsDNA, targets and unwinds its dsDNA target. After DNA unwinding, the cytidine deaminase catalyzes the deamination of cytidine to uridine (C to U) in the displaced non-target strand, which leads to replacement by thymidine after replication, hence C to T. In addition, the role of the UGI domain is to inhibit the uracil glycosylase enzyme and as such preventing base excision repair, thereby increasing the C to T editing efficiency (Fig. 1A)

Initially, dCas9 was used, because the role of Cas9 was just to specifically bind and unwind of a selected dsDNA target. In subsequent base editor designs, however, nickase Cas9 (nCas9) variants are often used instead as it was found that a break in the target strand results in elevated base editing efficiencies, most likely by promoting mismatch repair in which the edited non-target strand serves as template, resulting in the desired overall base pair substitution: C-G via T-G to T-A (159, 160).

Up until now, several designs of Cas9 C to T base editors have been generated to reduce the base editing range within the protospacer (base editing window) and to increase the base editing efficiency (162, 268, 269). In addition, also a dCas12a C to T base editor has been created to expand the base editing toolbox, allowing for targeting of sequences downstream a 5' (T)TTV PAM instead of sequences upstream a 3' NGG PAM in case of Cas9 (270, 271). Cas9 and Cas12a base editors also differ with respect to their editing windows. Base editing positions are numbered relative

to the PAM-distal end and the PAM proximal end of the protospacer for Cas9 and Cas12a, respectively. For example, the NGG PAM sequence of Cas9 is numbered 21 to 23 and the (T)TTV PAM sequence of Cas12a is numbered -4 to -1. Cas9 and Cas12a base editors target C's in positions 3-8 and 8-13, respectively (159, 160, 162, 270-272).

Despite their potential for several applications, a drawback of Cas9/Cas12a-associated base editors is the fact that the genes encoding these chimeric proteins and their guides are way too big (~ 6 kb) to be delivered in mammalian cells by adeno associated virus (AAV) vectors (maximal cargo size 4.8 kb) (266, 273). AAV is the preferred delivery method for gene therapy because AAV can infect a wide range of cell types in the human body (*in vivo*), and it is qualified as safe by the US Food and Drug Administration (FDA) (274). To solve this size problem, split Cas9 base editors were created, of which each half of the Cas9 fusion protein was delivered by separate AAV vectors (275, 276). A more efficient approach would be the use of smaller base editors.

We have recently revealed relevant details of a small novel Cas protein, currently known as MmuCas12u1 (1.8 kb), that forms a clade in the rapidly growing CRISPR-Cas Type V (Cas12) (18). Similarly to the type V archetype, Cas12a, MmuCas12u1 was found to recognize a 5' TTN PAM, and to use a Cas12a-like crRNA to bind dsDNA (chapter 6). Under a range of conditions (*in vivo* and *in vitro*), no cleavage of the target dsDNA has been observed; potential RNA cleavage activity of the RuvC nuclease domain of MmuCas12u1 is currently being investigated. Previous work has shown that a mutation in the catalytic site of the RuvC domain of MmuCas12u1 (MmudCas12u1) does not affect binding of dsDNA (chapter 6). Utilizing its specific dsDNA binding capacity, MmudCas12u1 was fused to a cytidine deaminase to create various MmudCas12u1 C to T base editors. In this work, we constructed and tested various Mmu(d)Cas12u1 base editors (MmuBE) in *E. coli* and *S. cerevisiae*.

Results

Smallest C to T base editor edits in two regions

The first MmuCas12u1 base editor constructed in this work consists of a catalytically inactive MmuCas12u1, termed dead MmuCas12u1 (MmudCas12u1), a 121-amino acid linker, a cytidine deaminase protein CDA, an uracil glycosylase inhibitor UGI, and an LVA degradation tag to reduce toxicity of the BE (Fig. 1A) (267). This first construct is termed MmuBE_E1, based on the nomenclature of the prokaryotic Cas9 base editors (267). A test for base editing was developed, by growing *E. coli* cells

harboring 3 plasmids: pCas, pCRISPR and pTarget. pCas and pCRISPR express the base editor and the CRISPR array, respectively, whereas pTarget plasmids contain the protospacer target sequence. We generated six variants of pTarget, which had six consecutive Cytosine bases at six different positions of the protospacers, termed C-tile plasmids (Fig. 1C). These boxes of six C's shift 3 positions towards the 3' end until the 20th position is reached (Fig. 1C). This method ensures overall C coverage on the protospacer. In addition to the C-tiles, a C at position 3 (C3) was also always present and served as an internal standard for base editing.

E. coli cells harboring all three plasmids were grown for 48 hours. Samples were taken at time points 16, 24 and 48 hours and were used for PCR amplification. Subsequent deep sequencing of the obtained amplicons was performed to assess base editing of the whole population. Sequence analysis revealed that base editing occurred in each of the six C-tile plasmids, with efficient C3 base editing (>90% editing) in all plasmids (Table S1). Next, the results of the uneven C-tile plasmids (1, 3 and 5) and of the even C-tile plasmids (2, 4 and 6) were merged to reveal the base editing window (Fig. 1D, S1A and B). Interestingly, it was found that MmuBE_E1 catalyzes base editing in two different regions within the base editing window instead of the one found for previously described base editors (266). These regions consist of a PAM-proximal (positions 2-5) and PAM-distal (positions 13-19) region. Positions 3 and 4 were found to have the highest base editing efficiency of >75%. Base editing efficiency for C15 varied between plasmids (C-tile 5 and C-tile 6). C15 base editing was found to be 41% for C-tile 5 and 94% for C-tile 6. These differences are most pronounced in position C15 but can also be seen for other positions, such as C6, C16 and C17. This may be caused by sequence specific base editing biases i.e., context dependent base editing. In an attempt to reduce the second base editing region, the spacer length was reduced, ranging from 14-17nt (Fig. S2A). The use of shorter spacers leads to shorter R-loops, which reduces the availability of ssDNA on the 3' of the protospacer and thereby the base editing in that region (Fig. S2B). From Sanger sequencing data of the whole population, a spacer length of 14nt was found to be able to reduce the extension of the 2nd base window to positions 14-16 (Fig. S1C, D and E). However, this approach also increases the likelihood of off-targeting. For that reason, a different approach was taken to reduce the base editing window, as described below.

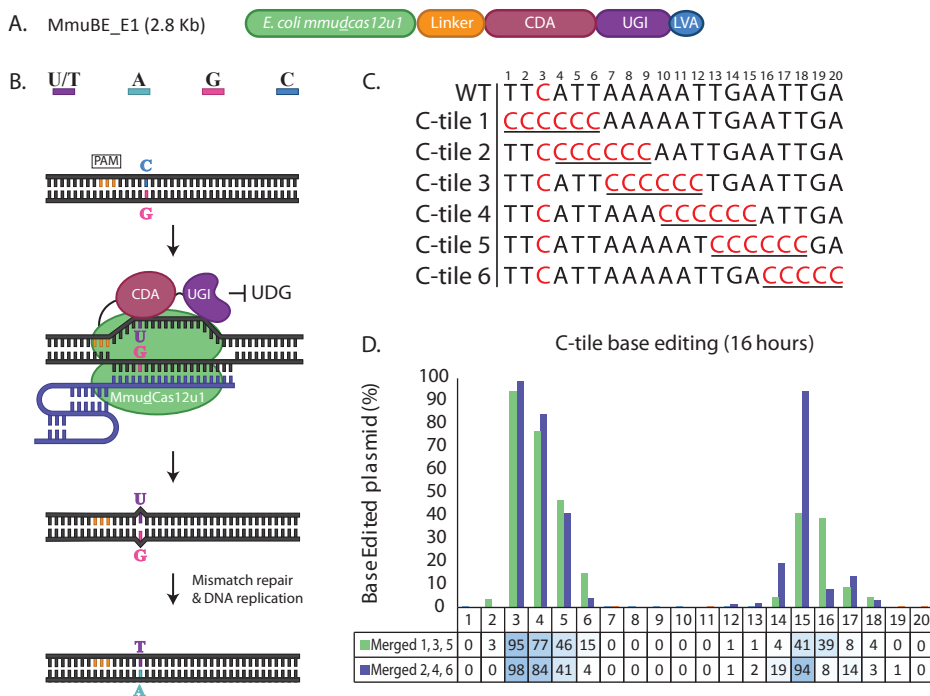


Figure 1 | C to T base editing by MmuBE1_E1. (A) Schematic of MmuBE_E1 gene. (B) Schematic of base editing process by MmuBE_E1. MmuBE_E1 recognizes a 5' TTN PAM (orange) and binds to its target. Once an R-loop is formed, CDA (bordeaux) deaminates a C to a U. Then mismatch repair and DNA replication generate a dsDNA containing a T instead of a C. (C) Overview of the C-tile targets used to characterize the editing window of MmuBE_E1. The wildtype sequence contains a C at position 3 and serves as an internal standard for base editing. C-tile 1 to C-tile 6 plasmids contain six consecutive C's in the sequence and shifts three position toward 3' end until position 20 is reached. (D) Deep sequencing results of MmuBE_E1 targeting the C-tile plasmids after 16 hours. Data from plasmids of uneven and even numbers were fused for easier data overview corresponding to 'Merged 1, 3, 5' and 'Merged 2, 4, 6', respectively. Y-axis represent base edited plasmids in % of the whole plasmid population and x-axis represent the C position within the protospacer.

Characterization of various MmuBEs in *E. coli*

Various MmuBEs were designed by varying the deaminase module as well as the linker length (Fig.1). Linker variation consisted of trimming down the flexible linker that was used in MmuBE_E1 from 121 to 97, 67 and 29 amino acids (aa). In addition, a rigid linker (33 aa) was tested as well (272). MmuBE_E1 base editors with these linkers were named MmuBE_E1.A-D (Fig. 2). Next to creating *E. coli* MmuBEs, several MmuBEs were also constructed for editing of mammalian cells. For constructing these MmuBE_H variants, we used *H. sapiens* codon harmonized *mmudcas12u1*,

H. sapiens optimized cytidine deaminases (CDA or rAPOBEC1) and *H. sapiens* optimized uracil glycosylase inhibitor (UGI), termed. The MmuBE_H1.A and MmuBE_H1.B variants contain CDA and UGI fused with a 121 aa or 16 aa linker, respectively. Using the same 16 aa linker, MmuBE_H2 and MmuBE_H2YE were constructed using rAPOBEC1 and rAPOBEC1 YE, respectively (159). rAPOBEC1_YE was previously shown to have a narrower editing window compared to WT rAPOBEC1 (162, 271). MmuBE_H variants were also tested in *E. coli* to validate their base editing potential, prior to testing in human cells.

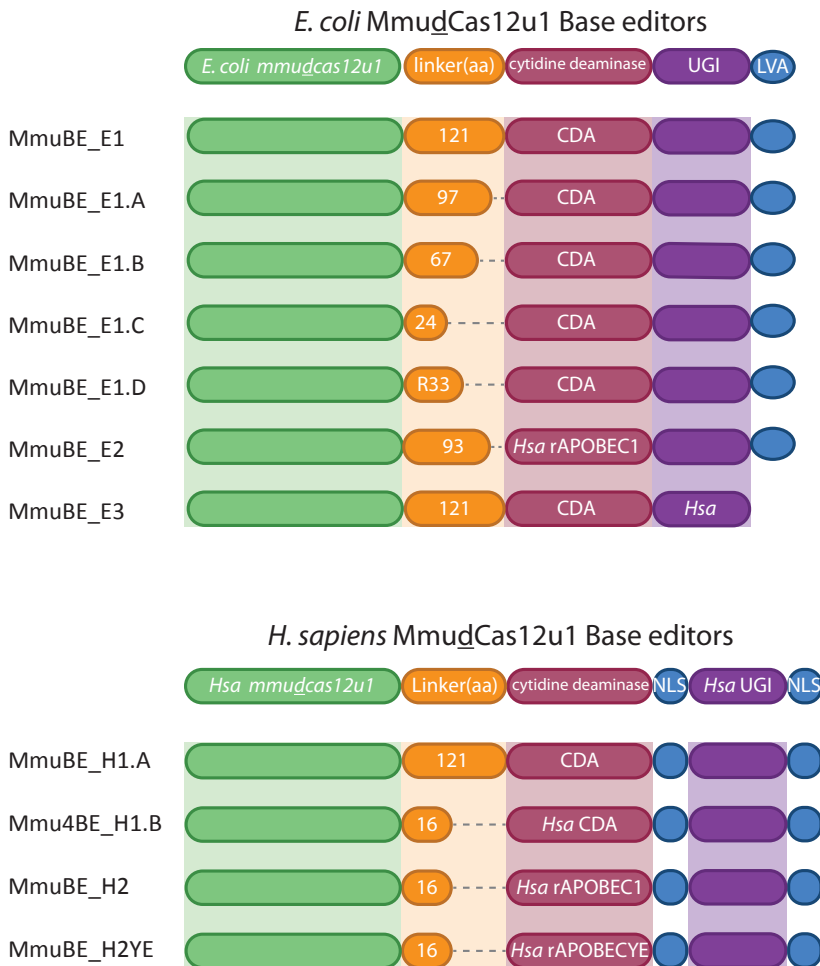


Figure 2 | Schematic of different MmuBEs. All MmuBE consists of a dMmu (green), linker (orange), cytidine deaminase (bordeaux), UGI (purple). *E. coli* and *H. sapiens* Mmu Base editors consist of genes harmonized or optimized for *E. coli* and *H. sapiens*, respectively. Linkers are indicated with a number, representing the aa length. In addition, MmuBE_E and MmuBE_H also have an LVA degradation tag or nuclear localization sequences (NLS), respectively (blue).

Prior to base editing, all MmuBEs were tested for binding activity of MmudCas12u1 *in vivo* using a GFP silencing assay. MmuBEs targeted a short *gfp* sequence containing no C nucleotide (only A, G or T nucleotides), so C-to-T base editing of the target sequence cannot occur (Fig.3A). A frame shift *E. coli* MmudCas12u1 (FSdMmu) and *E. coli* MmudCas12u1 were included to function as negative and positive controls, respectively. GFP fluorescence was measured and normalized to FSdMmu (Fig. 3B) and therefore, all percentages showed in figure 3B are relative to the fluorescence of this strain. It was found that all *E. coli* MmuBEs (MmuBE_E) were able to bind to the target DNA, i.e. decreasing the GFP levels to < 5% when compared to the negative control levels. MmuBE_E base editors silenced GFP similarly to the positive control dMmu (*E. coli* harmonized). As for MmuBE_H base editors, all MmuBE_H were found to have lower silencing activity when compared to the dMmu control, with 20-50% of GFP fluorescence still being detected. Out of the MmuBE_H base editors, MmuBE_H1.A and MmuBE_H2 show the best silencing activity with only 18% and 22% GFP fluorescence detected, respectively. This is followed by MmuBE_H1.B with 35% GFP fluorescence and then MmuBE_H2YE with the least silencing, with 47% of GFP fluorescence still being detected. Difference in silencing between MmuBE_E and MmuBE_H base editors can be due to expression differences affected by codon usage of *E. coli*. After testing the binding activity of various MmuBEs, base editing activity was tested.

The different C motif plasmids contain a tiled C motif (CxxCxxCxxCxxCxxCxxC), starting at every first (C1 motif), second (C2 motif) or third (C3 motif) nucleotide of the protospacer (Fig. 3C). Cells containing pCas (expressing Mmu BE), pCRISPR (expressing CRISPR array) and C-motif plasmids were grown for 48 hours and were used for a population PCR, which amplified the protospacer region on the C-motif plasmids. Amplified products were sent for Sanger Sequencing and results were analyzed by EditR (277). Base editing results obtained from all three C motif plasmids were merged and visualized in a heatmap (Fig.3D). It was found that trimming the MmuBE_E1 linker from 121 aa to 24 aa (MmuBE_E1.C) had no effect on editing of either of the two base editing regions (Fig.3D). However, MmuBE_E1.D, containing a 33 aa rigid linker showed slightly lower base editing activity in the PAM-distal region. Unexpectedly, also MmuBE_E2 and MmuBE_E3, which have long flexible linkers (93 aa and 121 aa), showed reduction of the PAM-distal region. MmuBE_E2 contains a *H. sapiens* optimized rAPOBEC1 instead of CDA and MmuBE_E3 contains a *H. sapiens* optimized UGI instead of the *E. coli* optimized UGI. Expression of these *H. sapiens* optimized genes in *E. coli* probably affect folding of the fusion proteins thereby changing the total number of active Mmu_BE proteins in the cell. Next, MmuBE_H base editors were also found to be active in *E. coli*, although they show lower base editing activity compared to MmuBE_E base editors (Fig.3E). MmuBE_H1.A and MmuBE_H1.B also have two base editing regions, but with reduced overall activities. MmuBE_H1.A edits C's at position 2-4 and 14-16, whereas MmuBE_H1.B (containing a shorter linker of 16 aa) edits C's at position 3-6 and 15-16. This suggests that, in these constructs, linker reduction from 93 to 16 aa results in a slight shift of the PAM-proximal base editing region.

The most precise MmuBEs in *E. coli* were found to be MmuBE_H2 and MmuBE_H2YE, with base editing detected only in the PAM-proximal region (Fig3.E). MmuBE_H2 edits C's at position 3, 5 and 6, whereas MmuBE_H2YE only edits at position 4 with little to no editing found at position 12 and 15. However, although MmuBE_H2 and MmuBE_H2YE have a narrow editing range, it should be mentioned that both base editors have a significantly lower base editing activity when compared to other MmuBEs. Hence, the detected narrow base editing window appears to be a consequence of a lower editing efficiency. The reduced editing activity may have different explanations: increased expression of human-codon optimized *mmuCas12u1* (in line with aforementioned reduction of silencing efficiency), of Hsa-APOBEC1-type cytosine deaminase, and of human-codon optimized uracil glycosylase inhibitor (Hsa-UGI). All these MmuBEs should still be analyzed by deep sequencing to validate the presented results obtained by Sanger Sequencing. Nonetheless, a variety of MmuBEs was created with differences in base editing windows, providing a wide selection of MmuBEs and further expanding the base editing toolbox in *E. coli*.

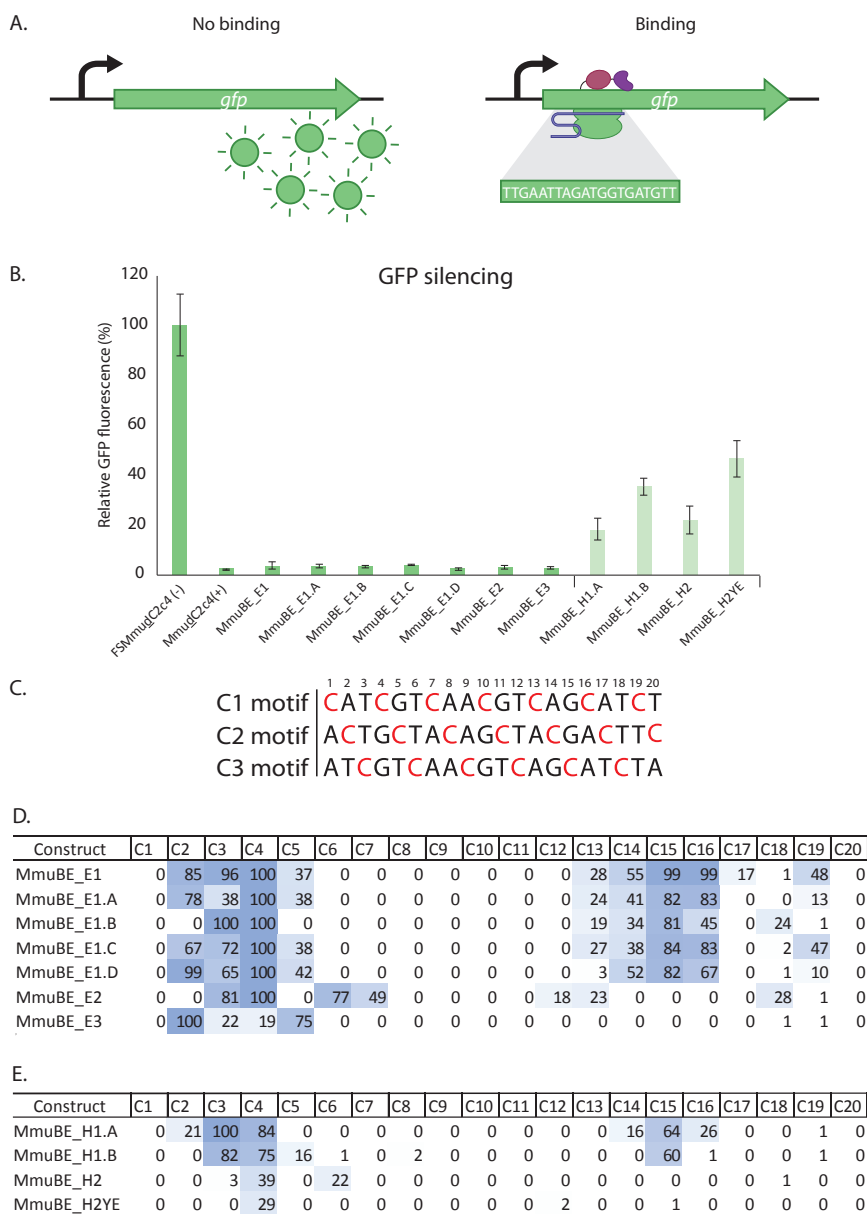


Figure 3 | Silencing and base editing by various MmuBEs. (A) Schematic of GFP silencing by MmuBE. **(B)** GFP silencing by various MmuBEs. Y-axis represents relative GFP fluorescence in % where negative control frameshift dMmu (FSDMmu) was used as 100%. X-axis represent the different MmuBEs tested. **(C)** Base editing targets consisting of a C on every first, second and third position of each trinucleotide. These plasmids were named C1, C2 and C3 motif, respectively. **(D)** Heat map representing % of base edited C's using different variants of *E. coli* MmuBEs (MmuBE_E). Data was obtained by fusion C1, C2 and C3 motif data. **(E)** Heat map representing % of base edited C's using different variants of *H. sapiens* MmuBEs (MmuBE_H). Data was obtained by fusion C1, C2 and C3 motif data.

MmuBE base edits in *S. cerevisiae*

To check whether a MmuBE can also function in eukaryotes, a MmuBE_S was constructed and tested in *Saccharomyces cerevisiae*. MmuBE_S, contains a *S. cerevisiae* codon-optimized *mmucas12u1*, a 93aa linker, and human codon-optimized variants of CDA and UGI (Fig 4A). Apart from the *S. cerevisiae* optimized *mmucas12u1*, MmuBE_S is similar to MmuBE_H1.A. MmuBE_S targeted the *ade2* reporter gene in the genome of *S. cerevisiae*. Targeted C to T mutation in certain positions in *ade2* results in the introduction of a premature stop codon, disrupting the *ade2* gene. In the absence of adenine and when *ade2* is knocked out, *S. cerevisiae* accumulates an intermediate of the adenine biosynthetic pathway (P-ribosylamino imidazole), which in aerobic conditions is oxidized to become a red pigment that can be visualized as red colonies on plates, easily discriminated from the white wild type (*ade2+*) colonies (Fig 3B). Red colonies were selected for colony PCR and subsequent analysis of the obtained amplicons was performed by Sanger sequencing to confirm targeted base editing of the *ade2* gene (Fig 3C). By varying the crRNA guides, MmuBE_S targeted three position in the *ade2* gene, of which C to T mutation in position 2, 3, or 4, respectively, leads to a nonsense mutation by converting a glutamine (Q) codon (CAA) to a stop codon (TAA) (Fig.3C). Selected colonies were sent for sequencing of the three different targets, ADE2_1, ADE2_2 and ADE2_3. The sequencing results of the three targets, revealed that two out of two (2/2), one out of five (1/5) and two out of two (2/2) were found to have the designed C to T base editing, respectively (Fig.3C). Some red colonies did not contain targeted C to T mutations, such as the ones found in ADE2_2 and non-targeting samples. These clones appeared to be *ade2* frame shift mutants, either due to spontaneous deletions or insertions. In addition, some red colonies were also found to have off-target base editing in the *ADE2* gene, causing missense mutations, P508L and P472L (data not shown). Based on These initial analyses demonstrate that targeted Mmu-dependent base editing in *S. cerevisiae* is possible. However, it is unclear how efficient and how specific this type of base editing is. Hence, more quantitative analysis still needs to be done by full population deep sequencing. Also, off-target base editing should be further investigated.

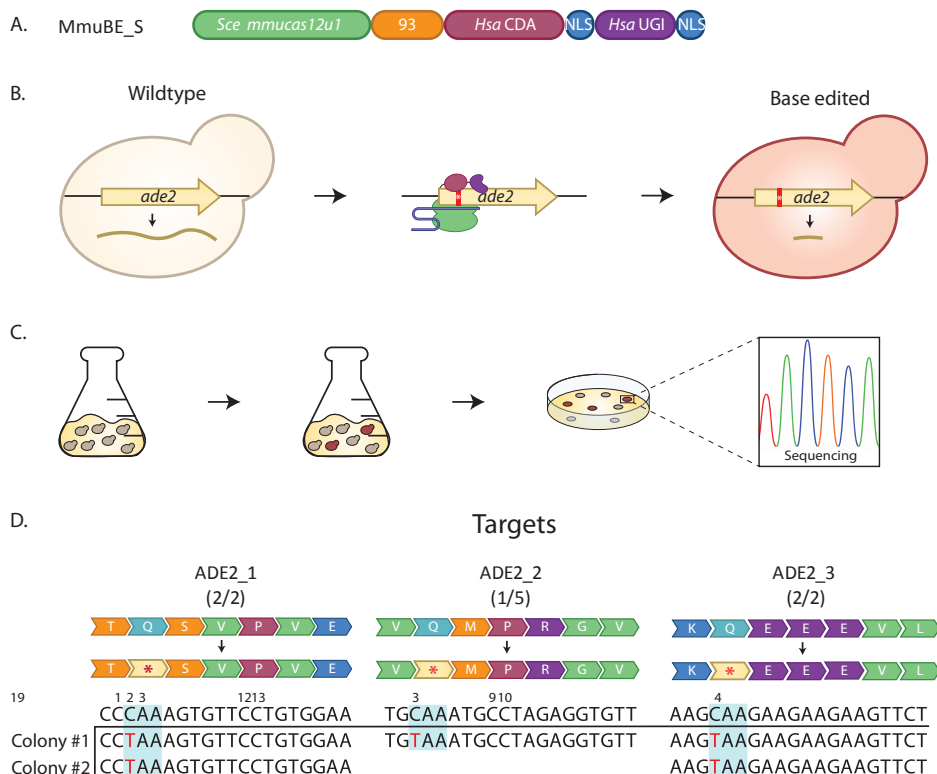


Figure 4 | Base editing in *S. cerevisiae* using MmuBE_S. (A) Schematic of MmuBE_S gene. (B) Experimental set-up for testing base editing in *S. cerevisiae*. *ade2* gene in *S. cerevisiae* is targeted by MmuBE_S and if successfully base edited, premature STOP codon is created (red line). If ADE2 is not expressed in the absence of adenine, a red pigment accumulates into the cells and the yeast colony will appear red on the plate. Red colonies were picked, *ade2* region amplified and sent for Sanger sequencing. (C) Schematic of base editing workflow for *S. cerevisiae*. Cells expressing the MmuBE_S are cultured in flasks for 24 hours, plated to distinguish between edited cells (red) and non-edited cells (white) and sent for sequencing. (D) Sequencing results of three MmuBE_S targets, ADE2_1, ADE2_2, ADE2_3. Numbers indicate number of red colonies that were base edited compared to the number of colonies sent for sequencing. Red indicated position where base editing took place.

Discussion

Previously, MmuCas12u1 was characterized to be a small nuclease (1.8 kb), guided by a crRNA to recognize and bind dsDNA (chapter 6). In this work MmuCas12u1 was used to construct various MmuBEs for *E. coli*, *H. sapiens* and *S. cerevisiae*. MmuBEs were constructed by fusing MmuCas12u1 to a cytidine deaminase (CDA, rAPOBEC) and an uracil glycosylase inhibitor (UGI) at its C-terminus end.

N-terminus fusions were also constructed but were found unable to bind dsDNA in our GFP silencing assay (data not shown). The MmuBEs characterized in this work have a base editing window consisting of two regions. The location of base editing is determined by several factors: the Cas protein structure, the linker length, and the type of cytidine deaminase. Different Cas proteins, such as Cas9 and Cas12a, were also found to have different base editing windows when using the same cytidine deaminase and linker (159, 162, 266, 271). The size and/or structure of MmuCas12u1 might be the cause for the two base editing regions, since a small protein would be unlikely to cover the entire protospacer, leaving outer ends of the protospacers exposed as exposed ssDNA and available for the cytidine deaminase. However, only by solving and studying the crystal structure of MmuCas12u1 can we further confirm this hypothesis.

Other studies have found that base editing windows can be modified by varying linkers in length and flexibility (272). Reducing the MmuBE-E1 base editor linker to 24 aa (MmuBE_E1.C) was found to have little to no effect on the PAM distal base editing region. However, when the MmuBE_H1.A linker was shortened to 16 aa (MmuBE_H1.B), a reduction of the PAM distal base editing region could be observed. The difference on the effect of linker reduction on the PAM distal editing region between MmuBE_E and MmuBE_H can be due to the overall lower base editing efficiency of MmuBE_H, which may allow for detection of small base editing differences in the PAM distal region. Another reason is the 24 aa linker length of MmuBE_E1.C was not short enough to reduce the second base editing region. Therefore, a shorter linker length of 16 aa or shorter should be tested. Contrary to Cas9 base editors, a rigid linker for MmuBE_E instead of a flexible linker resulted in similar base editing regions (272). However, also varying cytidine deaminases did result in slightly different base editing regions. MmuBE_H1.B, MmuBE_H2 and MmuBE_H2YE differ in their cytidine deaminase genes being CDA, rAPOBEC1 and rAPOBEC1(YE), respectively. All three base editors resulted in different base editing windows, with MmuBE_H1.B having the widest window, MmuBE_H2 with an intermediate window, and MmuBE_H2YE having the narrowest window, consistent with previous reports (162, 271). However the narrow window of MmuBE_H2YE can be caused by overall lower base editing efficiency of APOBEC1(YE) (162). Aside from proteins and linkers, expression of base editors proteins or dsDNA binding activity was found to influence base editing regions (278). Base editors that have lower silencing activity, such as the MmuBE_H base editors, were found to have lower base editing activity in *E. coli*, which led to narrower base editing regions, because only the most efficient positions will be base edited with moderate efficiency. Hence, fine tuning the expression of base editors can be an effective approach for more precise base editing. These results showcase efficient base editing in *E. coli* using a variety of MmuBEs with a wide range of base editing windows. In addition, deep sequencing analyses are still required to verify results obtained from Sanger sequencing. Even though no 'proximal off-targets' were detected in *E. coli*'s Sanger sequencing data, 'distal' off-target can still occur in other locations in the genome, especially for C to T base editors (279, 280). Therefore, full genome sequencing should be done to assess full genome of targeting of MmuBEs.

Preliminary results in *S. cerevisiae* have shown to have promising on-target base editing but must be further investigated to assess base editing efficiency and occurrence of off-target edits. At present, it is not known whether MmuBEs are functional in mammalian cell lines. Application of MmuBEs in mammalian cells can truly benefit from the small size of MmuBEs. MmuBEs are the smallest base editors known (2.7-2.8 kb) to fit in the AAV vector with a capacity of 4.7kb and leaves ample space for the CRISPR array. Also, MmuBEs can be used complementary to Cas12a base editors, as both proteins recognize a 5'TTN- PAM and base edit in complementary regions of the protospacer. In summary, this study shows the efficacy of various MmuBEs in *E. coli*, of which some may have great potential to be utilized for gene therapy in mammalian cells.

Methods

***E. coli* strains and growth conditions**

E. coli strains DH5- α and DH10- β were used for plasmid construction. *E. coli* BW25113 strain, lacking *lacI*, *lacZ* and the type I-E CRISPR-Cas system were used for all other experiments. Cells were cultured in 37°C at 220 rpm in Luria Bertani (LB) medium (10 g/L peptone (Oxoid), 10 g/L NaCl and 5 g/L yeast extract (BD)). Antibiotics, such as ampicillin (100 μ g/mL), kanamycin (50 μ g/mL) and chloramphenicol (35 μ g/mL) were added when required.

***S. cerevisiae* strains and growth conditions**

Yeast strains built in this study belong to the *S. cerevisiae* CEN.PK113-5D background. Strains can be found in Table x. YSTB164, contains a genome integrated *egfp* gene under control of the *Kluyvyromyces lactis* promoter of KLLA0F20031g (kl11p), in the INT1 site as previously described (281). YSTB164 was used as parental strain for all *S. cerevisiae* strains expressing MmuBE_5. *S. cerevisiae* was cultured in YPD media (10 g/L yeast extract (BD), 20 g/L peptone (Oxoid) and 20 g/L glucose) or synthetic medium (SMG) (3 g/L KH₂PO₄, 0.5 g/L MgSO₄·7H₂O, 5 g/L (NH₄)₂SO₄, 1 mL/L of a trace element solution, and 1 g/L of a vitamin solution as previously described (282)). When required, the media was supplemented with 200 mg/mL G418 (Geneticin). When required, selection with G418 on SMG media, (NH₄)₂SO₄ was replaced with 3 g L⁻¹ K₂SO₄ and 2.3 g/L urea to avoid pH drop (283).

***E. coli* Plasmid construction**

The plasmids constructed in this study and the oligonucleotides (IDT) used for cloning and sequencing can be found in Supplementary Tables S5 and S4, respectively. All

fluorescence and base editing assays in *E. coli* were performed in a three-plasmid system, which was based on the previously published PAM-SCNR screening method (242). The three-plasmid system consisted of pCas, pCRISPR and pTarget. pCas expresses MmuBEs under the control of the constitutive J23108 in a pBAD33 vector. pCRISPR expresses the Mmu CRISPR array under a J23119 promoter in pBAD18 backbone. pTarget contains the targeted protospacer and expresses a *gfp* gene under a constitutive promoter Placq in a pAU66 backbone. More in-depth cloning details of various pCas plasmids can be found in Table S6. The pCas-MmuBE_E1 was constructed using NEBuilder® HiFi DNA Assembly (NEB). DNA fragments used in the assembly were amplified by PCR using Q5® High-Fidelity 2X Master Mix (NEB). The pCas-MmuBE_E1 was then used to construct pCas-MmuBE_E1.A, pCas-MmuBE_E1.B and pCas-MmuBE_E1.C by Golden Gate cloning. The vector and linker were PCR amplified to introduce flanking SapI restriction sites. To enable more straightforward cloning of the other Mmu base editors, pCas-RFP-UGI-Entry was constructed. pCas-RFP-UGI-Entry which contains a *rfp* and an UGI gene. The *rfp* gene is flanked BbsI restriction sites, which can be used for Golden Gate cloning and visualization of correctly assembled plasmid by the absence of RFP fluorescence. pCas-RFP-UGI-Entry was used to construct pCas-MmuBE_E2, pCas-MmuBE_E3, pCas-MmuBE_H1.B, pCas-MmuBE_H2 and pCas-MmuBE_H2YE. Besides pCas-RFP-UGI-Entry, which was digested with restriction enzyme BbsI-HF® (NEB), all other fragments were PCR amplified to introduce BbsI restriction sites in each fragment. pCas-MmuBE_H1.B, pCas-MmuBE_H2 and pCas-MmuBE_H2YE were later found to contain a deletion within the linker, of which causes a frameshift in the fusion protein. These plasmids were then repaired by PCR amplification followed by blunt-end ligation of the linear fragment. Lastly, pCas-MmuBE_H1.A was constructed using NEBuilder® HiFi DNA Assembly (NEB), using fragment amplified from pCas-MmuBE_H1.B and pCas-MmuBE_E1.

The pCRISPR plasmids were constructed by restriction-digestion and ligation. pCRISPR_NT (chapter 6) contains an Mmu CRISPR array with a non-targeting spacer flanked by BbsI restriction sites and CRISPR repeats. To improve this cloning vector, the non-targeting spacer was replaced with an *mruby* gene by digestion and ligation, to create pCRISPR-Mmu-mRuby-Entry. All pCRISPR plasmids were then constructed by digesting pCRISPR-Mmu-mRuby-Entry with restriction enzyme BbsI-HF®, and subsequent ligating it with a short spacer sequence. Spacer sequences containing complementary overhangs were created by annealing two oligonucleotides (Table S4).

pTarget plasmids used for base editing such as C-tile and C-motif plasmids were constructed using a fragment of pTarget-divergent (Chapter 6) digested with the restriction enzymes, AatII and KpNI and subsequent ligating it to a short protospacer sequence. Protospacer sequences containing complementary overhangs were created by annealing two oligonucleotides (Table S4).

***S. cerevisiae* plasmid construction**

The plasmids constructed in this study and the oligonucleotides (IDT) used for cloning and sequencing can be found in Supplementary Tables S7 and S4, respectively. Mmu base editors in *S. cerevisiae* were genome integrated to generate various strains expressing different targeting guides expressed from a multicopy plasmid (Table S3). CRISPR arrays for Cas12a and MmuCas12u1 were expressed under control of the SNR52 promoter on a PL-074 backbone.

Initially, PL-074 was constructed to correct the SUP4 terminator sequence to its original length, by PCR amplification of pUD628 and subsequently re-circularizing it by blunt-end ligation. PL-098 was constructed by incorporation of the INT1 spacer as an overhang in the forward primer used for linearization of PL-074 by PCR amplification. In order to incorporate the MmuCas12u1 repeats, PL-162 was built by restriction digestion of pCRISPR_NT (BbsI) (chapter 6) with BbsI-HF® and ligation with a spacer created by annealing two oligonucleotides. PL-162 was then used to amplify the MmuCas12u1 CRISPR array containing a spacer flanked by BsaXI restriction sites instead of BbsI (fragment A0185). A0185 was digested in a two-step protocol with restriction enzyme KpnI and BtgZI. Afterwards, staggered ends were removed by T4 DNA polymerase (NEB). The blunted product was ligated into PCR amplified PL-074, to construct PL-163. PL-139 was constructed using the same protocol, except that a non-targeting spacer fragment obtained by annealing two oligonucleotides was used instead for ligation to BbsI restriction digested pCRISPR_NT (BbsI), obtaining the intermediate plasmid PL-138.

For easy screening correctly assembled plasmids, PL-196 was built which contains a *rfp* gene between the MmuCas12u1 repeats. PL-196 was constructed by HiFi® assembly of four PCR amplified fragments. Two backbone fragments were obtained from PL-163 and two RFP expression cassette fragments were obtained from pCRISPR-Cas12a-entry. Subsequently, MmuCas12u1 CRISPR array plasmids were built by BsaXI digestion of PL-196 and ligation of annealed oligonucleotide pairs with adequate overhangs.

Fluorescence repression assay

For the GFP silencing assays, *E. coli* cells harbouring pTarget-GFP (chapter 6) and pCRISPR-GFP were made chemically competent and transformed with the different Mmu base editor (pCas) plasmids. After recovery, the transformation mix was diluted 2 µL:200 µL M9TG medium in a 96 well 2 mL master block (Greiner). Master block was then sealed using a gas-permeable membrane (Sigma, AeraSeal™) and grown overnight at 37 °C at 900 rpm overnight. The following day, the cells were diluted 1:10000 in fresh M9TG medium in a 96-wells master block and grown overnight at 37°C. Overnight cultures were then used for fluorescence measurements.

Plate reader measurements

Overnight cultures were diluted 1:10 in 200 μ L PBS and measured on a Biotek Synergy MX microplate reader a Synergy MX microplate reader. Cell density was measured with 600 nm and GFP fluorescence was measured with an excitation of 405 nm and emission of 508 nm. GFP was measured using a gain of 50, 75 and 100.

Fluorescence was calculated as

$$\frac{\text{average} \left(\frac{F_{X_{\text{targeting}}} - F_{\text{Blank}}}{OD_{600_{X_{\text{targeting}}}} - OD_{600_{\text{Blank}}}} \right)}{\text{average} \left(\frac{F_{\text{FS}} - F_{\text{Blank}}}{OD_{600_{\text{FS}}} - OD_{600_{\text{Blank}}}} \right)}$$

Base editing assay

E. coli cells harboring pCRISPR-C-tile or pCRISPR-C motif plasmids and their corresponding pTarget plasmids were made chemically competent and transformed with the different Mmu base editor (pCas) plasmids. After recovery, the transformation mix was diluted 2 μ L:200 μ L M9TG medium in a 96 well 2 mL master block (Greiner). Master block was then sealed using a gas-permeable membrane (Sigma, AeraSeal™) and grown overnight at 37 °C at 900 rpm overnight. The following day, the cells were diluted 1:10000 in fresh M9TG medium in a 96-wells master block and grown overnight at 37°C. 20 μ L *E. coli* cultures were taken every at time point 16, 24 and 48 hours for C-tile base editing, whereas samples were only taken at 40 hours for C-motif base editing. Base edited region was PCR amplified by using 2 μ L cultures in a 50 μ L PCR reaction using Q5® High-Fidelity 2X Master Mix (NEB). Amplified fragments were purified using DNA Clean & Concentrator™-5 (Zymo Research) and sent for sequencing.

S. cerevisiae transformations

In order to construct a *S. cerevisiae* strain with genomic integration of *egfp*, an *egfp* expression cassette was integrated into integration site 1 (INT1) (281). A *S. cerevisiae* strain harboring pUDE731 (YSTB013) was transformed with 500 ng of PL-098 and four linear DNA fragments by the LiAc/SS carrier DNA/PEG method (284): one containing the *Kluyvyromyces lactis* promoter of KLLA0F20031g (*k111p*); another harboring the *egfp* gene and the CYCc terminator from pCFB2791 and two linear fragments homologous to the INT1 site as previously described (281). Correctly assembled and integrated cells were assessed by colony PCR and sequencing with primers listed in Table S4. After sequential sub-culturing in liquid YPD and a last culture on YPD-agar for plasmid curing. One colony isolate was selected and named YSTB164.

Subsequently, strains YSTB305 and YSTB211 were transformed with plasmids PL-242 to PL-246 and PL-139. Obtained colonies were investigated for phenotype change (red pigment accumulation in case of *ade2* knockouts).

Base editing assessment in *S. cerevisiae*

Red colonies were picked and re-streaked on YPD + G418 media until single red colonies were isolated. Individual colonies were picked for genomic DNA amplification using Q5® High-Fidelity 2X Master Mix (NEB). PCR products were analyzed with Sanger sequencing (Macrogen) with primers.

Acknowledgments

This study is supported by the Dutch Research Council (NWO) by a TOP grant (714.015.001) to J.v.d.O.

Author contributions

W.Y.W. and J.v.d.O conceived this study and the experimental design. W.Y.W., S.C.A.C, B.A.P., T.L. and K.K. conducted the experimental work. W.Y.W. and J.v.d.O. wrote the manuscript.

Competing interests

A patent application has been filed related to this work.

Corresponding author

Correspondence and requests for materials should be addressed to J.v.d.O. (john.vanderoost@wur.nl).

Supplementary Figures and Tables

Table S1 | C-tile base editing data of 16, 24 and 48 hours.

C-position #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16 hour																				
WT	0,00	0,00	94,63	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1. C-tile 1-6	0,00	3,46	94,82	76,85	46,45	14,68	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2. C-tile 4-9	0,00	0,00	99,24	84,41	41,10	3,74	0,26	0,11	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3. C-tile 7-12	0,00	0,00	99,24	0,00	0,00	0,00	0,29	0,11	0,03	0,02	0,09	0,54	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4. C-tile 10-15	0,00	0,00	99,40	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,10	0,73	1,58	19,31	94,47	0,00	0,00	0,00	0,00	0,00
5. C-tile 13-18	0,00	0,00	99,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,53	4,42	40,62	38,62	8,26	4,35	0,00	0,00
6. C-tile 16-20	0,00	0,00	99,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	92,28	86,26	97,26	99,46	99,82
24 hour																				
WT	0,00	0,00	94,29	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1. C-tile 1-6	0,00	3,38	94,71	76,33	46,16	14,56	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2. C-tile 4-9	0,00	0,00	99,39	84,58	41,23	3,89	0,28	0,10	0,15	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3. C-tile 7-12	0,00	0,00	99,39	0,00	0,00	0,00	0,23	0,09	0,02	0,01	0,06	0,39	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4. C-tile 10-15	0,00	0,00	99,36	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,12	0,79	1,64	19,31	94,24	0,00	0,00	0,00	0,00	0,00
5. C-tile 13-18	0,00	0,00	99,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,56	4,50	39,65	39,82	8,83	4,76	0,00	0,00
6. C-tile 16-20	0,00	0,00	99,35	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	92,76	86,46	97,39	99,48	99,79
48 hour																				
WT	0,00	0,00	94,47	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1. C-tile 1-6	0,00	3,99	96,64	81,43	49,47	16,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2. C-tile 4-9	0,00	0,00	99,20	91,54	47,49	9,34	0,77	0,24	0,24	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3. C-tile 7-12	0,00	0,00	99,20	0,00	0,00	0,00	0,72	0,29	0,08	0,03	0,27	1,35	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4. C-tile 10-15	0,00	0,00	99,22	0,00	0,00	0,00	0,00	0,00	0,00	0,04	0,48	2,37	4,19	38,13	98,01	0,00	0,00	0,00	0,00	0,00
5. C-tile 13-18	0,00	0,00	99,26	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,11	7,88	60,75	52,22	14,19	7,65	0,00	0,00
6. C-tile 16-20	0,00	0,00	99,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	91,84	84,57	97,07	99,35	99,77

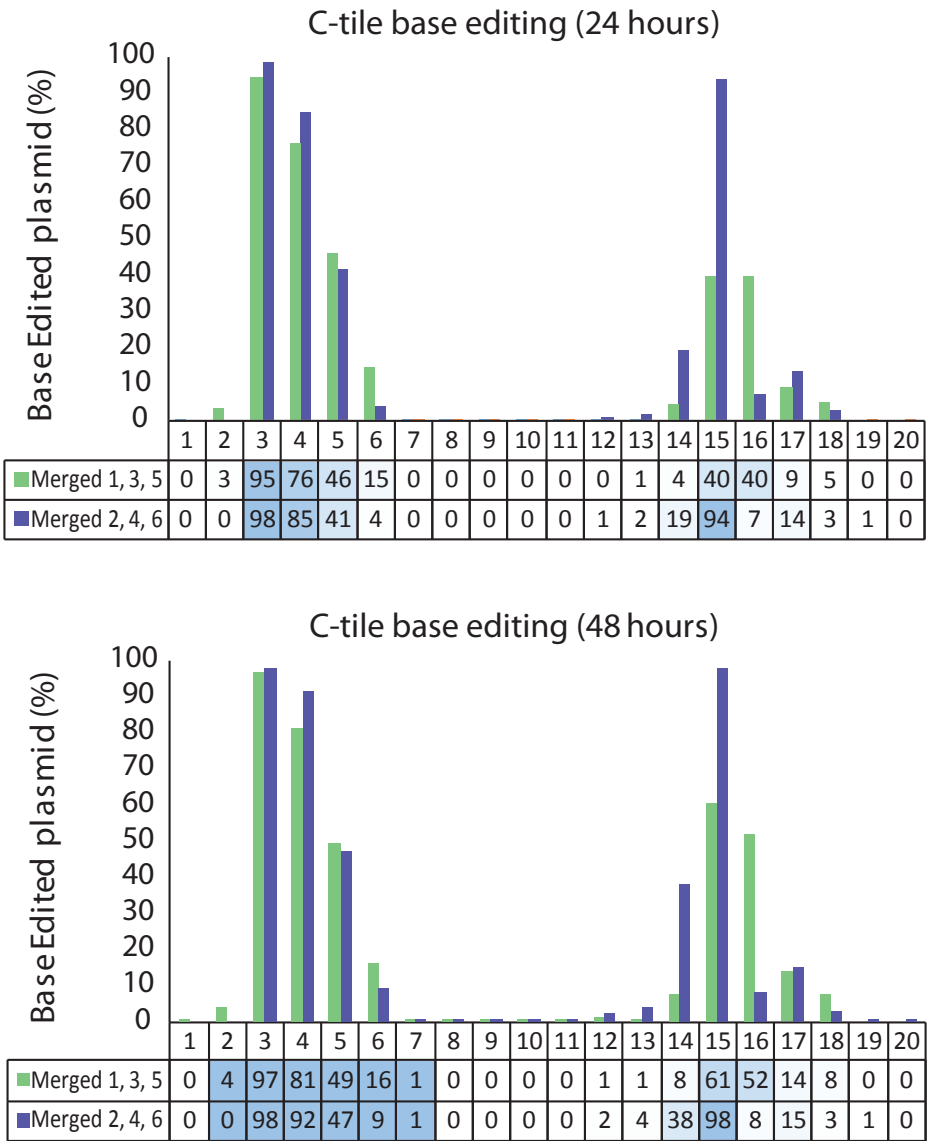


Figure S1 | Base editing C-tile plasmids in *E. coli* using MmuBE_E1 obtained from deep sequencing. Data from C-tile plasmids of uneven ('Merged 1, 3 and 5') and even numbers ('Merged 2, 4, and 6'). Y-axis represent base edited plasmids in % of the whole population and x-axis represent the C position within the protospacer. **(A)** Graphs showing base editing of C-tile plasmids after 24 hours. **(B)** Graphs showing base editing of C-tile plasmids after 48 hours.

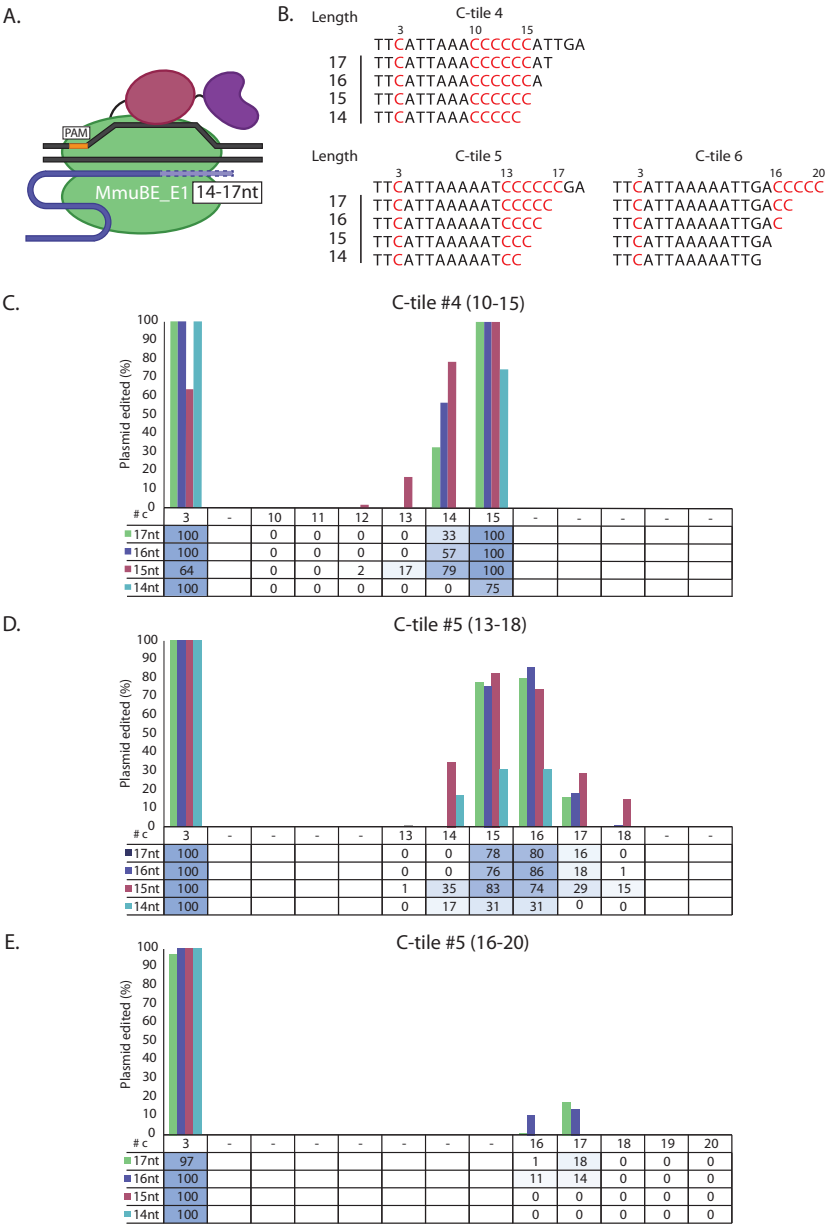


Figure S2 | Effect of spacer length on base editing. (A) Schematics of base editing using MmuBE_E1 and spacers varying in length (14-17 nt). (B, C, D) Sanger sequencing results for different spacer length (14-17 nt) targeting plasmids C-tile 4, 5 and 6, containing C's on position 10-15, 13-18 and 16-20, respectively. Y-axis represents edited plasmids in % of a whole population and x-axis represent the C position on the protospacer. Number of edited plasmids is represented in the table below each graph. C3 was also included as internal standard (B). Graph of edited plasmid from C-tile 4. (C). Graph of edited plasmid from C-tile 5. (D). Graph of edited plasmid from C-tile 6.

Table S2 | Various MmuCas12u1 base editors constructed

Name	MmudCas12u1	Linker	Cytidine deaminase	UGI
<i>E. coli</i> base editor				
MmuBE_E1	<i>E. coli</i>	Sh3 (121 aa)	CDA1 (<i>E. coli</i>)	<i>E. coli</i>
MmuBE_E1.A	<i>E. coli</i>	Sh3 (97 aa)	CDA1 (<i>E. coli</i>)	<i>E. coli</i>
MmuBE_E1.B	<i>E. coli</i>	Sh3 (67 aa)	CDA1 (<i>E. coli</i>)	<i>E. coli</i>
MmuBE_E1.C	<i>E. coli</i>	Sh3 (29 aa)	CDA1 (<i>E. coli</i>)	<i>E. coli</i>
MmuBE_E1.D	<i>E. coli</i>	PA (33 aa)	CDA1 (<i>E. coli</i>)	<i>E. coli</i>
EcMmuBE2	<i>E. coli</i>	Sh3 (93 aa)	rAPOBEC1 (<i>Hsa</i>)	<i>E. coli</i>
EcMmuBE3	<i>E. coli</i>	Sh3 (121 aa)	CDA1 (<i>E. coli</i>)	<i>Hsa</i>
<i>H. sapiens</i> base editor				
MmuBE_H1.A	<i>H. sapiens</i>	Sh3 (121 aa)	CDA1 (<i>Hsa</i>)	<i>Hsa</i>
MmuBE_H1.B	<i>H. sapiens</i>	XTEN (16aa)	CDA1 (<i>Hsa</i>)	<i>Hsa</i>
MmuBE_H2	<i>H. sapiens</i>	XTEN (16aa)	rAPOBEC1 (<i>Hsa</i>)	<i>Hsa</i>
MmuBE_H2YE	<i>H. sapiens</i>	XTEN (16aa)	rAPOBEC-YE (<i>Hsa</i>)	<i>Hsa</i>
<i>S. cerevisiae</i> base editor				
MmuBE_H1.A	(<i>MmuCas12u1</i>)	Sh3 (93 aa)	CDA1 (<i>Hsa</i>)	<i>Hsa</i>

Table S3 | *S. cerevisiae* strains used in the study

Strain name	Genotype	Obtained by transformation with	Origin
CEN.PK113-5D	MATa ura3–52		Euroscarf
YSTB013	MATa ura3–52 pUDE731		This study
YSTB164	MATa ura3–52 INT1::kl11p::eGFP::CYC1t	PL-098, A0135, A0136, A0195 and A0196	This study
YSTB305	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3	A0246, A0247, A0295, A0296 and A0297	This study
YSTB315	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-242	PL-242	This study
YSTB316	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-243	PL-243	This study
YSTB317	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-244	PL-244	This study
YSTB318	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-245	PL-245	This study
YSTB319	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-246	PL-246	This study
YSTB320	MATa ura3–52 INT1::kl11p::eGFP::CYC1t, INT2::TEF1p::MmuCas12u1::CYC1t:KIURA3 + PL-139	PL-139	This study

Table S4 | Oligonucleotides used in the study

oligo ID	sequence (5'-3')	description
Construction of RFP-UGI entry plasmid		
BG14064	GATGTCCTCCTGAGCTCGC	Rv pCas_PAMSCNR
BG14065	AAGCTTGGCTGTTTGGCG	Fw pCas_PAMSCNR
BG19000	ACGAGCTGTACAAGACTAGTCCCAAGAGAAACGGAAAGT	Fw Sv40 NLS
BG19001	CGCCAAAACAGACCAAGCTTTTAGACTTTCTCTTCTTCTTG	Rv Sv40 NLS2
BG19002	GGCAGCTCAGGAGGACCATATGGTGCTAAAGGGCGAAGAG	Fw Ndel mRuby
BG19003	ACTAGTCTTTGTACAGCTCGTCCATGC	Rv Spel mRuby
BG19102	CCCAAGAGAAACGGAAAGT	Fw2 Sv40 NLS
BG19104	CACTTTCCGTTTCTTCTTGGGAAGCTTCGTTAAGCACCGGTGGAGTG	Rv BbsI - RFP
BG19118	GGCAGCTCAGGAGGACATCTTGTCTTCTTGTGACAATTAAATCATCGGCTC	Fw BbsI - pTaq_RFP
Construction of base editor plasmids (MmuBE_E)		
BG14065	AAGCTTGGCTGTTTGGCG	Fw pCas Flank R
BG15295	GGGGTTCGAGGGGGCAGTTG	Rv E. coli Mmu(CDA flank)
BG15296	CAACTGCCCCCTCGAACCCCGGTGGAGGAGGTTCTGGAG	Fw CDA (mmu flanks)
BG15297	CGCCAAAACAGCAAGCTTTTATGCAACCAGTCTTAGCATC	Rv UGI (mmu flanks)
BG20896	ACACGCTCTTCTATGACCGACGCTGAGTACGTG	Fw SapI MmuBE1 vector
BG20897	ACACGCTCTTCTGGGGTTTCGAGGGGCGAGTTG	Rv SapI MmuBE1 vector
BG20898	ACACGCTCTTCTCCCGCTGCACAGCTCTGCTCCAGCACCTGCTCCAGCTCCAGCACCT GCACCTGCACACAGCTCCAGCACTGCTCCAGCTCTGCTCCT	Fw Pa33 Linker
BG20899	ACACGCTCTTCTTCATAGCAGCAGGAGCTGGAGCAGGTGCTGGAGCTGGTGCAGGTCGAGCTGG ACGAGGTGCTGGAGCAGGTGCTGGAGCAGGAGCTGGTCAGG	Rv Pa33 Linker
BG20900	ACACGCTCTTCTCCCTCCGGAGACTATAAGGACCAC	Fw 29aa SH3_kinase Linker
BG20901	ACACGCTCTTCTTCATGGACTCGAGCCTAGACTTATC	Rv 29aa SH3_kinase Linker
BG20902	ACACGCTCTTCTCCCGGTGGAGGAGGTTCTGGAGG	Fw 67aa/97aa SH3_kinase Linker
BG20903	ACACGCTCTTCTTCATATATATCTTCTCCACGTAAAGGAC	Rv 67aa SH3_kinase Linker

BG20904	ACACGCTCTTCTTTCAATTCGGGACTCGAGCCTAGACTT	Rv 97aa SH3_kinase Linker
BG20905	ACACGAAGACTTTTCATCATGACAACAATGACAGTACATAC	Fw BbsI dMmu (MmuBE_E2)
BG20906	ACACGAAGACTTTTCATCTCTAGACTTATCGTCATCG	Rv BbsI dMmu (MmuBE_E2)
BG20907	ACACGAAGACTTTATGAATGAGCTCAGAGACTGGCCC	Fw HsaApobec (MmuBE_E2)
BG20908	ACACGAAGACTTTTCATCATGACAACAATGACAGTACATAC	Fw dMmu-CDA (MmuBE_E3)
BG20909	ACACGAAGACAATGGGAACAGCAGGACTCTTTAGTGG	Rv dMmu-CDA (MmuBE_E3)
Construction of base editor plasmids (MmuBE_H)		
BG21058	CAAAGCAGATGACGATAAGTCTAGGATGACAGACGCCGAGTACGTG	Fw Backbone HsaMmu-CDA
BG21059	CTCCACCTCCAGAACCTCCTCCACCCGGATTACTCGTGCCGTGG	Rv Backbone HsaMmu-CDA
BG21060	CCAGGCACCGAGTAATCGGGTGGAGGAGTTCTTGGAGG	Fw Linker Gibson hsa Mmu-BE
BG21061	CACGTACTCGCGCTCTGTCAAT	Rv Linker Gibson hsa Mmu-BE
BG19700	ACACGAAGACTTTTCATCATGACCACCATGACCGTGAC	Fw BbsI dMmu
BG19702	ACACGAAGACAACACTGAGGTC CCGGGAGTCTCGCTGCCCGGATTACTCGGTGCCGTGG	Rv BbsI XTEN dMmu
BG19703	ACACGAAGACTTTTCAGAGTCGCCACACCCGAAAGTATGAGCTCAGAGACTGGCCCAAG	Fw BbsI XTEN-APOBEC(YE)
BG19704	ACACGAAGACAATGGGTTTCAACCCGGTGGCCGAG	Rv BbsI XTEN-APOBEC(YE)
BG19709	ACACGAAGACTTTTCAGAGTCGCCACACCCGAAAGTATGACAGACGCCGAGTACGTG	Fw BbsI Xten-CDA
BG19710	ACACGAAGACAATGGGACCACACCGCTGGAGACTTATGTG	Rv BbsI CDA
BG21160	AGCGGCAGCGAGACTCCC	Fw repair MmuBE_H(1.B/2/YE)
BG21161	CGGATTACTCGGTGCCGTGG	Rv repair MmuBE_H(1.B/2/YE)
Construction of GFP silencing guide		
BG20003	AGACTTGAATTAGATGGTGTGT	Fw spacer GFP Silence
BG20004	ACACAACATCACCATCTAATTCAA	Rv spacer GFP Silence
Construction of C-tile plasmids		
BG15892	GATCGGTACCCATTAACCTATAAAAATAGGCGTC	Fw CTTA-pPAM-SCNR KpnI
BG15893	CAATTTTAAATGAACGGAGCTCG	Rv CTTA-pPAM-SCNR AatI
BG15894	CTTATTCAATTAATAATTGAATTGAGGTAC	Fw protospacer C-tile (WT)
BG15895	CTCAATTCAATTTTAAATGAATAAGACGT	Rv protospacer C-tile (WT)

BG15896	CTTACCCCCCAAAAATTGAA TTGAGGTAC	Fw protospacer C-tile (1-6)
BG15897	CTCAATTCAA TTTTGTGGGGGTAAGACGT	Rv protospacer C-tile (1-6)
BG15898	CTTATTCCCCCCAATTGAA TTGAGGTAC	Fw protospacer C-tile (4-9)
BG15899	CTCAATTCAA TTGGGGGGGAATAAGACGT	Rv protospacer C-tile (4-9)
BG15900	CTTATTCAATCCCCCTGAA TTGAGGTAC	Fw protospacer C-tile (7-12)
BG15901	CTCAATTCAGGGGGAATCAATAAGACGT	Rv protospacer C-tile (7-12)
BG15902	CTTATTCAATTAACCCCCCA TTGAGGTAC	Fw protospacer C-tile (10-15)
BG15903	CTCAATGGGGGTTTAATCAATAAGACGT	Rv protospacer C-tile (10-15)
BG15904	CTTATTCAATAAAAATCCCCCGAGGTAC	Fw protospacer C-tile (13-18)
BG15905	CTCGGGGGA TTTTAAATCAATAAGACGT	Rv protospacer C-tile (13-18)
BG15906	CTTATTCAATTAATAATTGACCCCGGTAC	Fw protospacer C-tile (17-20)
BG15907	CGGGGGTCAA TTTTAAATCAATAAGACGT	Rv protospacer C-tile (17-20)
BG15908	AGACTTCATTAATAATAATTGAA TTGA	Fw spacer C-tile (WT)
BG15909	ACACTCAATTCAAATTTTAA TGAA	Rv spacer C-tile (WT)
BG15910	AGACCCCCCAAAAATTGAA TTGA	Fw spacer C-tile (1-6)
BG15911	ACACTCAATTCAAATTTTGGGGG	Rv spacer C-tile (1-6)
BG15912	AGACTTCCCCCCAATTGAA TTGA	Fw spacer C-tile (4-9)
BG15913	ACACTCAATTCAAATTTGGGGGAA	Rv spacer C-tile (4-9)
BG15914	AGACTTCATTCCTCCCTGAA TTGA	Fw spacer C-tile (7-12)
BG15915	ACACTCAATTCAGGGGGAA TGAA	Rv spacer C-tile (7-12)
BG15916	AGACTTCATTAACCCCTCA TTGA	Fw spacer C-tile (10-15)
BG15917	ACACTCAATGGGGGTTAA TGAA	Rv spacer C-tile (10-15)
BG15918	AGACTTCATTAATAATCCCCCGA	Fw spacer C-tile (13-18)
BG15919	ACACTCGGGGGATTTTAA TGAA	Rv spacer C-tile (13-18)
BG15920	AGACTTCATTAATAATTGACCCCC	Fw spacer C-tile(16-20)
BG15921	ACACGGGGTCAATTTTAA TGAA	Rv spacer C-tile (16-20)
BG18520	AGACTTCATTAATAATTG	Fw spacer 14 C-tile (WT)

BG18521	ACACCAATTTTAAATGAA	Rv spacer 14 C-tile (WT)
BG18522	AGACTTCATTAAAAATTGA	Fw spacer 15 C-tile (WT)
BG18523	ACACTCAATTTTAAATGAA	Rv spacer 15 C-tile (WT)
BG18524	AGACTTCATTAAAAATTGAA	Fw spacer 16 C-tile (WT)
BG18525	ACACTTCATTTTAAATGAA	Rv spacer 16 C-tile (WT)
BG18526	AGACTTCATTAAAAATTGAAT	Fw spacer 17 C-tile (WT)
BG18527	ACACATTCATTTTAAATGAA	Rv spacer 17 C-tile (WT)
BG18528	AGACTTCATTAAACCCCC	Fw spacer 14 C-tile (10-15)
BG18529	ACACGGGGTTTAAATGAA	Rv spacer 14 C-tile (10-15)
BG18530	AGACTTCATTAAACCCCC	Fw spacer 15 C-tile (10-15)
BG18531	ACACGGGGTTTAAATGAA	Rv spacer 15 C-tile (10-15)
BG18532	AGACTTCATTAAACCCCCA	Fw spacer 16 C-tile (10-15)
BG18533	ACACTGGGGTTTAAATGAA	Rv spacer 16 C-tile (10-15)
BG18534	AGACTTCATTAAACCCCCCAT	Fw spacer 17 C-tile (10-15)
BG18535	ACACATGGGGTTTAAATGAA	Rv spacer 17 C-tile (10-15)
BG18536	AGACTTCATTAAAAATCC	Fw spacer 14 C-tile (13-18)
BG18537	ACACGGATTTTAAATGAA	Rv spacer 14 C-tile (13-18)
BG18538	AGACTTCATTAAAAATCCC	Fw spacer 15 C-tile (13-18)
BG18539	ACACGGGATTTTAAATGAA	Rv spacer 15 C-tile (13-18)
BG18540	AGACTTCATTAAAAATCCC	Fw spacer 16 C-tile (13-18)
BG18541	ACACGGGGATTTTAAATGAA	Rv spacer 16 C-tile (13-18)
BG18542	AGACTTCATTAAAAATCCCC	Fw spacer 17 C-tile (13-18)
BG18543	ACACGGGGATTTTAAATGAA	Rv spacer 17 C-tile (13-18)
BG18544	AGACTTCATTAAAAATTGAC	Fw spacer 16 C-tile (16-20)
BG18545	ACACGTCAATTTTAAATGAA	Rv spacer 16 C-tile (16-20)
BG18546	AGACTTCATTAAAAATTGACC	Fw spacer 17 C-tile (16-20)
BG18547	ACACGGTCAATTTTAAATGAA	Rv spacer 17 C-tile (16-20)

CHAPTER 7

BG14555	ATCTACACAGTAGAAATTGATCAT	Rv for amplification of A0118 [linear PL-074 with INT1 spacer]
BG16700	CGCCGCGCGATGCCGCCAAACGTCCTGG	Fw for amplification of A0185 [Mmurepeat-BsaXI restriction sites-Mmurepeat]
BG14156	GGTGAATGTCGCGATATAGG	Rv for amplification of A0185 [Mmurepeat-BsaXI restriction sites-Mmurepeat]
BG17373	CACATAGCAATCTGGCTATATG	Fw for amplification of A0135 [upstream homologous region of INT1]
BG17374	AAACGCCCTGTGGTGTTGCTACTGGATATGCAAAAGCATTTGAAGTCGCTTGACTCCTCTGCCGTCAATTC	Rv for amplification of A0135 [upstream homologous region of INT1 + connector5]
BG17375	TTGCCCATCGAACGTAACAAGTACTCCTCTGTTCTCTCCTTTGCTTTAAAGCGTTGAAGTTTCTCTTTTG	Fw for amplification of A0136 [ConnectorA + downstream homologous region of INT1]
BG17376	TGTCAACTGGAGAGCTATCG	Rv for amplification of A0136 [downstream homologous region of INT1]
BG18470	AGACGGCTTGAGTTGGTGTATGCG	Fw NT crRNA in A0154
BG18471	ACACGCCATACACCAACTCAACGC	Rv NT crRNA in A0154
BG19774	AAACGACTTCCCAATCGCTTTTGGCATAATCCAGTACCAACACACAGGCGTTTCTTTTTCGCGTCACCCCC	Fw for amplification of A0195 [connector5-K11p-eGFP]
BG19775	TGAAAGTTCTTCTCCTTTGCTCATTTTGTGATAAGTATTTAAAGCGAGTCACTGAA	Rv for amplification of A0195 [connector5-K11p-eGFP]
BG19776	TCACTCGCTTAATACTTATCAAAAATGAGCAAGAGAGAACTTTTCACTGG	Fw for amplification of A0196 [eGFP-connectorA]
BG17378	AAAGCAAGGAAGGAGAGAAACAGAGAGTACTTGTACGTTTCGATGGCAACTTCGAGCGTCCCAAAACCTTC	Rv for amplification of A0196 [eGFP-connectorA]
BG20455	ACGGCTGCTCCCGGTAGC	Fw for amplification of A0250 [backbone_1 of PL-163]
BG13880	TGCTTCATTTTGTAGAACCAAAAATG	Rv for amplification of A0250 [backbone_1 of PL-163]
BG19061	AGACACCGTGCATACCGCTGCTCCCGGTAGC	Fw for integration of BsaXI restriction site in PL-162
BG19062	ACACGCTACGGGAGCAGCGGTATGCACGGT	Rv for integration of BsaXI restriction site in PL-162
BG13879	CATTTTGTGTTTACAAAAATGAAGCA	Fw for amplification of A0251 [backbone_2 of PL-163]
BG20456	CRACTTATATCGTATGGACACGGGTATGCACGGT	Rv for amplification of A0251 [backbone_2 of PL-163]
BG20457	GCTACGGGAGCAGCGGTAAAAAACCCCGCGAAGCG	Fw for amplification of A0252 [RFP fragment 1]
BG20239	GGACATCACCTCTCAACAGAAAGATT	Rv for amplification of A0252 [RFP fragment 1]
BG20240	AATCTTCGTTGTGAGAGGTGATGTCC	Fw for amplification of A0253 [RFP fragment 2]
BG20514	CGGCTCTCCATACGATATAAGTTGTAATTTCGTACCCCG	Rv for amplification of A0253 [RFP fragment 2]
BG20455	ACCGCTGCTCCCGGTAGC	Fw for amplification of A0250 [backbone PL-074]
BG13880	TGCTTCATTTTGTAGAACCAAAAATG	Rv for amplification of A0250 [backbone PL-074_1]

BG13879	CATTTTGTCTACAAATGAAGCA	Fw for amplification of A0251 [backbone PL-074_2]
BG20456	CAACTTATATCGTATGGAGCAGCGGTATGCACGGT	Rv for amplification of A0251 [backbone PL-074_2]
BG20457	GTTACGGGAGCAGCGGTAAAAACCCCGCCGAAAGCG	Fw for amplification of A0252 [RFPcassette_1]
BG20239	GGACATCACTCTCTCAACAAGAGATT	Rv for amplification of A0252 [RFPcassette_1]
BG20240	AATCTTCGTTGTGAGAGGTGATGTC	Fw for amplification of A0253 [RFPcassette_2]
BG20514	CCGCTGCTCCATACGATATAAGTTGTAATTCGGTACCCCG	Rv for amplification of A0253 [RFPcassette_2]
BG20383	CGGATCGATGTACACAACGACTGCACCAACGAAACAAATCTTAGCATCATAGCTTCAAAA TGTTTCTACT	Fw for amplification of A0295 and A0245 [MmuCas12u1]
BG21049	GCGGTACCTCTCTCACCCACCTTCTCTTCTTCTTCTTGGGCTTTTCTTTTGGCCGTGGC	Rv for amplification of A0295 [MmuCas12u1]
BG21050	GGCCAGGCAAAAAGAAAAAGCCCAAGAAAGAGGAAGGT	Fw for amplification of A0296 [CDA-UGI]
BG21051	GGCGGTGAATGTAAAGCGTGACATACTAATTTATAGCATCTTGAATCTTGTTT	Rv for amplification of A0296 [CDA-UGI]
BG21052	AATGGAGAAACAAGATCAAGATGCTATAAAATTAGTTATGTCACGCTTACAT	Fw for amplification of A0297 [KIURA3]
BG20390	CAACAGGAGCGGATGGATATATCTGTGGTC TCGAAGATGCCGGAAGCGTGATCCCAATACAACAGATCAC	Rv for amplification of A0297 and A0248 [KIURA3]
BG20384	CGATGAACCTTAATTAACAGAGCTCAAAATTAAGCCCTTCGAGCG	Rv for amplification of A0245 [MmuCas12u1]
BG20385	AGAAGATTTCTCTTCAATCTCCT	Fw for amplification of A0246 (INT2-5')
BG20386	TGCTAAGATTTGTGTTCTGTTTGGGTGCAGTCGGTTGTGTACATCGATCCGCCCTTATCAAGGATACCTGG	Rv for amplification of A0246 (INT2-5')
BG20387	ACGCTTTCGGGATCTTCCAGACCAAGATATATCCATCCGCTCTGTTGGGGCGATTACACAAGCG	Fw for amplification of A0247 (INT2-3')
BG20388	TCTCC TCTTTCGATGACC	Rv for amplification of A0247 (INT2-3')
BG20389	TGGGACGCTCGAAGGCTTTAATTTGAGCTCGTTTATTTAGGTTCTATC	Fw for amplification of A0248
BG21235	AAGATGGTTCCGTTCAACTAGTG	Fw for crRNA targeting eGFP15
BG21236	TAGTTGAACGGAACCATCTTGTC	Rv for crRNA targeting eGFP15
BG21237	CCCAAAGTTCCTGTGGAAATG	Fw for crRNA targeting ade2.3.1
BG21238	TTTCCACAGAAACACTTTGGGGTC	Rvfor crRNA targeting ade2.3.1
BG21239	TGCAATGCCTAGAGGTGTTGTG	Fw for crRNA targeting ade2.3.2
BG21240	AACACCTTAGGCATTTGCAGTC	Rv for crRNA targeting ade2.3.2
BG21241	ATACAAGACAAATATATTCAATG	Fw for crRNA targeting ade2.4.1
BG21242	TGAATATATTTGTCTTGTAATGTC	Rv for crRNA targeting ade2.4.1
BG21243	AAGCAAGAAGAAGATTTCTGTG	Fw for crRNA targeting ade2.4.2
BG21244	AGAACTTCTTCTTCTTGCTTGTC	Rv for crRNA targeting ade2.4.2

Table S5 | plasmids and fragments used in the study

Name	Description	Source
<i>E. coli</i>		
pCMV-BE3	Cas9-APOBEC BE3 under CMV promoter	addgene #73021
pSI-Target-AID-NG	Cas9-CDA TargetAID	addgene #119861
pScI_dCas9-CDA-UL	Prokaryotic Cas9 Base editor	addgene #108551
pCMV-dCpf1-BE	Cas12a-APOBEC base editor	addgene #107685
pCMV-dCpf1-BE-YE	Cas12a-APOBEC(YE) base editor	addgene #107686
pCas-dMmu	PJ23108-MmudCas12u1 (<i>E. coli</i> harmonized)	chapter 6
pCas-mRuby-UGI-Entry	mRuby flanked by BbsI restriction sites for cloning fusion proteins to UGI	this study
pCas-RFP-UGI-Entry	RFP flanked by BbsI restriction sites for cloning fusion proteins to UGI	this study
pCas-MmuBE_E1	PJ23108-MmudCas12u1-CDA-UGI (121 aa SH3 linker)	this study
pCas-MmuBE_E1.A	PJ23108-MmudCas12u1-CDA-UGI (96 aa SH3 linker)	this study
pCas-MmuBE_E1.B	PJ23108-MmudCas12u1-CDA-UGI (67 aa SH3 linker)	this study
pCas-MmuBE_E1.C	PJ23108-MmudCas12u1-CDA-UGI (24 aa SH3 linker)	this study
pCas-MmuBE_E1.D	PJ23108-MmudCas12u1-CDA-UGI (33 aa PAPA rigid linker)	this study
pCas-MmuBE_E2	PJ23108-MmudCas12u1-HsaAPOBEC-UGI (93 aa SH3 linker)	this study
pCas-MmuBE_E3	PJ23108-MmudCas12u1-CDA-HsaUGI (121 aa Sh3linker)	this study
pCas-MmuBE_H1.B	PJ23108-HsaMmudCas12u1-HsaCDA-HsaUGI (121 aa SH3 linker)	this study
pCas-MmuBE_H1.A	PJ23108-HsaMmudCas12u1-HsaAPOBEC-HsaUGI (16 aa XTEN linker)	this study
pCas-MmuBE_H2	PJ23108-HsaMmudCas12u1-HsaAPOBEC-HsaUGI (16 aa XTEN linker)	this study
pCas-MmuBE_H2YE	PJ23108-HsaMmudCas12u1-HsaAPOBEC(YE)-HsaUGI (121 aa Sh3linker)	this study
pCRISPR-Mmu-NT (BbsI)	PJ23119-CRISPR array (repeat-spacer-repeat). 30 nt non-targetting spacer flanked by BbsI	this study
pCRISPR-Mmu-NT	PJ23119-CRISPR array: non-targetting spacer (20nt)	this study
pCRISPR-Mmu-GFP	PJ23119-CRISPR array: GFP spacer	this study
pCRISPR-Mmu-C-tile (WT)	PJ23119-CRISPR array: C-tile (WT) spacer	this study
pCRISPR-Mmu-C-tile (1-6)	PJ23119-CRISPR array: C-tile (1-6) spacer	this study
pCRISPR-Mmu-C-tile (4-9)	PJ23119-CRISPR array: C-tile (4-9) spacer	this study
pCRISPR-Mmu-C-tile (7-12)	PJ23119-CRISPR array: C-tile (7-12) spacer	this study
pCRISPR-Mmu-C-tile (10-15)	PJ23119-CRISPR array: C-tile (10-15) spacer	this study
pCRISPR-Mmu-C-tile (13-18)	PJ23119-CRISPR array: C-tile (13-18) spacer	this study
pCRISPR-Mmu-C-tile(16-20)	PJ23119-CRISPR array: C-tile (16-20) spacer	this study
pCRISPR-Mmu-14 C-tile (WT)	PJ23119-CRISPR array: 14 nt C-tile (WT) spacer	this study
pCRISPR-Mmu-15 C-tile (WT)	PJ23119-CRISPR array: 15 nt C-tile (WT) spacer	this study
pCRISPR-Mmu-16 C-tile (WT)	PJ23119-CRISPR array: 16 nt C-tile (WT) spacer	this study
pCRISPR-Mmu-17 C-tile (WT)	PJ23119-CRISPR array: 17 nt C-tile (WT) spacer	this study
pCRISPR-Mmu-14 C-tile (10-15)	PJ23119-CRISPR array: 14 nt C-tile (10-15) spacer	this study
pCRISPR-Mmu-15 C-tile (10-15)	PJ23119-CRISPR array: 15 nt C-tile (10-15) spacer	this study
pCRISPR-Mmu-16 C-tile (10-15)	PJ23119-CRISPR array: 16 nt C-tile (10-15) spacer	this study

Name	Description	Source
<i>E. coli</i>		
pCRISPR-Mmu-17 C-tile (10-15)	PJ23119-CRISPR array: 17 nt C-tile (10-15) spacer	this study
pCRISPR-Mmu-14 C-tile (13-18)	PJ23119-CRISPR array: 14 nt C-tile (13-18) spacer	this study
pCRISPR-Mmu-15 C-tile (13-18)	PJ23119-CRISPR array: 15 nt C-tile (13-18) spacer	this study
pCRISPR-Mmu-16 C-tile (13-18)	PJ23119-CRISPR array: 16 nt C-tile (13-18) spacer	this study
pCRISPR-Mmu-17 C-tile (13-18)	PJ23119-CRISPR array: 17 nt C-tile (13-18) spacer	this study
pCRISPR-Mmu-16 C-tile (16-20)	PJ23119-CRISPR array: 16 nt C-tile (16-20) spacer	this study
pCRISPR-Mmu-17 C-tile (16-20)	PJ23119-CRISPR array: 17 nt C-tile (16-20) spacer	this study
pCRISPR-Mmu-C-motif_1	PJ23119-CRISPR array: C-motif_1 spacer	this study
pCRISPR-Mmu-C-motif_2	PJ23119-CRISPR array: C-motif_2 spacer	this study
pCRISPR-Mmu-C-motif_3	PJ23119-CRISPR array: C-motif_3 spacer	this study
pTarget-divergent	pTaq-RFP and pLacIq-GFP (divergent expression)	chapter 6
pTarget-GFP	pLacIq-GFP	chapter 6
pTarget-C-tile (WT)	pTarget-GFP containing C-tile (WT) spacer	this study
pTarget-C-tile (1-6)	pTarget-GFP containing C-tile (1-6) spacer	this study
pTarget-C-tile (4-9)	pTarget-GFP containing C-tile (4-9) spacer	this study
pTarget-C-tile (10-15)	pTarget-GFP containing C-tile (10-15) spacer	this study
pTarget-C-tile (13-18)	pTarget-GFP containing C-tile (13-18) spacer	this study
pTarget-C-tile (16-20)	pTarget-GFP containing C-tile (16-20) spacer	this study
pTarget-Mmu-C-motif_1	pTarget-GFP containing C-motif_1 spacer	this study
pTarget-Mmu-C-motif_2	pTarget-GFP containing C-motif_2 spacer	this study
pTarget-Mmu-C-motif_3	pTarget-GFP containing C-motif_3 spacer	this study
<i>S. cerevisiae</i>		
pCfB2791	integrative plasmid - Ty4Cons PTEF1::GFP KI.URA3	addgene #63654
pCSN068	CEN/ARS4 ampR KanMX TRP1 KI11p::Fncpf1::GND2t	addgene #101749
pUDE731	2µm ampR KIURA3 TEF1p::Fncpf1::CYC1t	addgene #103008
pUD628	2µm KanMX ampR SNR52p::Cas12aRP::crADE2-3.S::SUP4t	addgene #103018
PL-074	2µm KanMX ampR SNR52p:: Cas12aRP::crADE2-3.S::SUP4t	this study
PL-098	2µm KanMX ampR SNR52p:: Cas12aRP::crINT1::SUP4t	this study
PL-162	PJ23119-CRISPR array: BsaXI spacer	this study
PL-163	2µm KanMX ampR SNR52p::MmuRP:: crINT1::MmuRP::SUP4t	this study
PL-196	2µm KanMX ampR SNR52p::MmuRP::blap-RFP-t::MmuRP::SUP4t	this study
PL-242	2µm KanMX ampR SNR52p::MmuRP::eGFP.15::MmuRP::SUP4t	this study
PL-243	2µm KanMX ampR SNR52p::MmuRP::ADE2.3.1::MmuRP::SUP4t	this study
PL-244	2µm KanMX ampR SNR52p::MmuRP::ADE2.3.2::MmuRP::SUP4t	this study
PL-245	2µm KanMX ampR SNR52p::MmuRP::ADE2.4.1::MmuRP::SUP4t	this study
PL-246	2µm KanMX ampR SNR52p::MmuRP::ADE2.4.2::MmuRP::SUP4t	this study
PL-138	PJ23119-CRISPR array: <i>S. cerevisiae</i> non-target spacer	this study
PL-139	2µm KanMX ampR SNR52p::MmuRP::NT::MmuRP::SUP4t	this study

Table S6 | Cloning strategy of *E. coli* plasmids

Name	Cloning strategy	Template
pCas-mRuby-UGI-Entry	NEBuilder® HiFi DNA Assembly (NEB)	mRuby
		pCMV-BE3 (addgene #73021)
		pCas (chapter 6)
pCas-RFP-UGI-Entry	NEBuilder® HiFi DNA Assembly (NEB)	pCas-mRuby-UGI-Entry
		RFP
pCas-MmuBE_E1	NEBuilder® HiFi DNA Assembly (NEB)	pCas-dMmu (chapter 6)
		pScl_dCas9-CDA-UL (addgene #108551)
pCas-MmuBE_E1.A	Golden Gate (SapI)	pCas-MmuBE_E1
		pCas-MmuBE_E1
pCas-MmuBE_E1.B	Golden Gate (SapI)	pCas-MmuBE_E1
		pCas-MmuBE_E1
pCas-MmuBE_E1.C	Golden Gate (SapI)	pCas-MmuBE_E1
		pCas-MmuBE_E1
pCas-MmuBE_E1.D	Golden Gate (SapI)	pCas-MmuBE_E1
		-
pCas-MmuBE_E2	Golden Gate (BbsI)	pCas-RFP-UGI-Entry
		pCas-dMmu (chapter 6)
		pCMV-dCpf1-BE (addgene #107685)
pCas-MmuBE_E3	Golden Gate (BbsI)	pCas-RFP-UGI-Entry
		pCas-MmuBE_E1
pCas-MmuBE_H1.B	Golden Gate (BbsI)	pCas-RFP-UGI-Entry
		pCas-dMmu (chapter 6)
		pScl_dCas9-CDA-UL (addgene #108551)
pCas-MmuBE_H1.A	NEBuilder® HiFi DNA Assembly (NEB)	pCas-MmuBE_H1.B
		pCas-MmuBE_H1.B
		pCas-MmuBE_E1
pCas-MmuBE_H2	Golden Gate (BbsI)	pCas-RFP-UGI-Entry
		pCas-dMmu (chapter 6)
		pCMV-dCpf1-BE (addgene #107685)
pCas-MmuBE_H2YE	Golden Gate (BbsI)	pCas-MmuBE_H1.B
		pCas-RFP-UGI-Entry
		pCas-dMmu (chapter 6)
		pCMV-dCpf1-BE-YE (addgene #107686)
	PCR and blunt-end ligation*	pCas-MmuBE_H1.B

Description
PCR <i>mruby</i> using BG19002 and BG19003
PCR UGI using BG19000 and BG19001
PCR vector using BG14064 and BG14065
PCR vector using BG14064 and BG19102
PCR <i>rfp</i> using BG19104 and BG19118
PCR vector using BG14065 and BG15296
PCR CDA-UGI using BG15296 and BG15297
PCR vector using BG20896 and BG20897
PCR linker using BG20902 and BG20904
PCR vector using BG20896 and BG20897
PCR linker using BG20902 and BG20903
PCR vector using BG20896 and BG20897
PCR linker using BG20900 and BG20901
PCR vector using BG20896 and BG20897
anneal oligo's BG20898 and BG20899 to create Pa33 linker
digest vector with BbsI
PCR dMmu using BG20905 and BG20906
PCR APOBEC using BG20907 and BG19704
digest vector with BbsI
PCR hsaUGI with BG20908 and BG20909
digest vector with BbsI
PCR dMmu-XTEN with BG19700 and BG19702
PCR CDA-UGI using BG19709 and BG19710
repair frameshift using BG21160 and BG21161
PCR vector with BG21058 and BG21059
PCR linker with BG21060 and BG21061
digest vector with BbsI
PCR dMmu-XTEN with BG19700 and BG19702
PCR APOBEC using BG19703 and BG19704
repair frameshift using BG21160 and BG21161
digest vector with BbsI
PCR dMmu-XTEN with BG19700 and BG19702
PCR APOBEC-YE using BG19703 and BG19704
repair frameshift using BG21160 and BG21161

Table S7 | Cloning strategy of *S. cerevisiae* plasmids

Name	Cloning strategy	Template
PL-074	PCR and blunt-end ligation	pUD628
PL-098	PCR and blunt-end ligation	PL-074
PL-162	Digestion (BbsI) and ligation (T4 ligase)	pCRISPR_NT Oligo inserts annealed
PL-163	PCR and blunt-end ligation	PL-162 PL-074
PL-196	NEBuilder® HiFi DNA Assembly (NEB)	PL-163 pGuide Cas12a mRFP(b) BbsI entry pGuide Cas12a mRFP(b) BbsI entry
PL-242	Digestion (BsaXI) and ligation (T4 ligase)	PL-196 Oligo inserts annealed
PL-243	Digestion (BsaXI) and ligation (T4 ligase)	PL-196 Oligo inserts annealed
PL-244	Digestion (BsaXI) and ligation (T4 ligase)	PL-196 Oligo inserts annealed
PL-245	Digestion (BsaXI) and ligation (T4 ligase)	PL-196 Oligo inserts annealed
PL-246	Digestion (BsaXI) and ligation (T4 ligase)	PL-196 Oligo inserts annealed
PL-138	Digestion (BbsI) and ligation (T4 ligase)	pCRISPR_NT Oligo inserts annealed
PL-139	Digestion (BsaXI) and ligation (T4 ligase)	PL-138 PL-074

Description
PCR vector using BG14031 and BG14032
PCR vector using BG16236 and BG14555
digested with BbsI
BG19061 & BG19062
PCR CRISPR array with BG16700 and BG14156. Digest with KpnI and BtgZI and make blunt with T4 PNK
PCR backbone with BG12945 and BG16493
PCR with BG20455 and BG13880
PCR with BG13879 and BG20456
PCR with BG20457 and BG20239
PCR with BG20240 and BG20514
digested with BsaXI
BG21235 & BG21236
digested with BsaXI
BG21237 & BG21238
digested with BsaXI
BG21239 & BG21240
digested with BsaXI
BG21241 & BG21242
digested with BsaXI
BG21243 & BG21244
digested with BbsI
BG18470-BG18471
PCR CRISPR array with BG16700 and BG14156. Digest with KpnI and BtgZI and make blunt with T4 PNK
PCR backbone with BG12945 and BG16493

G	T	C	T	A	A	G	A
T	A	C	T	G	T	T	G
A	C	C	A	C	T	A	T
G	T	A	T	C	A	C	A
C	T	A	T	T	A	G	G

A T A A T T T C
T C H A P T E R
C A G C A C G A
A T T A C G A 8
A A A A C T T T

Summary and general discussion

Summary

CRISPR-Cas is an extraordinary prokaryotic adaptive immune system, divided into two classes that each contain three types and a wide variety of subtypes. Each system is unique and has distinct features in the different steps of the adaptive immunity process. Besides being part of a sophisticated adaptive immune system, CRISPR-associated (Cas) proteins have also been applied in groundbreaking technologies ranging from diagnostics to genome editing. Class 2 effector proteins are the most exploited because of their compact structures with multi-functional properties. Amongst the class 2 systems, the more recently discovered type V systems appears to be the most diverse type, with new mechanistic features still to be uncovered. The research described in this thesis focusses on the characterization and subsequent development of applications of type V CRISPR-Cas systems, more specifically types V-A and V-U1.

Chapter 1 introduces CRISPR-Cas by starting with a brief history of some key discoveries, which started with the serendipitous finding of repetitive sequences in the genomes of bacteria interspaced by variable DNA fragments. This variable DNA appeared to correspond to phage DNA and functions as a “memory bank” that allows for targeting of invading phages (and other mobile genetic elements), making CRISPR-Cas a sophisticated prokaryotic adaptive immune system. CRISPR-Cas systems are divided into classes, types and subtypes. Each CRISPR-Cas system has unique features, but all actively participate in adaptive immunity through a three-step mechanism: adaptation, expression and interference. Although all variants follow the same steps, the different CRISPR-Cas (sub)types are highly diverse with unique structural and functional features at all levels of the mechanism. A more recently discovered type of CRISPR-Cas systems, is type V. Type V consists of eleven characterized subtypes A to K and four uncharacterized subtypes U1-U4.

After describing the underlying molecular mechanism of CRISPR-Cas, an overview is given in **chapter 2** on the different genome editing applications. The main focus is on DNA-targeting class 2 effector proteins, such as Cas9, Cas12a and Cas12b. Despite their distant similarities, these proteins differ in both structural and functional features. Apart from natural variants, engineered CRISPR-Cas nuclease variants that increase editing precision or regulate nuclease activity were presented as well. One of the biggest bottle necks of genome editing in eukaryotes is the delivery of a specific nuclease. Therefore, different approaches of nuclease/guide delivery were discussed. After successful DNA cleavage at a selected genomic location, different host repair systems may be involved in repairing the DNA damage, which may result in different types of genome editing. Various host repair pathways were examined based on the type of DNA damage and the type of repair template available. This chapter provides the information required to make an informed decision regarding nuclease variant, delivery type and repair system for a given application.

In **chapter 3**, the first step of the CRISPR-Cas adaptive immunity, adaptation of the CRISPR memory, is studied in two type V systems, namely type V-A and V-B from *Francisella novicida tularensis* subsp. *novicida* U112 and *Alicyclobacillus acidoterrestris* ATCC 49025, respectively. The type V-A locus encodes Cas12a, Cas4, Cas1 and Cas2, whereas that of V-B encodes Cas12b, a Cas4/1 fusion protein and Cas2. This chapter describes the study of the Cas proteins that are required and their role in the adaptation process. The CRISPR-Cas loci were overexpressed in *Escherichia coli* and adaptation was detected by PCR amplification of the CRISPR array. Spacers were then extracted and analyzed by deep sequencing. After failing to find the previously established T-rich 5'-PAM from spacers obtained in a large-scale experiment, it was realized that truncated Cas4 proteins were being expressed in both systems. This truncation was due to a mutation found in the genome, which led to an incorrect prediction of the Cas4 open reading frame. After correcting the sequences of both Cas4 proteins, they were used in smaller scale adaptation study for both V-A and V-B. It was found that in type V-A, only Cas1 and Cas2 are required for adaptation, and in type V-B, Cas4/1 and Cas2 are required for adaptation, but Cas4 activity is dispensable. Spacers acquired without a functional Cas4, appeared to target protospacers containing mostly non-conical PAMs. Thus, Cas4 activity is required for PAM selection and acquisition of suitable spacers in both type V-A and V-B. The role of Cas12a in the adaptation process has not been elucidated yet in this chapter but will be addressed in future studies.

Following adaptation, **chapter 4** describes crRNA maturation in type V-A. Type V-A crRNA maturation is distinct to that found in type II, where Cas9 requires both a crRNA and a tracrRNA, and gets processed by endogenous RNaseIII after ribonucleoprotein complex formation. This study demonstrated that Cas12a does not require a tracrRNA nor RNase III for crRNA maturation. Instead Cas12a itself is able to process pre-crRNA into mature crRNA using a previously unknown RNase domain found in Cas12a. Cas12a cleaves pre-crRNA by recognition of secondary structures found on repeat of the pre-crRNA. More specifically, Cas12a recognizes specific nucleotides in the 5' end of the repeat, just upstream the stem loop forming sequence. Having Cas12a able to process its own crRNA is greatly advantageous for genome editing applications. It allows for simple simultaneous multi-gene (multiplex) editing using a single CRISPR array. Using a single CRISPR array containing four spacers, Cas12a was able to simultaneously edit up to four genes in mammalian cells (*ex vivo*) and up to three genes in mouse brain cells (*in vivo*).

Apart from processing its own pre-crRNA, another distinct feature of Cas12a is the generation of staggered ends after cleavage of dsDNA. These staggered ends were exploited in **chapter 5** to create in a novel genome editing approach in *E. coli*, termed "cut and paste". Cas12a targets and cleaves at two selected locations within the genome. Cleavage by Cas12a generates double-stranded DNA breaks with 4-5 nt compatible staggered ends. These staggered ends can be repaired by ligation using T4 ligase. A plasmids reconstruction *in vivo* by cut & paste was attempted, but failed due to a low cleavage efficiency in one of the two spacers. Several spacers

sequences were screened for cleavage efficiency and the best was selected. In addition, a terminator like sequence in the T4 ligase gene was removed to increase expression of T4 ligase. With these improvements, a genomic deletion in *E. coli* by cut & paste was successfully achieved, albeit with a relatively low editing efficiency.

To further explore other type V systems, **chapter 6** focusses on the characterization of a novel compact type V systems, type V-U1 from *Mycolicibacterium mucogenicum* CCH10. The type V-U1 CRISPR-Cas locus express a small effector protein MmuCas12u1. MmuCas12u1 is roughly half the size of Cas12a. Despite its small size, MmuCas12u1 seems to retain some functional features also found in Cas12a. Features such as processing its own pre-crRNA, targeting dsDNA and recognizing a 5'-TTN-3' PAM. The RuvC domain of MmuCas12u1 does not cleave dsDNA, but instead is hypothesized to be involved dsDNA-activated transcriptional silencing. Apart from the RuvC, evidence is presented that also the zinc-finger domains contribute to this transcriptional silencing activity. By leveraging this property, MmuCas12u1 has been used for single- and multiplex- transcriptional silencing in *E. coli*.

Chapter 7 described how the fundamental knowledge gained on MmuCas12u1 is used to develop small Mmu base editors (MmuBE). MmuBEs are fusion proteins consisting of MmuCas12u1, cytidine deaminase and uracil glycosylase inhibitor, which is a tool for RNA-guided targeted nucleotide (C → T) substitution. By varying the linker lengths, the deaminase protein sequence, and the codon usage of MmuBEs, several MmuBE variants were constructed and characterized in *E. coli*. Most variants are relatively efficient, with a base editing window consisting of two regions, a PAM-proximal (2-5) and a PAM-distal (13-19) region, with the PAM-proximal region having more edits. It was found that less favorable codon usage reduces over editing efficiency, which can eliminate base editing in the PAM-distal region. In addition, a small-scale pilot experiment also demonstrated on-target base-editing by MmuBE in eukaryotic cells, namely in *Saccharomyces cerevisiae*. MmuBEs are currently the smallest base editors (genes ~2.8 kb) known, further expanding the current toolbox for prokaryotic base editing, and with great promise for eukaryotic base editing.

Discussion

The discovery of CRISPR-Cas started when scientists got curious about the physiological role of unique repetitive sequences that were found in the genomes of many bacteria. It is the drive, motivation, and hard work from these pioneers all around the world that led to these great discoveries we have today: from groundbreaking fundamental research of CRISPR-Cas being an adaptive immune system found in prokaryotes and archaea towards astonishing and life changing CRISPR-Cas-based technologies. In this thesis, research is described that adds some more stones to the previously established road. This final chapter discusses the different aspects of type V systems that are not covered in the previous chapters of this thesis, as well as an outlook on the future of type V CRISPR-Cas systems

Classification of type V CRISPR-Cas systems

The current type V system holds eleven characterized subtypes, V-A to V-K, and four uncharacterized subtypes, VU1 to VU4, making type V the most diverse type of all CRISPR-Cas systems (Fig. 1). A new subtype is established when the amino acid sequence of the effector protein is significantly different to that of an already characterized subtypes (18). Such analysis includes sequence similarity-based clustering and bipartite networks of gene sharing using modular structure (18). Also, the composition of the set of cas genes, and potential accessory genes, present on the CRISPR-locus are considered (18). A new subtype is classified with a new alphabetical letter, a letter that follows the most recent subtype. Meaning each type may only contain 26 subtypes (the number of letters in the alphabet). In four years, the type V systems has expanded to contain 11 subtypes. In the future, many more subtypes may be characterized as sequencing becomes more accessible and more metagenome becomes available. If type V goes beyond 26 subtypes, the classification will need to be adjusted in how subtypes are named. A possibility is to continue the alphabet, is by using the Greek alphabet, which adds an additional 24 subtypes. Another is to change the alphabetical system to a numbering system. Applying the Greek alphabet can be problematic as Cas12j is currently also named CasΦ. However, this can be solved by changing the name CasΦ to Cas12j in future publication, like how Cpf1, C2c1 and C2c2 were renamed to Cas12a, Cas12b and Cas13a, respectively (43, 58, 285). Nomenclature in CRISPR-Cas is crucial with this current rate of novel subtype characterization. For example, Cas14 is according the classification Cas12f (18). Cas14 should only be given the effector protein of a new type of CRISPR-Cas system, which is type VII. Even though Cas14 was later remained in other publication, still the name Cas14 remains in the scientific community. In addition, Cas12j (CasΦ) from type V-J was claimed when discovered in a metagenome data but was still uncharacterized and should have been named type V-U6 (285). Cas12k was characterized before Cas12j but took the following letter k instead (69). In short, naming and classifying novel type V systems should

be communicated, discussed, and agreed upon in the scientific community prior to publication. This is to adhere to consistency of the classification and to avoid confusion within the field.

Type V CRISPR-Cas systems

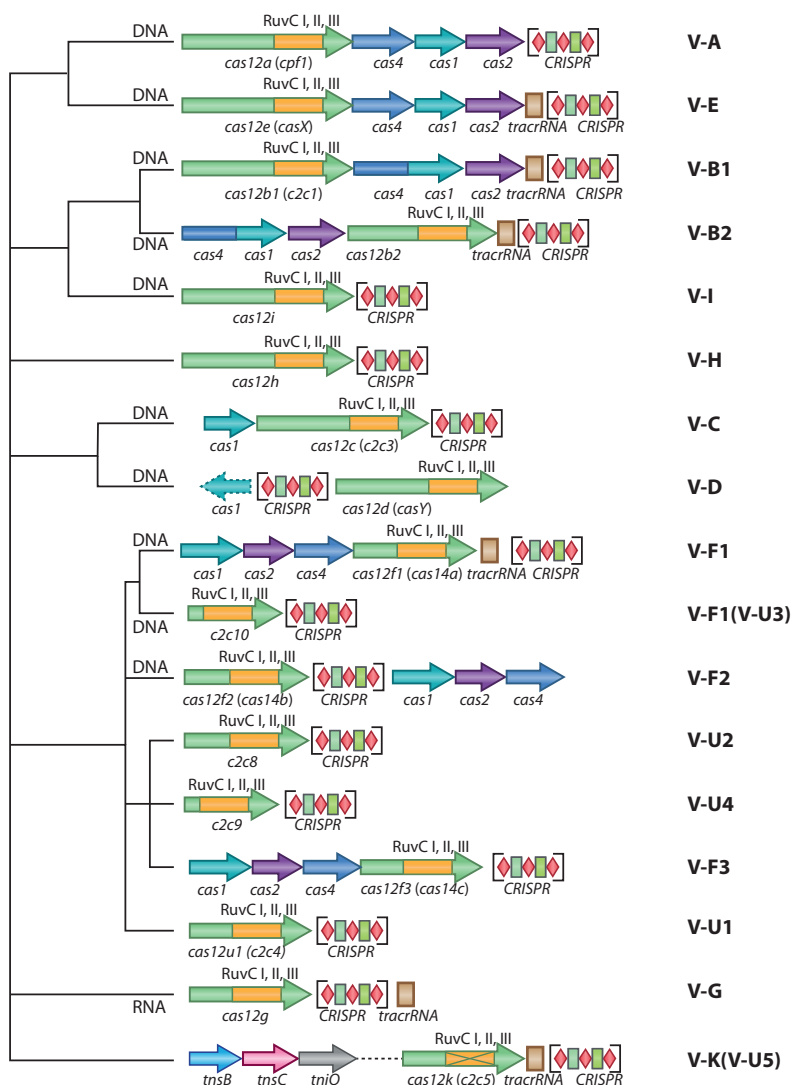


Figure 1 | Schematic classifications of type V CRISPR-Cas systems. Dendrogram shows likely evolutionary relationships between the different type V subtypes. Asterisk (*) indicates estimated placement of the type V-J system based on (286). Dashed lines in *cas1* of type V-D indicates presence of *cas1* in a subset of type V-D systems. The RuvC nuclease domain is indicated in orange; for clarity, this domain is depicted as one continuous motif although it is interrupted by other sequences (see Fig. 1, chapter 6). A cross in the RuvC indicates an inactive RuvC domain. Figure was adapted from (18).

Adaptation in type V

The adaptation module found in type V systems often consist of Cas4, Cas1 and Cas2. The role of Cas4 was elucidated in **chapter 3** for type V-A and V-B to select for PAM containing pre-spacer (PAM-scanning). Though found in the same type of CRISPR-Cas system, Cas4 functionality can differ between type V subtypes. For example, Cas4 found in type I systems have been reported not to have the exact same molecular mechanism for different subtypes (26-28). Therefore, Cas4 has been investigated for the different type V subtypes. Also, it is interesting to know if there is a biological advantage in having Cas4 and Cas1 either as separate proteins or as a fusion of the two. More insight could be gained by solving the structure of Cas4-Cas1-Cas2 complexes of different type V subtypes, as has been done for type I-C (28, 287).

Some type V systems, such as subtypes V-C and V-D, contain only Cas1, hence missing both Cas4 and Cas2. In type V-C, Cas1 alone has been demonstrated to be capable of acquiring functional PAM containing spacers (174). For V-D systems, Cas1 is present in some variants, so it would also be interesting to explore differences in CRISPR-Cas immunity in V-D variants with and without Cas1 (18). Furthermore, it is unclear whether Cas12 itself plays a role in adaptation like that of Cas9 in type II-A and remains to be explored (29, 30, 37).

Surprisingly, many type V subtypes do not contain an adaptation module (Fig. 1), and are thought to be ancestral Cas12 proteins that existed prior to the introduction of an adaptation module (288). How these type V systems acquire new spacers is still not yet known. It can be that the inherited CRISPR array is already fully equipped against, and that their mobile genetic the adaptation is no longer required resulting in loss of these genes. If these “adaptation-less” type V systems, co-occur with other CRISPR-Cas system, they still may be able to utilize the adaptation module or the CRISPR array from the other system, as has been demonstrated to occur in type III (289).

crRNA processing in type V

Some Cas12 nucleases do not require a tracrRNA and are able to process its own pre-crRNA, e.g. Cas12a from type V-A. Chapter 4 describes the importance of the sequence and the structure of the stem loop (also referred as the pseudoknot) for Cas12a pre-crRNA processing. It was found that spacer sequences within the crRNA can destabilize the pseudoknot by favoring other RNA structures. Destabilization of the crRNA structure can lead to lower cleavage efficiency by Cas12a. In other words, spacer sequences may affect the cleavage efficiency. Stabilization can be improved by changing 3' end spacer sequence, e.g. position 19-24, which does not base pair with the protospacer. The 3' end sequence should be modified to fold back and base pair with the spacer sequence (Fig. 2) (210). The drawback of this solution is the need

to assess and design each 3' flank sequence for each spacer used. However, that might still be better than the current solution, which is designing and testing three or more spacers. Apart from it is still hard to predict whether a given spacer sequence will have a high or a low cleavage efficiency. RNA prediction tools are still not accurate to accurately predict the RNA structure of your crRNA (290). The goal is to create such a crRNA for Cas12a that guarantees high cleavage efficiency, no matter the spacers sequence. For a universal solution, repeat sequences can be modified to allow for more stable pseudoknot. Modification of the stem loop or swapping a U-A pair in the stem to a G-C pair was found to increase Cas12a editing efficiency (291, 292). Another solution would be increasing the stem length of the pseudoknot. Some initial attempts to modify the pseudoknot were unsuccessful (Creutzburg & Van der Oost, unpublished), strongly suggesting that the corresponding RNA-binding pocket of Cas12a protein should be adjusted as well. Apart from crRNA structure, other factors may also play a role in cleavage efficiency, such as target accessibility, Cas nuclease and crRNA delivery and Cas nuclease precision (210).

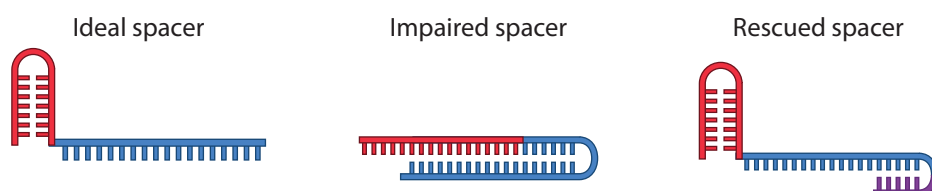


Figure 2 | Rescuing type V-A impaired crRNA. The crRNA consists of a repeat sequence (red), a spacer sequence (blue). A spacer can be “rescued” by modifying position 21-24 of the protospacer (purple), of which base pairs with the protospacer sequence to restore the pseudoknot structure.

Currently there is not conserved domain found in type V nucleases that indicates processing of pre-crRNA. In Cas12a, the residues responsible for pre-crRNA processing resides in the wedge domain (WED, an oligonucleotide binding domain (OBD)) (53). In Cas12j, substitution of one or more catalytic residues of the RuvC domain has been reported to abolish crRNA processing (285). For MmuCas12u1, residues involved in pre-crRNA processing have also been investigated through mutations in the OBD-like domain (R241A; R249A; H269A; R270A; R287A; H269A + R270A). Cas12u1 OBD mutants were tested in an *in vivo* silencing assay with either one spacer CRISPR array (repeat-spacer-repeat) or a four spacer CRISPR array. Surprisingly, no indication of reduced pre-crRNA processing was found in any of the mutants (data not shown). This implies that, based on sequence similarity and/or domain architecture, it is not straightforward to predict the key residues responsible for pre-crRNA processing of a Cas12 protein. Attempts to crystallize MmuCas12u1 are ongoing.

Not all Cas12 proteins can auto-process pre-crRNA, e.g. Cas12b from type V-B. Like Cas9, these Cas12 proteins require a tracrRNA as well as RNaseIII for crRNA maturation. The presence of a tracrRNA on a CRISPR-locus can be located using

a prediction tool (293). In type II-A of *Francisella Novicida*, a small CRISPR-Cas associated RNA (scaRNA) can also base pair with the tracrRNA. Cas9 containing a tracrRNA:scaRNA duplex can target and regulate transcription to aid in the virulence of *F. novicida* (41, 42). Recently, a long form tracrRNA (tr_L) has been found to guide Cas9 to downregulated its own CRISPR locus (42). Also, a new type of tracrRNA was found in subtypes V-C and V-D, known as short-complementarity untranslated RNA (scoutRNA) (70). The scoutRNA contains a very short sequence complementary to the repeat (anti-repeat) of the pre-crRNA and the secondary structure is predicted to be different to a tracrRNA (70). Other external RNAs, similar to scoutRNA might also be required the uncharacterized type V-U systems, where no predicted tracrRNA could be detected. Transcriptome analysis of type V-U containing organisms can aid in the detection of these elusive, “scout-like” RNAs. Another approach would be to express new CRISPR-Cas loci in *E. coli*, and then systematically deleting non-coding regions, as has been done to demonstrate that Cas12a does not require a tracrRNA (43).

Cas12 nucleases

Apart from different crRNA maturation strategies, Cas12 proteins also have features that are well conserved throughout type V; in many cases, these features are unique in that they are not shared with any other type of CRISPR-Cas system. All Cas12 proteins (except Cas12g) recognize a 5'-T-rich PAM, and all possess a single RuvC-like nuclease domain that is involved in target interference. In contrast, Cas9 possesses two nuclease domains: a RuvC domain that is responsible for cleavage of the non-target strand, and a HNH domain that cleaves the target strand of the dsDNA. In Cas12a, the RuvC cleaves both the non-target strand and the targeted strand (294, 295). However, the RuvC activity has been demonstrated to vary substantially between different Cas12 effectors (53). The majority of Cas12 proteins use the RuvC domain to target and cleave both strands of dsDNA, but variants have been described that do not follow this trend. For example, Cas12i predominantly nicks the non-target strand of dsDNA, Cas12f1 cleaves ssDNA in a PAM independent manner, and Cas12g cleaves ssRNA in a PAM independent manner (57, 67). The target of Cas12f1 is not entirely clear, as one study finds Cas12f1 to cleave ssDNA and another study reports cleavage of dsDNA (71). Although it may be that the rather large V-F clade (Fig. 1) includes variants with different target preferences, it is also possible that Cas12f1 cleaves both ssDNA and dsDNA but differs in activity based on cleavage conditions; future analyses are required to validate these findings. Also, a crystal structure of a Cas12f1 trapped while cleaving ssDNA and another one trapped cleaving dsDNA would help to explain the differences in target specificity. In addition, MmuCas12u1 was demonstrated to bind but not cleave dsDNA (**chapter 6**). Perhaps MmuCas12u1 is missing a component to cleave dsDNA, such as an external RNA such as the scoutRNA or a different buffer composition. Still, the activity of the RuvC domain of MmuCas12u1 remains elusive, although it is tempting to speculate that the observed expression silencing relies on binding and/or cleavage of the protein's

RuvC domain. Next, Cas12k also binds but does not cleave dsDNA, because Cas12k has an inactive RuvC domain (64, 69). Instead, Cas12k coupled with its accessory proteins (*tnsB*, *tnsC*, *tniQ*) target specific sites in genomic bacterial DNA as part of an RNA-guided transposition process (69). Overall, it is hard to predict the activity of the RuvC domain for these type V nucleases. The RuvC activity can be determined by its active site composition or determined by its surrounding domains. By solving the crystal structure of various Cas12 proteins and by studying the catalytic sites, insight should be gained that may explain the different RuvC activities found in the rapidly growing set of Cas12 nucleases.

Another interesting protein to characterize would be TnpB of the IS605 transposon family. Based on bioinformatic analysis, in particular of RuvC-like sequences, TnpB has been proposed to be the ancestral protein of Cas12 nucleases (64, 296). When co-expressed, the *tnpB* gene and the *tnpA* gene allow for autonomous transposition. However, when expressed alone, *tnpB* cannot support transposition (296, 297). Little is known about the activity of TnpB, since it is not required for transposition (298). It has been proposed that at several independent evolutionary events TnpB variants were associated with a CRISPR array, after which the acquisition of additional domain insertions led to the emergence of different RNA-guided Cas12 nucleases (64). Characterization of TnpB should focus on identifying the poly-nucleotide target of RuvC, the activity and molecular mechanism of RuvC and affinity towards a possible RNA or DNA “guide” (64). Apart from TnpB, characterization of other type V-U nucleases will also reveal pieces of the evolutionary path(s) from the proposed TnpB ancestor. The pool of identified TnpB and Cas12u nucleases depends on the currently available (meta)genomes in the database. This database will be further expanded, most likely resulting in new type V subtypes to study.

Current application of type V nucleases

The repurposing of CRISPR-Cas nucleases as a genome editing tool has been, and still is, revolutionary to the field of life science. It allowed for simple and quick genome editing in a wide range of prokaryotes and eukaryotes (299–302). Class 1 nuclease have been applied for genome editing by fusing the Cascade complex to a FokI domain (303). However, class 2 nucleases are still more widely used because of their single multi-domain effector proteins (303). Cas9 was the first class 2 nuclease to be characterized and applied in genome editing (76, 191). A few years later, Cas12a was characterized as a CRISPR-associated nuclease with distinct features, and with potential for genome editing of mammalian cells (43). One of the biggest advantages of Cas12a, is its very high target specificity (less off-target issues) as well as its ability to process its own pre-crRNA, which is beneficial for multiplexing (**chapter 4**). In addition, Cas12 recognizes a 5'-T-rich PAM, which (together with SpCas9 and its 3'-G-rich PAM) increases the targeting scope for genome editing (Fig. 4A) (100). Another unique feature of Cas12a is the generation of staggered ends after cleavage of dsDNA. These staggered ends can potentially be used for precision repair by

microhomology-dependent targeted integration, demonstrated in eukaryotes (206). In **Chapter 5** a similar methodology is described for editing in prokaryotes, termed cut & paste, resulting in proof of concept by generating a successful deletion in the *E. coli* genome. Although compared to Cas9 it lags behind, Cas12a publications have been steadily increasing ever since its discovery, meaning the Cas12a toolbox is also expanding (Fig. 3), especially in cases where Cas9 did not generate to desired results, such as in microorganisms and plants, or in certain mammalian cell types (304).

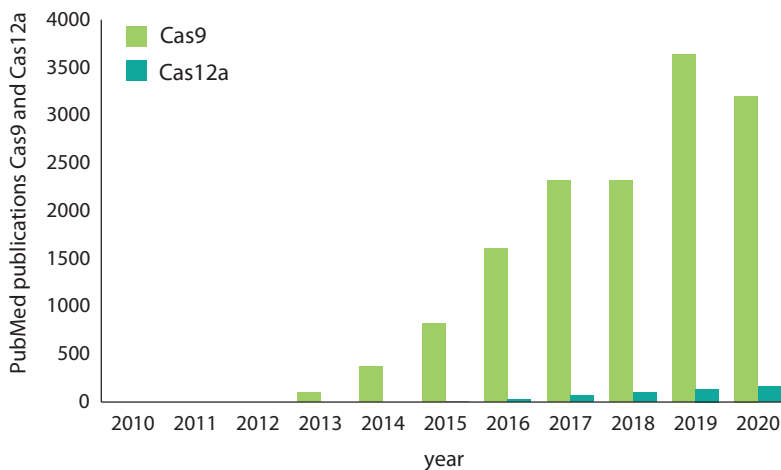


Figure 3 | NCBI PubMed publications containing “Cas9 AND CRISPR” and “Cas12a OR Cpf1 AND CRISPR”. Data was collected from PubMed on October 2020.

Apart from genome editing, Cas12a has also been utilized for (multiplex) transcriptional regulation such as silencing and activation (Fig. 4B and C)(305-307). Like Cas9, Cas12a based C to T base editors have also been developed using dCas12a, which does not cause DNA damage, unlike Cas9 base editors that nicks the target strand (Fig. 4D) (271, 308). A more recently identified feature of Cas12a is that target binding (dsDNA or ssDNA) activates indiscriminate ssDNA degradation (228, 309). This means that upon binding of its DNA target, it cleaves the targeted DNA in *cis* but also “collateral” ssDNA in *trans* (228, 309). This mechanism was also found in other Cas12 proteins, such as Cas12b, c, f, g, h, and i (chapter1, table 1). Using this feature of Cas12a, a nucleic acid detection tool was developed (DETECTR) (228, 310). Binding of Cas12a to a specific dsDNA sequence, cascades into cleavage of ssDNA-fluorescently quenched (FQ) reporters, which results in a fluorescent signal when cleaved (Fig. 4E) (228, 310). This type of CRISPR-Cas based detection technology was first developed for Cas13 from type VI (SHERLOCK) (311, 312). Cas13 targets and cleaves RNA in *cis*, but also cleaved “collateral” ssRNA in *trans* (44). SHERLOCK version-2 (SHERLOCKv2) combines Cas13, Cas12a and Csm6 from types VI, V and III, respectively to achieve multiplex nucleotide detection (313).

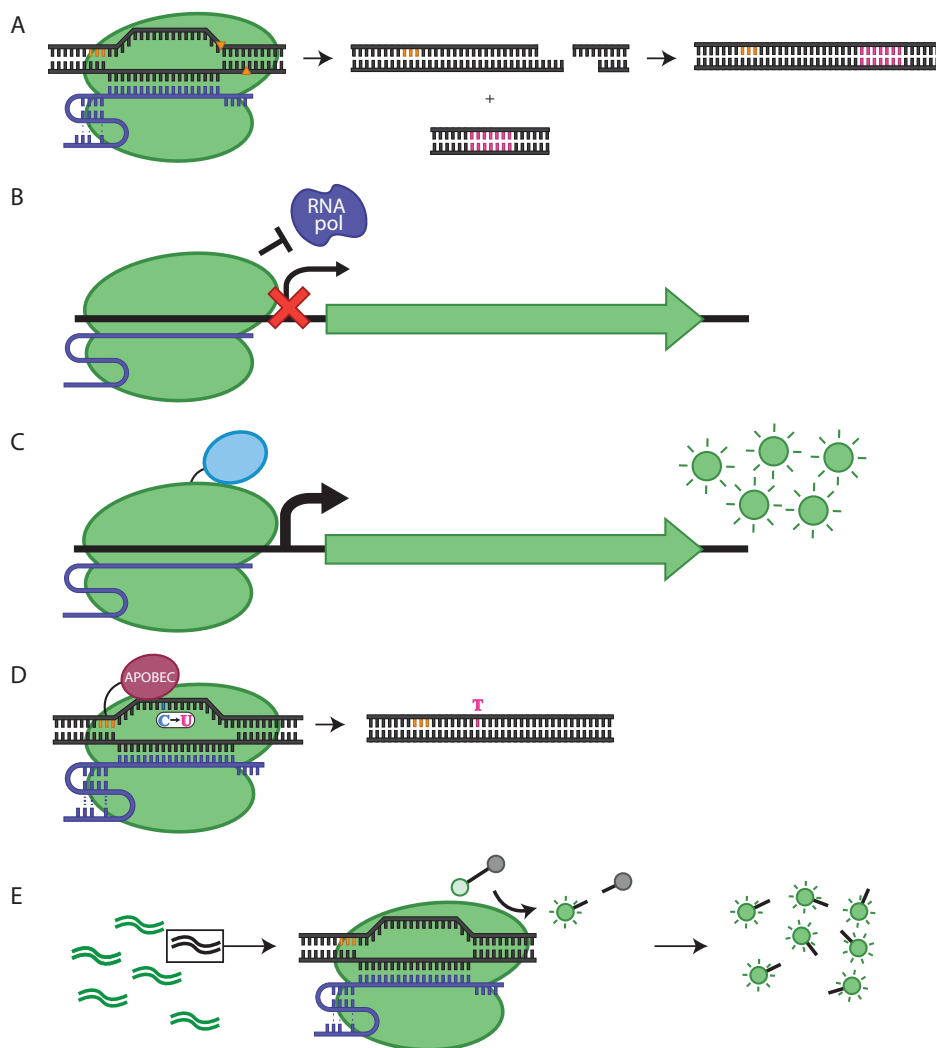


Figure 4 | Applications of using Cas12a. The PAM nucleotide motif is indicated in orange. Mutations are indicated in pink nucleotides. **(A)** Cleavage of dsDNA by Cas12a causes a dsDNA break. If supplemented with a repair template, homologous recombination takes place and uses the repair template to repair the double stranded break. Mutations are incorporated in the genome during repair **(B)** Transcriptional inhibition. Cas12a binds on the promoter region of a gene of interest and inhibits RNA polymerase (RNA pol) from binding to its recognition site to start transcription. **(C)** Transcriptional activation. Cas12a fused to a transcriptional activator (blue), of which can upregulation transcription and thereby expression. **(D)** Base editing. dsDNA binding and R-loop formation allows APOBEC to deaminate cytosine (C) to uracil (U) on the non-target strand. After replication uracil is turned into thymine (T), and on the complementary strand a guanine (G) into an adenine (A). **(E)** Nucleic acid detection. Binding or cleavage of the target, activates indiscriminate ssDNA collateral cleavage in *trans*. Cleavage of ssDNA containing a quencher (grey) and fluorophore (green), will release the quencher and allows a fluorescence signal to be detected.

Next to Cas12a applications, similar applications have also been developed using Cas12b such as genome editing and development of a nucleotide detection tool (314, 315). One of these smaller Cas12 nucleases is MmuCas12u1, which was characterized and further developed into a small C to T base editor (chapter 7). In addition to the C to T base editor, MmuCas12u1 was also used to create an A to G base editor by fusing MmuCas12u1 to an adenosine deaminase (316, 317). Two MmuCas12u1 adenosine base editors (MmuABEs) were tested by targeting three different A motif plasmids in *E. coli*. The different A motif plasmids contain a tiled A motif (AxxAxxAxxAxxAxxAxxA), starting at every first (A1 motif), second (A2 motif) or third (A3 motif) nucleotide of the protospacer (Fig. 5A). To normalize for base editing efficiency, all three A motif plasmids contain an A on position 4. Recent results show both A to G and T to C base editing by MmuABEs, meaning MmuABEs base edit on the non-target and the target strand, respectively. Like MmuBEs, base editing occurred in two editing regions, a PAM proximal and a PAM distal region (Fig. 5B) (chapter 7). A to G mutation was detected in the PAM proximal region (position 3, 4, 6 and 8) and T to C mutation was detected in the PAM distal region (position 16, 18 and 20).

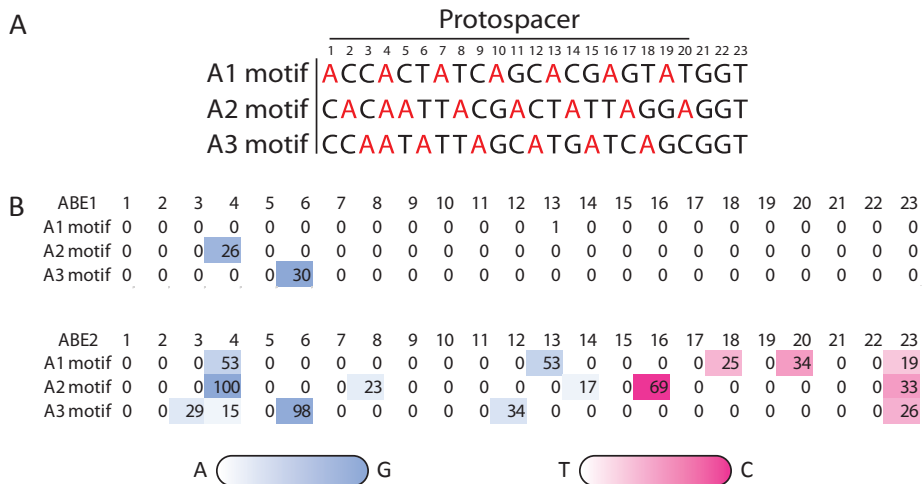


Figure 5 | A to G base editing by MmuABEs. (A) Base editing targets consisting of an A on every first, second and third position of each trinucleotide. These plasmids were named C1, C2 and C3 motif, respectively. A on position four was present in all A motif targets **(B)** Heat map representing % of base edited C's using different variants of MmuABEs. White to blue gradient indicate A to G mutation and white to pink gradient indicates T to C mutation.

Future application and prospective of type V nucleases

In the future, more Cas12 nuclease will get characterized, meaning more applications with Cas12 nucleases will be developed. The more recently discovered Cas12 nucleases are also smaller than the currently used nucleases, e.g. Cas12a or Cas12b. Smaller nucleases (e.g. Cas12j) can be advantageous for eukaryotic genome editing as they are better suited for delivery using Adeno-associated virus (AAV) vectors (274, 285). More effort will be put into the mining, characterizing, and repurposing of compact type V nucleases. Systems such as type V-U2 and U4 still remain to be characterized and exploited (64). Depending on their biochemical features, Cas12 nucleases will be engineered to cleave more efficiently, precisely and recognize other PAM sequences (270, 318). To better control activity of Cas12a genome editing, similar to Cas9, Cas12a can also be engineered to be regulated by small molecules or light (**chapter 2**). Another way to increase cleavage specificity is using engineered Cas12a nickases, so that two targets are required to generate a double stranded (84, 319, 320). More applicable would be to use Cas12i, which naturally predominantly nicks dsDNA (57). Cas12a nickases can also be beneficial in base editing as nickases were found to increase base editing efficiency, because nicking the non-base edited strand stimulates the cell to repair the non-edited strand using the edited strand as template (159, 160). However, this increased base editing by nicking comes at a cost of increased DNA damage in the cell (308). Currently only C to T base editors exist for Cas12a and is anticipated to also include A to G and C to G base editors. Cas12 base editors can also be expanded by using other Cas12 nucleases, as different Cas12 base editors can have different base editing windows (266). The goal is to have a complete arsenal of Cas12 base editors, that can be used for different base editing application. Current research into natural or engineered deaminases will likely give rise to new nucleotide conversion that are currently not available.

A to-be-developed nickase variant of Cas12a can also be used for developing a Cas12a based prime editor (321). Prime editor allows for a novel genome editing technique and in which dsDNA is edited precisely, without generating a dsDNA break and without requiring a separate donor template (Fig. 6). The recently developed prime editor consists of Cas9 fused to a reverse transcriptase. Apart from the fusion protein, the sgRNA has also been modified into an extended, prime editing guide RNA (pegRNA). A pegRNA consists of an sgRNA, a reverse transcriptase template containing the edit and a primer binding sequence. Nicking of the non-target strand by Cas9 allows binding of the non-target strand to the primer binding site of the pegRNA. Reverse transcriptase will then continue to elongate and generate complementary DNA (cDNA) of the RNA template. Once finished, the newly generated cDNA can base pair with the target strand, which results in an equilibrium between the 5' edited flap and the 3' unedited flap. Cleavage and ligation of the 5' edited flap followed by DNA repair results in successful edited DNA (Fig. 6). Using a similar design, a prime editor can potentially also be constructed using nickase Cas12a or Cas12i.

Another type V protein with unique functionalities is Cas12k, that can be utilized to incorporate large gene clusters in the genome of production strains, e.g. for incorporation of novel metabolic pathways for the dedicated production of desired compounds such as antibiotics. Other Cas12 nucleases that can cleave collateral oligonucleotides in *trans*, such as Cas12c, f, g, h and i, can be utilized as a nucleotide detection tool similar to Cas12a and Cas13.

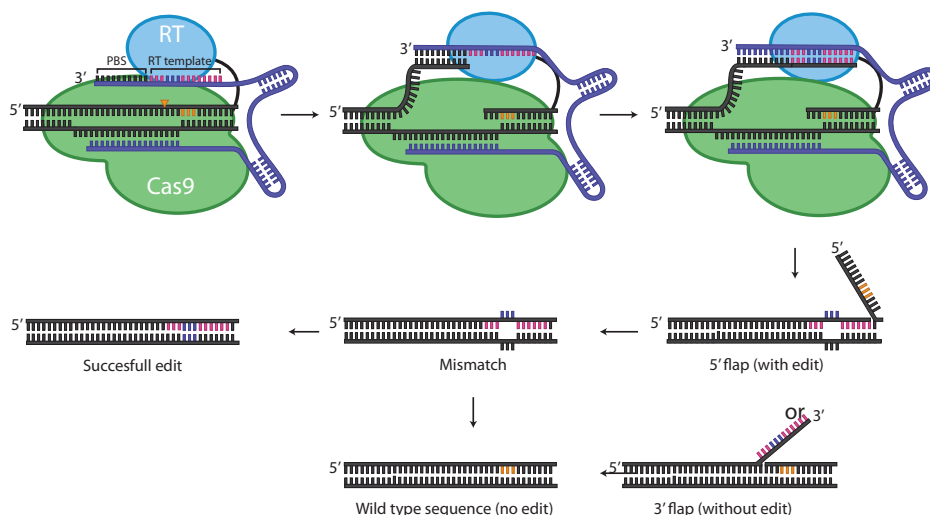


Figure 6 | Schematic of genome editing by a Cas9 prime editor. dsDNA containing a 5'-NGG PAM (orange) is targeted by Cas9 fused to a reverse transcriptase (RT). Cas9-RT is guided by a pegRNA (purple) containing a spacer, a primer binding site (PBS) and a reverse transcriptase template (RT template) (pink) containing the mutation (purple). The non-target strand gets nicked by Cas9 and base pairs with the PBS of the pegRNA. RT then transcribes the cDNA using the RT template. After reverse transcription, the cDNA can base pair with the DNA creating a 5' flap. The cDNA can also be excluded, creating a 3' flap. Repair of the 3' flap results in unedited wild type DNA. Repair of the 5' flap results in a mismatch between the cDNA and the target strand DNA. Resolving the mismatch can result in an unedited wild type sequence or a successful edited sequence (321).

Currently, many Cas12 proteins have been characterized but only few Cas12 structures have been solved. Those include: Cas12a, Cas12b, Cas12e and cas12i (66, 83, 84, 322). Once all Cas12 protein structures become available, structures can be studied and compared, and used as basis to investigate the structural mechanism behind the molecular features of the different Cas12 nucleases. Interesting cases would be Cas12g's RNA targeting and cleavage, Cas12b's requirement for a tracrRNA and its lack of ability to process its own pre-crRNA, Cas12f's ambiguous targeting of ssDNA and dsDNA, and Cas12i's increased nicking activity compared to Cas12h. Answering these fundamental questions will lead to increased understanding of these diverse Cas12 nucleases and will ultimately allow for improved engineering efficiencies. Possibly, new Cas12 nucleases can be created by combining domains derived from different Cas12 nucleases. Like creating a Cas12g nuclease able to

target and both ssRNA and ssDNA, or Cas12f1 and Cas12e variants that does not require a tracrRNA.

It is an exciting time to be active in the field of CRISPR-Cas, especially that of the growing type V CRISPR-Cas systems. The type V systems is expected to further increase with new subtypes in the near future, which most likely will bring along new features with potential for improved application and tools. Natural and synthetic Cas12 variants will further expand the CRISPR-Cas toolbox in both genome editing and diagnostics. The field of CRISPR-Cas has come a long way and with no end in sight.

It is good to remember that it all started by being curious and use that as a main driving force to conduct scientific research.

References

1. A. Fyfe, C. M. Røstvik. (Nature Publishing Group, 2018).
2. S. Shapin, *A social history of truth: Civility and science in seventeenth-century England*. (University of Chicago Press, 2011).
3. E. Rivers, Women, minorities, and persons with disabilities in science and engineering. *National Science Foundation*, (2017).
4. J. Agar, 2016 Wilkins–Bernal–Medawar lecture The curious history of curiosity-driven research. *Notes and Records: The Royal Society Journal of the History of Science* **71**, 409–429 (2017).
5. E. S. Lander, The heroes of CRISPR. *Cell* **164**, 18–28 (2016).
6. Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of bacteriology* **169**, 5429–5433 (1987).
7. F. J. Mojica, G. Juez, F. Rodriguez–Valera, Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Molecular microbiology* **9**, 613–621 (1993).
8. R. Jansen, J. D. v. Embden, W. Gaastra, L. M. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology* **43**, 1565–1575 (2002).
9. A. Bolotin, B. Quinquis, A. Sorokin, S. D. Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
10. F. J. Mojica, J. García-Martínez, E. Soria, Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution* **60**, 174–182 (2005).
11. C. Pourcel, G. Salvignol, G. Vergnaud, CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
12. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct* **1**, 7 (2006).
13. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
14. S. J. Brouns *et al.*, Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
15. P. Mohanraju *et al.*, Diverse evolutionary roots and mechanistic variations of the CRISPR–Cas systems. *Science* **353**, aad5147 (2016).
16. J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
17. K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR–Cas systems. **13**, 722–736 (2015).
18. K. S. Makarova *et al.*, Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*, 1–17 (2019).
19. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

20. J. Van Der Oost, E. R. Westra, R. N. Jackson, B. Wiedenheft, Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Reviews Microbiology* **12**, 479–492 (2014).
21. J. Van der Oost, M. M. Jore, E. R. Westra, M. Lundgren, S. J. Brouns, CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences* **34**, 401–407 (2009).
22. I. Yosef, M. G. Goren, U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic acids research* **40**, 5569–5576 (2012).
23. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
24. J. K. Nuñez *et al.*, Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature structural & molecular biology* **21**, 528 (2014).
25. B. Wiedenheft *et al.*, Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912 (2009).
26. M. Shiimori, S. C. Garrett, B. R. Graveley, M. P. Terns, Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci. *Molecular cell* **70**, 814–824. e816 (2018).
27. S. N. Kieper *et al.*, Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell reports* **22**, 3377–3384 (2018).
28. H. Lee, Y. Zhou, D. W. Taylor, D. G. Sashital, Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays. *Molecular cell* **70**, 48–59. e45 (2018).
29. Y. Wei, R. M. Terns, M. P. Terns, Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes & development* **29**, 356–361 (2015).
30. R. Heler *et al.*, Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519**, 199–202 (2015).
31. Y. Koo, D.-k. Jung, E. Bae, Crystal structure of *Streptococcus pyogenes* Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS One* **7**, e33401 (2012).
32. M. Wilkinson *et al.*, Structure of the DNA-bound spacer capture complex of a type II CRISPR–Cas system. *Molecular cell* **75**, 90–101. e105 (2019).
33. S. A. Jackson *et al.*, CRISPR–Cas: Adapting to change. *Science* **356**, (2017).
34. P. C. Fineran *et al.*, Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences* **111**, E1629–E1638 (2014).
35. D. C. Swarts, C. Mosterd, M. W. Van Passel, S. J. Brouns, CRISPR interference directs strand specific spacer acquisition. *PLoS one* **7**, e35888 (2012).
36. T. Künne *et al.*, Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. *Molecular cell* **63**, 852–864 (2016).
37. P. M. Nussenzweig, J. McGinn, L. A. Marraffini, Cas9 cleavage of viral genomes primes the acquisition of new immunological memories. *Cell host & microbe* **26**, 515–526. e516 (2019).
38. J. Carte, N. T. Pfister, M. M. Compton, R. M. Terns, M. P. Terns, Binding and cleavage of CRISPR RNA by Cas6. *Rna* **16**, 2181–2188 (2010).
39. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
40. M. Jinek *et al.*, A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816 (2012).
41. H. K. Ratner *et al.*, Catalytically active Cas9 mediates transcriptional interference to facilitate bacterial virulence. *Molecular cell* **75**, 498–510. e495 (2019).

-
42. T. R. Sampson, S. D. Saroj, A. C. Llewellyn, Y.-L. Tzeng, D. S. Weiss, A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254-257 (2013).
 43. B. Zetsche *et al.*, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771 (2015).
 44. O. O. Abudayyeh *et al.*, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, (2016).
 45. I. Fonfara, H. Richter, M. Bratovič, A. Le Rhun, E. Charpentier, The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517-521 (2016).
 46. B. Zetsche *et al.*, Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. **35**, 31-34 (2017).
 47. S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67 (2014).
 48. J. N. Vink *et al.*, Direct visualization of native CRISPR target search in live bacteria reveals Cascade DNA surveillance mechanism. *Molecular Cell* **77**, 39-50. e10 (2020).
 49. R. T. Leenay *et al.*, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Molecular cell* **62**, 137-147 (2016).
 50. O. O. Abudayyeh *et al.*, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
 51. J. R. Elmore *et al.*, Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes & development* **30**, 447-459 (2016).
 52. X. Wu *et al.*, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature biotechnology* **32**, 670-676 (2014).
 53. D. C. Swarts, J. van der Oost, M. Jinek, Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Molecular cell* **66**, 221-233. e224 (2017).
 54. L. Loeff, S. J. Brouns, C. Joo, Repetitive DNA reeling by the Cascade-Cas3 complex in nucleotide unwinding steps. *Molecular cell* **70**, 385-394. e383 (2018).
 55. E. R. Westra *et al.*, CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Molecular cell* **46**, 595-605 (2012).
 56. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences* **109**, E2579-E2586 (2012).
 57. W. X. Yan *et al.*, Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88-91 (2019).
 58. S. Shmakov *et al.*, Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Molecular cell* **60**, 385-397 (2015).
 59. D. Wu, X. Guan, Y. Zhu, K. Ren, Z. Huang, Structural basis of stringent PAM recognition by CRISPR-C2c1 in complex with sgRNA. *Cell research* **27**, 705-708 (2017).
 60. K. Johnson, B. A. Learn, M. A. Estrella, S. Bailey, Target sequence requirements of a type III-B CRISPR-Cas immune system. *Journal of Biological Chemistry* **294**, 10290-10299 (2019).
 61. O. Niewoehner *et al.*, Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548 (2017).
 62. M. Kazlauskienė, G. Kostiuk, Č. Venclovas, G. Tamulaitis, V. Siksnys, A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609 (2017).
 63. A. J. Meeske, S. Nakandakari-Higa, L. A. Marraffini, Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241-245 (2019).

64. S. Shmakov *et al.*, Diversity and evolution of class 2 CRISPR–Cas systems. *Nature reviews microbiology* **15**, 169-182 (2017).
65. D. C. Swarts, M. Jinek, Mechanistic Insights into the cis-and trans-Acting DNase Activities of Cas12a. *Molecular cell* **73**, 589-600. e584 (2019).
66. J.-J. Liu *et al.*, CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218-223 (2019).
67. L. B. Harrington *et al.*, Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839-842 (2018).
68. D. Burstein *et al.*, New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237-241 (2017).
69. J. Strecker *et al.*, RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48-53 (2019).
70. L. B. Harrington *et al.*, A scoutRNA Is Required for Some Type V CRISPR-Cas Systems. *Molecular Cell*, (2020).
71. T. Karvelis *et al.*, PAM recognition by miniature CRISPR–Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic acids research* **48**, 5016-5023 (2020).
72. J. A. Doudna, E. Charpentier, The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
73. M. L. Hochstrasser *et al.*, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proceedings of the National Academy of Sciences* **111**, 6618-6623 (2014).
74. A. A. Smargon *et al.*, Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Molecular cell* **65**, 618-630. e617 (2017).
75. J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).
76. P. Mali *et al.*, RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823 (2013).
77. D. Kim *et al.*, Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature biotechnology* **34**, 863-868 (2016).
78. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31**, 233 (2013).
79. B. Zetsche *et al.*, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771 (2015).
80. C. Anders, O. Niewoehner, A. Duerst, M. Jinek, Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569-573 (2014).
81. P. Gao, H. Yang, K. R. Rajashankar, Z. Huang, D. J. Patel, Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell research* **26**, 901-913 (2016).
82. D. C. Swarts, J. van der Oost, M. Jinek, Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell* **66**, 221-233. e224 (2017).
83. L. Liu *et al.*, C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. *Molecular Cell* **65**, 310-322 (2016).
84. T. Yamano *et al.*, Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949-962 (2016).
85. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827-832 (2013).

-
86. V. Pattanayak *et al.*, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. **31**, 839-843 (2013).
 87. X. Wu *et al.*, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol* **32**, 670-676 (2014).
 88. H. O'Geen, I. M. Henry, M. S. Bhakta, J. F. Meckler, D. J. Segal, A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Research* **43**, 3389-3404 (2015).
 89. J.-P. Zhang *et al.*, Different Effects of sgRNA Length on CRISPR-mediated Gene Knockout Efficiency. **6**, 28566 (2016).
 90. Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, J. K. Joung, Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology* **32**, 279-284 (2014).
 91. R. Chari, P. Mali, M. Moosburner, G. M. Church, Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. **12**, 823-826 (2015).
 92. J. G. Doench *et al.*, Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature biotechnology* **32**, 1262-1267 (2014).
 93. I. M. Slaymaker *et al.*, Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88 (2015).
 94. B. P. Kleinstiver *et al.*, High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495 (2016).
 95. J. S. Chen *et al.*, Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407-410 (2017).
 96. S. W. Cho *et al.*, Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research* **24**, 132-141 (2014).
 97. S. Q. Tsai *et al.*, Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology* **32**, 569-576 (2014).
 98. B. P. Kleinstiver *et al.*, Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature biotechnology* **34**, 869-874 (2016).
 99. W. Tang, J. H. Hu, D. R. Liu, Aptazyme-embedded guide RNAs enable ligand-responsive genome editing and transcriptional activation. *Nature Communications* **8**, 15939 (2017).
 100. B. Zetsche, S. E. Volz, F. Zhang, A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nature biotechnology* **33**, 139-142 (2015).
 101. K. M. Davis, V. Pattanayak, D. B. Thompson, J. A. Zuris, D. R. Liu, Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nature chemical biology* **11**, 316-318 (2015).
 102. J. Hemphill, E. K. Borchardt, K. Brown, A. Asokan, A. Deiters, Optical control of CRISPR/Cas9 gene editing. *Journal of the American Chemical Society* **137**, 5642-5645 (2015).
 103. G. Petris *et al.*, Hit and go CAS9 delivered through a lentiviral based self-limiting circuit. *Nature Communications* **8**, 15334 (2017).
 104. A. Pawluk *et al.*, Naturally occurring off-switches for CRISPR-Cas9. *Cell* **167**, 1829-1838. e1829 (2016).
 105. B. L. Oakes *et al.*, Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nature biotechnology* **34**, 646-651 (2016).
 106. K. I. Liu *et al.*, A chemical-inducible CRISPR-Cas9 system for rapid control of genome editing. *Nature chemical biology* **12**, 980-987 (2016).
 107. B. Maji *et al.*, Multidimensional chemical control of CRISPR-Cas9. *Nature chemical biology* **13**, 9 (2017).

108. Y. Nihongaki, F. Kawano, T. Nakajima, M. Sato, Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nature biotechnology* **33**, 755-760 (2015).
109. J. C. Rose *et al.*, Rapidly inducible Cas9 and DSB-ddPCR to probe editing kinetics. *Nature* **201**, 7 (2017).
110. K. Xie, B. Minkenberg, Y. Yang, Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences* **112**, 3570-3575 (2015).
111. M. A. Świat *et al.*, FnCpf1: a novel and efficient genome editing tool for *Saccharomyces cerevisiae*. *Nucleic Acids Research* **45**, 12585-12598.
112. A. Vasileva, R. Jessberger, Precise hit: adeno-associated virus in gene targeting. *Nature Reviews Microbiology* **3**, 837-847 (2005).
113. F. Mingozi, K. A. High, Therapeutic in vivo gene transfer for genetic disease using AAV: progress and challenges. *Nature reviews genetics* **12**, 341-355 (2011).
114. G. Gao, L. H. Vandenberghe, J. M. Wilson, New recombinant serotypes of AAV vectors. *Current gene therapy* **5**, 285-297 (2005).
115. M. A. Kay, State-of-the-art gene-based therapies: the road ahead. *Nature Reviews Genetics* **12**, 316-328 (2011).
116. F. A. Ran *et al.*, In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186-191 (2015).
117. M. Tabebordbar *et al.*, In vivo gene editing in dystrophic mouse muscle and muscle stem cells. *Science* **351**, 407-411 (2016).
118. C. E. Nelson *et al.*, In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science* **351**, 403-407 (2016).
119. A. V. Wright *et al.*, Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences* **112**, 2984-2989 (2015).
120. D.-J. J. Truong *et al.*, Development of an intein-mediated split-Cas9 system for gene therapy. *Nucleic acids research* **43**, 6450-6458 (2015).
121. E. J. Fine *et al.*, Trans-spliced Cas9 allows cleavage of HBB and CCR5 genes in human cells using compact expression cassettes. *Scientific reports* **5**, 10777 (2015).
122. W. L. Chew *et al.*, A multifunctional AAV-CRISPR-Cas9 and its host response. *Nature methods* **13**, 868-874 (2016).
123. J. Li, W. Sun, B. Wang, X. Xiao, X.-Q. Liu, Protein trans-splicing as a means for viral vector-mediated in vivo gene therapy. *Human gene therapy* **19**, 958-964 (2008).
124. C. Zincarelli, S. Soltys, G. Rengo, J. E. Rabinowitz, Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Molecular Therapy* **16**, 1073-1080 (2008).
125. D. S. D'Astolfo *et al.*, Efficient intracellular delivery of native proteins. *Cell* **161**, 674-690 (2015).
126. J. A. Zuris *et al.*, Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature biotechnology* **33**, 73-80 (2015).
127. S. Kim, D. Kim, S. W. Cho, J. Kim, J.-S. Kim, Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome research* **24**, 1012-1019 (2014).
128. S. Ramakrishna *et al.*, Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome research* **24**, 1020-1027 (2014).
129. J. Mueller, I. Kretzschmar, R. Volkmer, P. Boisguerin, Comparison of cellular uptake using 22 CPPs in 4 different cell lines. *Bioconjugate chemistry* **19**, 2363-2374 (2008).

-
130. M. Lundberg, M. Johansson, Is VP22 nuclear homing an artifact? *Nature biotechnology* **19**, 713-713 (2001).
 131. K. Schumann *et al.*, Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proceedings of the National Academy of Sciences* **112**, 10437-10442 (2015).
 132. J. Derdelinckx, Z. N. Berneman, N. Cools, GMP-grade mRNA electroporation of dendritic cells for clinical use. *Synthetic mRNA: Production, Introduction Into Cells, and Physiological Consequences*, 139-150 (2016).
 133. L. Xu *et al.*, CRISPR-mediated genome editing restores dystrophin expression and function in mdx mice. *Molecular Therapy* **24**, 564-569 (2016).
 134. O. Boussif *et al.*, A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proceedings of the National Academy of Sciences* **92**, 7297-7301 (1995).
 135. A. J. Mahiny *et al.*, In vivo genome editing using nuclease-encoding mRNA corrects SP-B deficiency. *Nature biotechnology* **33**, 584-586 (2015).
 136. P. L. Felgner *et al.*, Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. *Proceedings of the National Academy of Sciences* **84**, 7413-7417 (1987).
 137. M. Wang *et al.*, Efficient delivery of genome-editing proteins using bioreducible lipid nanoparticles. *Proceedings of the National Academy of Sciences* **113**, 2868-2873 (2016).
 138. W. Sun *et al.*, Self-Assembled DNA Nanoclews for the Efficient Delivery of CRISPR-Cas9 for Genome Editing. *Angewandte Chemie International Edition* **54**, 12029-12033 (2015).
 139. H. Yin *et al.*, Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. *Nature biotechnology* **34**, 328-333 (2016).
 140. R. Maurisse *et al.*, Comparative transfection of DNA into primary and transformed mammalian cells from different lineages. *BMC biotechnology* **10**, 9 (2010).
 141. S. Dokka, D. Toledo, X. Shi, V. Castranova, Y. Rojanasakul, Oxygen radical-mediated pulmonary toxicity induced by some cationic liposomes. *Pharmaceutical research* **17**, 521-525 (2000).
 142. S. Armeanu *et al.*, Optimization of nonviral gene transfer of vascular smooth muscle cells in vitro and in vivo. *Molecular Therapy* **1**, 366-375 (2000).
 143. B. T. Staahl *et al.*, Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. *Nature Biotechnology* **35**, 431-434 (2017).
 144. K. Lee *et al.*, Nanoparticle delivery of Cas9 ribonucleoprotein and donor DNA in vivo induces homology-directed DNA repair. *Nature Biomedical Engineering* **1**, 889-901 (2017).
 145. M. Jasin, J. E. Haber, The democratization of gene editing: Insights from site-specific cleavage and double-strand break repair. *DNA repair* **44**, 6-16 (2016).
 146. M. Bétermier, P. Bertrand, B. S. Lopez, Is non-homologous end-joining really an inherently error-prone process? *PLoS genetics* **10**, e1004086 (2014).
 147. S. Nakade *et al.*, Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nature communications* **5**, (2014).
 148. K. Suzuki *et al.*, In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144 (2016).
 149. Y. Kan, B. Ruis, T. Takasugi, E. A. Hendrickson, Mechanisms of precise genome editing using oligonucleotide donors. *Genome research* **27**, 1099-1111 (2017).
 150. J.-B. Renaud *et al.*, Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases. *Cell reports* **14**, 2263-2272 (2016).
 151. D. Paquet *et al.*, Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125-125 (2016).

152. C. D. Richardson, G. J. Ray, M. A. DeWitt, G. L. Curie, J. E. Corn, Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature biotechnology* **34**, 339-344 (2016).
153. L. Davis, N. Maizels, Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. *Proceedings of the National Academy of Sciences* **111**, E924-E932 (2014).
154. S. Lin, B. Staahl, R. K. Alla, J. A. Doudna, Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766 (2014).
155. T. Gutschner, M. Haemmerle, G. Genovese, G. F. Draetta, L. Chin, Post-translational regulation of Cas9 during G1 enhances homology-directed repair. *Cell reports* **14**, 1555-1566 (2016).
156. A. Orthwein *et al.*, A mechanism for the suppression of homologous recombination in G1 cells. *Nature* **528**, 422 (2015).
157. A. N. Zelensky, J. Schimmel, H. Kool, R. Kanaar, M. Tijsterman, Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nature Communications* **8**, (2017).
158. K. Kim *et al.*, Highly efficient RNA-guided base editing in mouse embryos. *Nature biotechnology* **35**, 435-437 (2017).
159. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).
160. K. Nishida *et al.*, Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
161. C. Kuscu *et al.*, CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nature methods* **14**, 710 (2017).
162. Y. B. Kim *et al.*, Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nature biotechnology* **35**, 371 (2017).
163. D. B. Cox *et al.*, RNA editing with CRISPR-Cas13. *Science* **358**, 1019-1027 (2017).
164. A. H. Badran *et al.*, Programmable base editing of A• T to G• C in genomic DNA without DNA cleavage. *Nature* **551**, 464 (2017).
165. R. H. Staals *et al.*, Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nature communications* **7**, 1-13 (2016).
166. T. J. Nicholson *et al.*, Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. *RNA biology* **16**, 566-576 (2019).
167. I. Fonfara, H. Richter, M. Bratovic, A. Le Rhun, E. Charpentier, The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517-521 (2016).
168. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *science* **337**, 816-821 (2012).
169. F. J. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740 (2009).
170. C. Rollie, S. Graham, C. Rouillon, M. F. White, Prespacer processing and specific integration in a Type IA CRISPR system. *Nucleic acids research* **46**, 1007-1020 (2018).
171. D. Ka, D. M. Jang, B. W. Han, E. Bae, Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. *Nucleic acids research* **46**, 9805-9815 (2018).
172. Y. He *et al.*, Cas1 and Cas2 from the type II-C CRISPR-Cas system of *Riemerella anatipestifer* are required for spacer acquisition. *Frontiers in cellular and infection microbiology* **8**, 195 (2018).

-
173. S. Hooton, I. F. Connerton, *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Frontiers in microbiology* **5**, 744 (2015).
 174. A. V. Wright *et al.*, A functional mini-integrase in a two-protein type VC CRISPR system. *Molecular cell* **73**, 727-737. e723 (2019).
 175. J. Zhang, T. Kasciukovic, M. F. White, The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PLoS One* **7**, e47232 (2012).
 176. S. Hudaiberdiev *et al.*, Phylogenomics of Cas4 family nucleases. *BMC evolutionary biology* **17**, 232 (2017).
 177. H. Yang, P. Gao, K. R. Rajashankar, D. J. Patel, PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease. *Cell* **167**, 1814-1828. e1812 (2016).
 178. R. E. McKenzie, C. Almendros, J. N. Vink, S. J. Brouns, Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nature protocols* **14**, 976-990 (2019).
 179. S. Lemak *et al.*, The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic acids research* **42**, 11144-11155 (2014).
 180. A. Levy *et al.*, CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510 (2015).
 181. A. M. Kabadi, D. G. Ousterout, I. B. Hilton, C. A. Gersbach, Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res* **42**, e147 (2014).
 182. L. Nissim, S. D. Perli, A. Fridkin, P. Perez-Pinera, T. K. Lu, Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol Cell* **54**, 698-710 (2014).
 183. T. Sakuma, A. Nishikawa, S. Kume, K. Chayama, T. Yamamoto, Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Scientific reports* **4**, 5400 (2014).
 184. S. Q. Tsai *et al.*, Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32**, 569-576 (2014).
 185. K. Xie, B. Minkenberg, Y. Yang, Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3570-3575 (2015).
 186. T. Yamano *et al.*, Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949-962 (2016).
 187. C. Ostlund *et al.*, Dynamics and molecular interactions of linker of nucleoskeleton and cytoskeleton (LINC) complex proteins. *Journal of cell science* **122**, 4099-4108 (2009).
 188. L. Swiech *et al.*, In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nature Biotechnology* **33**, 102-U286 (2015).
 189. S. Konermann *et al.*, Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472-476 (2013).
 190. N. Heidrich, G. Dugar, J. Vogel, C. M. Sharma, Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. *Methods in molecular biology* **1311**, 1-21 (2015).
 191. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832 (2013).
 192. B. Giardine *et al.*, Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15**, 1451-1455 (2005).

193. B. Cousineau *et al.*, Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**, 451-462 (1998).
194. Y. G. Yoon, J. H. Cho, S. C. Kim, Cre/loxP-mediated excision and amplification of large segments of the Escherichia coli genome. *Genetic analysis: biomolecular engineering* **14**, 89-95 (1998).
195. N. Eroshenko, G. M. Church, Mutants of Cre recombinase with improved accuracy. *Nature communications* **4**, 1-10 (2013).
196. S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, D. L. Court, Recombineering: a homologous recombination-based method of genetic engineering. *Nature protocols* **4**, 206 (2009).
197. J. A. Mosberg, M. J. Lajoie, G. M. Church, Lambda red recombineering in Escherichia coli occurs through a fully single-stranded intermediate. *Genetics* **186**, 791-799 (2010).
198. S. Swaminathan *et al.*, Rapid engineering of bacterial artificial chromosomes using oligonucleotides. *genesis* **29**, 14-21 (2001).
199. K. A. Datsenko, B. L. Wanner, One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences* **97**, 6640-6645 (2000).
200. M. E. Pyne, M. Moo-Young, D. A. Chung, C. P. Chou, Coupling the CRISPR/Cas9 system with lambda red recombineering enables simplified chromosomal gene replacement in Escherichia coli. *Appl. Environ. Microbiol.* **81**, 5103-5114 (2015).
201. S. Shuman, M. S. Glickman, Bacterial DNA repair by non-homologous end joining. *Nature Reviews Microbiology* **5**, 852-861 (2007).
202. Y. Jiang *et al.*, Multigene editing in the Escherichia coli genome via the CRISPR-Cas9 system. *Applied and environmental microbiology* **81**, 2506-2514 (2015).
203. W. Y. Wu, J. H. Lebbink, R. Kanaar, N. Geijsen, J. Van Der Oost, Genome editing by natural and engineered CRISPR-associated nucleases. *Nature chemical biology* **14**, 642 (2018).
204. X. Ao *et al.*, A multiplex genome editing method for Escherichia coli based on CRISPR-Cas12a. *Frontiers in microbiology* **9**, 2307 (2018).
205. M.-Y. Yan *et al.*, CRISPR-Cas12a-assisted recombineering in bacteria. *Appl. Environ. Microbiol.* **83**, e00947-00917 (2017).
206. P. Li *et al.*, Cas12a mediates efficient and precise endogenous gene tagging via MITI: microhomology-dependent targeted integrations. *Cellular and Molecular Life Sciences*, 1-10 (2019).
207. Z. Ren, R. Baumann, L. Black, Cloning of linear DNAs in vivo by overexpressed T4 DNA ligase: construction of a T4 phage hoc gene display vector. *Gene* **195**, 303-311 (1997).
208. S.-Y. Li, G.-P. Zhao, J. Wang, C-Brick: a new standard for assembly of biological parts using Cpf1. *ACS synthetic biology* **5**, 1383-1388 (2016).
209. M. Naville, A. Guillot-Gaudeffroy, A. Marchais, D. Gautheret, ARNold: a web tool for the prediction of Rho-independent transcription terminators. *RNA biology* **8**, 11-13 (2011).
210. S. C. Creutzburg *et al.*, Good guide, bad guide: spacer sequence-dependent cleavage efficiency of Cas12a. *Nucleic acids research* **48**, 3228-3243 (2020).
211. C. Liao, R. A. Slotkowski, T. Achmedov, C. L. Beisel, The Francisella novicida Cas12a is sensitive to the structure downstream of the terminal repeat in CRISPR arrays. *RNA biology* **16**, 404-412 (2019).
212. T. J. Bodine *et al.*, Escherichia coli DNA ligase B may mitigate damage from oxidative stress. *PloS one* **12**, e0180800 (2017).
213. T. Su *et al.*, The phage T4 DNA ligase mediates bacterial chromosome DSBs repair as single component non-homologous end joining. *Synthetic and systems biotechnology* **4**, 107-112 (2019).

-
214. A. Erental, I. Sharon, H. Engelberg-Kulka, Two programmed cell death systems in *Escherichia coli*: an apoptotic-like death is inhibited by the mazEF-mediated death pathway. *PLoS biology* **10**, e1001281 (2012).
 215. K. N. Kreuzer, DNA damage responses in prokaryotes: regulating gene expression, modulating growth patterns, and manipulating replication forks. *Cold Spring Harbor perspectives in biology* **5**, a012674 (2013).
 216. C. Anderson. (2017).
 217. C. Lei *et al.*, The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro. *Nucleic acids research* **45**, e74-e74 (2017).
 218. M. Maresca *et al.*, Single-stranded heteroduplex intermediates in λ Red homologous recombination. *BMC molecular biology* **11**, 54 (2010).
 219. P. Mohanraju, J. v. d. Oost, M. Jinek, D. C. Swarts, Heterologous Expression and Purification of the CRISPR-Cas12a/Cpf1 Protein. *Bio-protocol* **8**, e2842 (2018).
 220. R. Barrangou, P. Horvath, A decade of discovery: CRISPR functions and applications. *Nat Microbiol* **2**, 17092 (2017).
 221. E. V. Koonin, K. S. Makarova, Y. I. Wolf, Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol* **71**, 233-261 (2017).
 222. K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736 (2015).
 223. S. Shmakov *et al.*, Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* **15**, 169-182 (2017).
 224. G. J. Knott, J. A. Doudna, CRISPR-Cas guides the future of genetic engineering. *Science* **361**, 866-869 (2018).
 225. W. Y. Wu, J. H. G. Lebbink, R. Kanaar, N. Geijsen, J. van der Oost, Genome editing by natural and engineered CRISPR-associated nucleases. *Nat Chem Biol* **14**, 642-651 (2018).
 226. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-2586 (2012).
 227. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
 228. J. S. Chen *et al.*, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436-439 (2018).
 229. O. O. Abudayyeh *et al.*, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
 230. A. East-Seletsky *et al.*, Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270-273 (2016).
 231. Z. Wu, H. Yang, P. Colosi, Effect of genome size on AAV vector packaging. *Mol Ther* **18**, 80-86 (2010).
 232. C. A. Lino, J. C. Harper, J. P. Carney, J. A. Timlin, Delivering CRISPR: a review of the challenges and approaches. *Drug Deliv* **25**, 1234-1257 (2018).
 233. W. X. Yan *et al.*, Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88-91 (2019).
 234. D. Burstein *et al.*, New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237-241 (2017).
 235. S. Shmakov *et al.*, Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol Cell* **60**, 385-397 (2015).

236. K. E. Watters, C. Fellmann, H. B. Bai, S. M. Ren, J. A. Doudna, Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science* **362**, 236-239 (2018).
237. E. V. Koonin, K. S. Makarova, F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* **37**, 67-78 (2017).
238. F. A. Ran *et al.*, In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186-U198 (2015).
239. B. Zetsche *et al.*, Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat Biotechnol* **35**, 31-34 (2017).
240. C. Liao *et al.*, Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. *Nature communications* **10**, 1-14 (2019).
241. F. J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740 (2009).
242. R. T. Leenay *et al.*, Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol Cell* **62**, 137-147 (2016).
243. H. Yang, P. Gao, K. R. Rajashankar, D. J. Patel, PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell* **167**, 1814-1828 e1812 (2016).
244. Y. Zhang, R. Rajan, H. S. Seifert, A. Mondragon, E. J. Sontheimer, DNase H Activity of *Neisseria meningitidis* Cas9. *Mol Cell* **60**, 242-255 (2015).
245. E. Ma, L. B. Harrington, M. R. O'Connell, K. Zhou, J. A. Doudna, Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Mol Cell* **60**, 398-407 (2015).
246. W. X. Yan *et al.*, Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88-+ (2019).
247. D. C. Swarts *et al.*, DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258-261 (2014).
248. J. S. Parker, S. M. Roe, D. Barford, Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *The EMBO journal* **23**, 4727-4737 (2004).
249. G. Sheng *et al.*, Structure-based cleavage mechanism of *Thermus thermophilus* Argonaute DNA guide strand-mediated DNA target cleavage. *Proceedings of the National Academy of Sciences* **111**, 652-657 (2014).
250. C. Liao *et al.*, Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. *Nat Commun* **10**, 2948 (2019).
251. J. Garamella, R. Marshall, M. Rustad, V. Noireaux, The all *E. coli* TX-TL toolbox 2.0: a platform for cell-free synthetic biology. *ACS synthetic biology* **5**, 344-355 (2016).
252. R. Marshall *et al.*, Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription-translation system. *Molecular cell* **69**, 146-157. e143 (2018).
253. K. G. Wandera *et al.*, An enhanced assay to characterize anti-CRISPR proteins using a cell-free transcription-translation system. *Methods* **172**, 42-50 (2020).
254. K. S. Whinn *et al.*, Nuclease dead Cas9 is a programmable roadblock for DNA replication. *Scientific reports* **9**, 1-9 (2019).
255. B. P. Kleinstiver *et al.*, Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature Biotechnology* **34**, 869-+ (2016).
256. E. Semenova *et al.*, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108**, 10098-10103 (2011).
257. D. B. T. Cox *et al.*, RNA editing with CRISPR-Cas13. *Science* **358**, 1019-1027 (2017).

-
258. S. Konermann *et al.*, Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* **173**, 665-676 e614 (2018).
 259. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).
 260. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).
 261. W. Dawson, K. Fujiwara, G. Kawai, Y. Futamura, K. Yamamoto, A method for finding optimal RNA secondary structures using a new entropy model (vsfold). *Nucleosides, Nucleotides, and Nucleic Acids* **25**, 171-189 (2006).
 262. A. V. Anzalone, L. W. Koblan, D. R. Liu, Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, 1-21 (2020).
 263. C. Bertrand, A. Thibessard, C. Bruand, F. Lecoite, P. Leblond, Bacterial NHEJ: a never ending story. *Molecular microbiology* **111**, 1139-1151 (2019).
 264. H. H. Chang, N. R. Pannunzio, N. Adachi, M. R. Lieber, Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature reviews Molecular cell biology* **18**, 495 (2017).
 265. P. Verma, R. A. Greenberg, Noncanonical views of homology-directed DNA repair. *Genes & development* **30**, 1138-1154 (2016).
 266. H. A. Rees, D. R. Liu, Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature reviews genetics* **19**, 770-788 (2018).
 267. S. Banno, K. Nishida, T. Arazoe, H. Mitsunobu, A. Kondo, Deaminase-mediated multiplex genome editing in Escherichia coli. *Nature microbiology* **3**, 423-429 (2018).
 268. A. C. Komor *et al.*, Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C: G-to-T: A base editors with higher efficiency and product purity. *Science advances* **3**, eaao4774 (2017).
 269. B. W. Thuronyi *et al.*, Continuous evolution of base editors with expanded target compatibility and improved activity. *Nature biotechnology* **37**, 1070-1079 (2019).
 270. B. P. Kleinstiver *et al.*, Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nature biotechnology* **37**, 276-282 (2019).
 271. X. Li *et al.*, Base editing with a Cpf1-cytidine deaminase fusion. *Nature biotechnology* **36**, 324 (2018).
 272. J. Tan, F. Zhang, D. Karcher, R. Bock, Engineering of high-precision base editors for site-specific single nucleotide replacement. *Nature communications* **10**, 1-10 (2019).
 273. J.-Y. Dong, P.-D. Fan, R. A. Frizzell, Quantitative analysis of the packaging capacity of recombinant adeno-associated virus. *Human gene therapy* **7**, 2101-2112 (1996).
 274. D. Wang, P. W. Tai, G. Gao, Adeno-associated virus vector as a platform for gene therapy delivery. *Nature reviews Drug discovery* **18**, 358-378 (2019).
 275. J. Winter *et al.*, Targeted exon skipping with AAV-mediated split adenine base editors. *Cell discovery* **5**, 1-12 (2019).
 276. J. M. Levy *et al.*, Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nature Biomedical Engineering* **4**, 97-110 (2020).
 277. M. G. Kluesner *et al.*, EditR: a method to quantify base editing from Sanger sequencing. *The CRISPR journal* **1**, 239-250 (2018).
 278. L. W. Koblan *et al.*, Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nature biotechnology* **36**, 843-846 (2018).

279. S. Jin *et al.*, Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292-295 (2019).
280. E. Zuo *et al.*, Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289-292 (2019).
281. R. Verwaal, N. Buiting–Wiessenhaan, S. Dalhuijsen, J. A. Roubos, CRISPR/Cpf1 enables fast and simple genome editing of *Saccharomyces cerevisiae*. *Yeast* **35**, 201-211 (2018).
282. C. Verduyn, E. Postma, W. A. Scheffers, J. P. Van Dijken, Effect of benzoic acid on metabolic fluxes in yeasts: a continuous–culture study on the regulation of respiration and alcoholic fermentation. *Yeast* **8**, 501-517 (1992).
283. J. T. Pronk, Auxotrophic yeast strains in fundamental and applied research. *Applied and environmental microbiology* **68**, 2095-2100 (2002).
284. R. D. Gietz, R. A. Woods, in *Methods in enzymology*. (Elsevier, 2002), vol. 350, pp. 87-96.
285. P. Pausch *et al.*, CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* **369**, 333-337 (2020).
286. B. Al-Shayeb *et al.*, Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-431 (2020).
287. H. Lee, Y. Dhingra, D. G. Sashital, The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, e44248 (2019).
288. K. S. Makarova, L. Aravind, Y. I. Wolf, E. V. Koonin, Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology direct* **6**, 38 (2011).
289. S. Silas *et al.*, Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. *Elife* **6**, e27601 (2017).
290. J. Wang *et al.*, Limits in accuracy and a strategy of RNA structure prediction using experimental information. *Nucleic acids research* **47**, 5563-5572 (2019).
291. P. Liu *et al.*, Enhanced Cas12a editing in mammalian cells and zebrafish. *Nucleic acids research* **47**, 4169-4180 (2019).
292. F. Teng *et al.*, Enhanced mammalian genome editing by new Cas12a orthologs with optimized crRNA scaffolds. *Genome biology* **20**, 1-6 (2019).
293. T.-y. Chyou, C. M. Brown, Prediction and diversity of tracrRNAs from type II CRISPR-Cas systems. *RNA biology* **16**, 423-434 (2019).
294. H. Nishimasu *et al.*, Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014).
295. C. Huai *et al.*, Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nature communications* **8**, 1-9 (2017).
296. E. V. Koonin, K. S. Makarova, Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome biology and evolution* **9**, 2812-2825 (2017).
297. W. Bao, J. Jurka, Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mobile DNA* **4**, 12 (2013).
298. V. V. Kapitonov, K. S. Makarova, E. V. Koonin, ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *Journal of bacteriology* **198**, 797-807 (2016).
299. I. Mougias, E. F. Bosma, J. Ganguly, J. van der Oost, R. van Kranenburg, Hijacking CRISPR-Cas for high-throughput bacterial metabolic engineering: advances and prospects. *Current opinion in biotechnology* **50**, 146-157 (2018).
300. B. Adiego-Pérez *et al.*, Multiplex genome editing of microorganisms using CRISPR-Cas. *FEMS microbiology letters* **366**, frnz086 (2019).

-
301. P. D. Hsu, E. S. Lander, F. Zhang, Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278 (2014).
 302. A. Pickar-Oliver, C. A. Gersbach, The next generation of CRISPR-Cas technologies and applications. *Nature reviews Molecular cell biology* **20**, 490-507 (2019).
 303. P. Cameron *et al.*, Harnessing type I CRISPR-Cas systems for genome engineering in human cells. *Nature biotechnology* **37**, 1471-1477 (2019).
 304. E. De Dreuzy *et al.* (American Society of Hematology Washington, DC, 2019).
 305. J. Wang, A. Lu, J. Bei, G. Zhao, J. Wang, CRISPR/ddCas12a-based programmable and accurate gene regulation. *Cell discovery* **5**, 1-4 (2019).
 306. Y. Liu *et al.*, Engineering cell signaling using tunable CRISPR-Cpf1-based transcription factors. *Nature communications* **8**, 1-8 (2017).
 307. Y. E. Tak *et al.*, Inducible and multiplex gene regulation using CRISPR-Cpf1-based transcription factors. *Nature methods* **14**, 1163-1166 (2017).
 308. X. Wang *et al.*, Cas12a Base Editors Induce Efficient and Specific Editing with Low DNA Damage Response. *Cell Reports* **31**, 107723 (2020).
 309. S.-Y. Li *et al.*, CRISPR-Cas12a has both cis- and trans-cleavage activities on single-stranded DNA. *Cell research* **28**, 491-493 (2018).
 310. J. P. Broughton *et al.*, CRISPR-Cas12-based detection of SARS-CoV-2. *Nature Biotechnology*, 1-5 (2020).
 311. J. S. Gootenberg *et al.*, Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438-442 (2017).
 312. M. J. Kellner, J. G. Koob, J. S. Gootenberg, O. O. Abudayyeh, F. Zhang, SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature protocols* **14**, 2986-3012 (2019).
 313. J. S. Gootenberg *et al.*, Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439-444 (2018).
 314. L. Li *et al.*, HOLMESv2: a CRISPR-Cas12b-assisted platform for nucleic acid detection and DNA methylation quantitation. *ACS synthetic biology* **8**, 2228-2237 (2019).
 315. J. Strecker *et al.*, Engineering of CRISPR-Cas12b for human genome editing. *Nature communications* **10**, 1-8 (2019).
 316. N. M. Gaudelli *et al.*, Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017).
 317. A. Lapinaite *et al.*, DNA capture by a CRISPR-Cas9-guided adenine base editor. *Science* **369**, 566-571 (2020).
 318. L. Gao *et al.*, Engineered Cpf1 variants with altered PAM specificities. *Nature biotechnology* **35**, 789-792 (2017).
 319. T. Hyodo *et al.*, Tandem paired nicking promotes precise genome editing with scarce interference by p53. *Cell Reports* **30**, 1195-1207. e1197 (2020).
 320. A. E. Trevino, F. Zhang, in *Methods in enzymology*. (Elsevier, 2014), vol. 546, pp. 161-174.
 321. A. V. Anzalone *et al.*, Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019).
 322. H. Zhang, Z. Li, R. Xiao, L. Chang, Mechanisms for target recognition and cleavage by the Cas12i RNA-guided endonuclease. *Nature Structural & Molecular Biology*, 1-8 (2020).

About the author



Wen Ying Wu (16 - 11 - 1991), of Chinese descent, was born and raised in Aruba. In 2010 She came to Wageningen for the biotechnology BSc programme. During her BS, she went to Cornell University in the United States for a minor in wine and beer technology. For her BSc thesis, she studied the effects of ATP citrate lyase on the acid production in *Aspergillus niger*. After obtaining her BSc degree in biotechnology in 2013, she started her MSc in Biotechnology at Wageningen University. During her MSc, she participated in the 2014 iGEM competition - the international Genetically Engineered Machine Competition / a

synthetic biology competition, of which she was also team captain. As part of her MSc thesis, she worked on the project known as “BananaGuard”. Her team was very successful and got 2nd place worldwide for the graduate division. After iGEM, Wen did her internship at Imperial College London and finished her MSc in Molecular Biotechnology in the fall of 2015.

In March 2016, she started her PhD research in the Laboratory of Microbiology at Wageningen University under the supervision of Prof. John van der Oost and Dr. Raymond Staals. During her PhD, she studied type V CRISPR-Cas systems. In particular, she studied the molecular mechanism of type V CRISPR-Cas systems, and how they function as an adaptive immune system in bacteria against viruses. Using the proteins that she studied, she also repurposed them into tools for genome engineering. Most of her work can be found in this thesis.

Completed training activities

Discipline specific activities

Meetings & conferences

- CRISPR Conference, Rehovot, Israel (2016)
- Host Pathogen Conference, Wageningen, The Netherlands (2016)
- CRISPR Conference, Montana, U.S.A. (2017)* + **
- NWO Chains, Veldhoven, The Netherlands (2017) **
- Microbiology Centennial, Wageningen, The Netherlands (2017)
- Host Microbe Genetics Meeting, NWO, Wageningen, The Netherlands (2017)
- CRISPR Conference, Vilnius, Lithuania (2018) *
- CRISPR Conference, Quebec, Canada (2019) * + **

Courses

- Python for Life scientist, Enpicom, Amsterdam, The Netherlands (2016)
 - Hands-on Flow Cytometry - Learning by Doing!, EMBL, Heidelberg, Germany (2016)
 - Synthetic Biology in Action: Programming Bacteria to Do Amazing Things, EMBL, Heidelberg, Germany (2017) * + **
 - Bioinformatics in Linux and Python, Wageningen, The Netherlands (2020)
- *poster presentation ** poster pitch ** oral presentation

General courses

- Presenting with Impact, Wageningen, The Netherlands (2016)
- Competent assessment, Wageningen, The Netherlands (2016)
- Scientific Publishing, Wageningen, The Netherlands (2017)
- Scientific Writing, Wageningen, The Netherlands (2017)
- Famelab Wageningen pitch workshop, Wageningen, The Netherlands (2018)
- Famelab Netherlands pitch workshop, Amsterdam, The Netherlands (2018)
- Faces of Science, science communication workshop, Amsterdam, The Netherlands (2018)
- Writing Grant Proposals, Wageningen, The Netherlands (2018)
- Career assessment, Wageningen, The Netherlands (2018)
- Brain friendly working and writing, Wageningen, The Netherlands (2018)
- Career Perspective, Wageningen, The Netherlands (2019)

Optionals

- Preparation of research proposal
- Bacterial Genetics group meetings, Wageningen, The Netherlands
- Microbiology PhD meeting, Wageningen, The Netherlands
- Microbiology seminars
- PhD representative at the Microbiology Daily Board, Wageningen, The Netherlands (2018)
- Organizing committee, PhD study trip to Germany, Sweden and Denmark (2017)
- PhD study trip to Boston and New York, U.S.A. (2019)

Co-author affiliation

Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University

6703 HB Wageningen, The Netherlands

John van der Oost, Raymond H.J. Staals, Prarthana Mohanraju^a, Sjoerd C.A., Rob Joosten, Karlijn Keessen, Suzan Yilmaz, Jorik Bot, Tahseen S. Khan, Stijn Prinsen, Belén Adiego-Pérez, Timon Lindeboom

Hubrecht Institute for Developmental Biology and Stem Cell Research, University Medical Center Utrecht and Faculty of Veterinary Medicine, Clinical Sciences of Companion Animals, Utrecht University

3584 CT Utrecht, The Netherlands

Niels Geijsen ^a

Department of Molecular Genetics Erasmus MC, University Medical Center Rotterdam

3015 GD Rotterdam, the Netherlands

Joyce H. G. Lebbink, Roland Kanaar, Charlie Laffeber

Department of Radiation Oncology, Erasmus MC, University Medical Center Rotterdam

3015 GD Rotterdam, the Netherlands

Joyce H. G. Lebbink, Roland Kanaar

Oncode Institute

3521 AL Utrecht, The Netherlands

Roland Kanaar

Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology

2629 HZ, Delft, the Netherlands

Stan Brouns, Cristóbal Almendros^b

Department of Microbiology and Immunology, University of Otago

9054 Dunedin, New Zealand

Simon A. Jackso

Broad Institute of MIT and Harvard

Cambridge, MA 02142, USA

Bernd Zetsche, Feng Zhang, Matthias Heidenreich, Jeroen Kneppers, Ellen M. DeGennaro, Nerges Winblad, Sourav R. Choudhury, Omar O. Abudayyeh, Jonathan S. Gootenberg, David A. Scott

McGovern Institute for Brain Research, MIT

Cambridge, MA 02139

Bernd Zetsche, Matthias Heidenreich, Iana Fedorova, Nerges Winblad, Sourav R. Choudhury, Omar O. Abudayyeh, Jonathan S. Gootenberg

Department of Brain and Cognitive Sciences, MIT

Cambridge, MA 02139

Bernd Zetsche, Matthias Heidenreich, Iana Fedorova, Nerges Winblad, Sourav R. Choudhury, Omar O. Abudayyeh, Jonathan S. Gootenberg

Department of Biological Engineering, MIT

Cambridge, MA 02139

Bernd Zetsche, Matthias Heidenreich, Iana Fedorova, Nerges Winblad, Sourav R. Choudhury, Omar O. Abudayyeh, Jonathan S. Gootenberg

Skolkovo Institute of Science and Technology

Skolkovo, 143025, Russia

Iana Fedorova, Konstantin Severinov

Harvard-MIT Division of Health Sciences and Technology, MIT

Cambridge, MA 02139

Ellen M. DeGennaro, Omar O. Abudayyeh

Waksman Institute for Microbiology, Rutgers, The State University of New Jersey

Piscataway, NJ 08854, USA

Konstantin Severinov

Institute of Molecular Genetics, Russian Academy of Sciences

Moscow, 123182, Russia

Konstantin Severinov

Arbor Biotechnologies

Cambridge, MA 02139, USA

Winston X. Yan, David A. Scott

Helmholtz Institute of RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI)

97080 Würzburg, Germany

Chase Beisel, Chunyu Liao

National Center for Biotechnology Information, National Library of Medicine, National Institute

Bethesda, MD 20894, USA

Kira S. Makrova, Eugene V. Koonin

Current Addresses

^a Developmental Biology and Regenerative Medicine and Head, Department of Anatomy and Embryology, Leiden University Medical Center

2333 ZA Leiden, The Netherlands

Niels Geijsen, Prarthana Mohanraju

^b GenScript Biotech

2333 CG, Leiden, The Netherlands

Cristóbal Almendros Romero

List of publications

Wen Y. Wu, Simon A. Jackson, Cristóbal Almendros, Suzan Yilmaz, Rob Joosten, Stan J.J. Brouns, John van der Oost, Raymond H.J. Staals. *Adaptation in type V-A and type V-B CRISPR-Cas systems. Manuscript in preparation.*

Wen Y. Wu*, Prarthana Mohanraju*, Sjoerd C. A. Creutzburg, Karlijn Keessen, Tahseen S. Khan, Stijn Prinsen, Winston X. Yan, Chunyu Liao, Kira S. Makarova, David A. Scott, Chase L. Beisel, Charlie Laffeber, Joyce H.G. Lebbink, Eugene V. Koonin & John van der Oost. *Characterizing a compact CRISPR-Cas12u1 enzyme. Manuscript in preparation.*

Wen Y. Wu, Sjoerd C.A. Creutzburg, Belén Adiego-Pérez, Timon Lindeboom, Karlijn Keessen, John van der Oost. *Small and mighty: MmuCas12u1 C-to-T and A-to-G base editors. Manuscript in preparation.*

Sjoerd C A Creutzburg, **Wen Y Wu**, Prarthana Mohanraju, Thomas Swartjes, Ferhat Alkan, Jan Gorodkin, Raymond H J Staals, John van der Oost (2020). *Good guide, bad guide: spacer sequence-dependent cleavage efficiency of Cas12a. Nucleic Acids Research*, 48(6), 3228-3243. DOI 10.1093/nar/gkz1240.

Bernd Zetsche*, Matthias Heidenreich*, Prarthana Mohanraju*, Iana Fedorova, Jeroen Kneppers, Ellen M DeGennaro, Nerges Winblad, Sourav R Choudhury, Omar O Abudayyeh, Jonathan S Gootenberg, **Wen Y Wu**, David A Scott, Konstantin Severinov, John van der Oost & Feng Zhang (2017). *Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nature Biotechnology*, 35(1), 31-34. DOI 10.1038/nbt.3737.

Wen Y. Wu, Joyce H. G. Lebbink, Roland Kanaar, Niels Geijsen & John van der Oost (2018). *Genome editing by natural and engineered CRISPR-associated nucleases. Nature Chemical Biology*, 14, 642–651. DOI 10.1038/s41589-018-0080-x.

*equal contribution

Patent Applications

J. van der Oost, **W.Y. Wu**, P. Mohanraju, S.C.A. Creutzburg. *Type V-U CRISPR system. UK1909597.5 (filed July 2019)*

Acknowledgements

As I am writing this piece, I feel many things. I feel relieved that the hardest part of the PhD is behind me, I feel glad that I am closer to obtaining my doctorate title, I feel sad that my PhD journey is coming to an end, but I also feel ecstatic to have gotten this far! Most importantly, I feel grateful to the people that were with me throughout this amazing journey. I believe that your surroundings shape who you are, and I would have to say, I am surrounded by a great group of people. In this chapter I give my gratitude to those great people.

John, thank you for hiring me as one your PhD students in the BacGen group, a group that I am so proud to be a part of! I always say that BacGen is like family, a science-family that you've had put together. Thank you for all the great opportunities and wonderful memories that you have given me throughout the years. Some examples are, me presenting at the opening of the academic year, our short talk to the king where you doubted whether you should write "Lieve Koning....". Our excursion with Mark Young to get samples in Yellowstone and then hanging out in the "hot springs". My favorite memories are the endless long meetings, discussing odd scientific results, thinking of experimental designs to prove our hypothesis and brainstorming crazy ideas for future projects. Thank you for your time and patience over the years. You have taught me a great deal about science, I am who I am today largely due to you and I am glad to have had you as my supervisor! I wish you a future full of exciting scientific breakthroughs, but also relaxing family time with **Paulien**, your four **sons** and the sweetest dog in the world, **Charlie**.

Raymond, a.k.a. my adaptation bruh/swa, thanks for being my co-promoter and mentor. For teaching me all about the world of CRISPR-Cas adaptation! I always appreciate your love for fundamental research and finding the biological relevance in everything. You have also sparked my interest in programming by showing me how a few lines of codes can do so much for you. I have thoroughly enjoyed our time together talking about science, teaching AMM and our occasional lunch hang outs (in other words, Wen barging into Raymond's office while he is having lunch), where I watch you defeat all the balloons. It is always great joy and fun working together with you. Thanks for all the fun and science **Ray**. KThnxBye!

Stan, thank you for taking me in into your CRISPR-group meetings when I started my PhD. I have learned a lot about CRISPR-Cas during those meetings and from your advice and suggestions during the time you were here in Wageningen. **Thijs E.**, thank you for being the new chair of the laboratory of Microbiology. It has been a pleasure working together with you in the daily board. You have done such an amazing job managing Microbiology while also establishing MicEvo. **Willem**, it has been an honor being part of Microbiology while you were chair. Never have I met such a productive, quick thinking and efficient person. You sir, are a rare specimen. You have told me once the reason why science communication is very important. It is taxpayers fund our research, therefore we are responsible in communicating that research back to

them. This is something I always keep in my mind and continue to strive to be a good science communicator. Thank you to the thesis committee: **Michiel Kleerebezem, Dolf Weijers, Gorben Pijlman and Chirlmin Joo** that took their time to evaluate my PhD thesis.

To my collaborator **Niels, Joyce and Ronald**, it was a nice experience writing the review together you, I have learned so much about Cas protein delivery and repair during that time. Looking forward to our future collaborations together. **Simon**, thank you helping with analyzing the NGS data, I have a really good feeling that the next data set is going to be amazing. Thank you for teaching me how to use R studio and the in and outs of analyzing spacers. I have always enjoyed our interactions during all the CRISPR meetings. Can't wait to some day visit you in New Zealand. **Cristobal**, thank for the great collaboration of the adaptation project and it was very nice working with. Whenever we are not in meetings or working with the blue pippin, we would be have conversation on which video games we should play. **Feng**, thank you for the great collaboration on the Cas12a multiplex project. I am always so impressed on how fast you lab moves. **Eugene and Kira**, Thank you for all the help regarding the bioinformatics of MmuCas12u1. **Winston and David**, thank you for sequencing the PAM library, because of that we were able to start further characterizing the MmuCas12u1 protein. **Christian S.**, thank you my tall friend for teaching me how to use the cell sorter at BPE, a.k.a. Sonny. It has been lots of fun with working with you, while we wait for Sonny to calibrate. **Chase, Chunyu, Anzhela and Katarina**, thank you for running the TXTL and northern blot experiments for the Mmu project, I really believe that we can crack it! Also, thank you for taking me in for two weeks and teaching how to properly run TXTL. I have really enjoyed my time there together the rest of the group. **Johannes, Koen and Simon**, it is great to be able to work on the single molecule tracking with all of you. Looking forward to the great results to come.

Costas, the brave person that goes to CrossFit with me every day at 7:00 am, leaving 6:30 am at Helix and coming back at 8:30 am at work, 5 days a week. Thank you for always being there for me, through tough physical workouts in CrossFit where you yell "COME ON WEN!" and through some sad emotional times where you hear me out. I admire your motivation and drive to be the best that you can possibly be, both in life and in science. You are talented in many ways, both intellectually and physically. You can do a handstand walk while being intoxicated and you also have the best biceps (dear reader, please ask Costas about his biceps). Sometimes, we can have heated discussions, but discussions remain discussions and that is what true friends are. I can't wait to see you launch a successful start-up and bring it to market one day. I truly believe that it can and will happen, it's only a matter of time. doooooeeeeiiiiiiilolololol.

Jeroen, my dear friend, you are that person that I would always go to "hang out" whenever I have some waiting time for my experiments. I appreciate your modesty and willingness to help those who need it. I am always amazed on how stress-free you are in everything that you do and that you have such a great work life balance. Often in stressful situations I ask myself: "what would **Jeroen** do?". I have thoroughly enjoyed our lunches together where we talk about the environment and how butter

is the best topping to put on raisin bread (not). We may not agree on the raisin bread debate, but I really look up to you as a person and I can't wait to have our path cross again in our scientific career.

Prarthana(aaaaaaa), I know you do not like hearing it, but I will say it anyways, thank you for being my work-wife. *add Shakespeare voice here*. "You are the moon to my sun, my *in vitro* to my *in vivo*!". It has been a truly amazing time working together with you, a time full of fun, hard work and cool science all mixed in together. We can always jump from a serious scientific discussion, to casual talks, to joking around, to deep societal issues in just one conversation. Thank you for showing me the world of protein purification. Even after you have left the lab, you have constantly supported me throughout the writing of my thesis, answering me silly questions, giving me advice, and helping me wrap up. I have always enjoyed our work/catch-up skype sessions that we have. You know, those meetings that never last just 15 min. Thank you for being such a great colleague/work-wife and a great friend. I wish you and **Ashwin** all the best in the future!

SJOERD, "Guten morgen / guten tag mein Freund!" Since the beginning of my PhD we have always been working side by side, with both our computer desks and lab benches being next to each other. This such closeness can lead to two things, us getting along well or us hating each other. I am happy to say that for us, it was the first and with added bonus laughter and fun. You always surprise me with fun facts about random things in our endless conversations. We played this game a lot, where we would constantly argue and discuss what the "best" approach is for a project. Together I think we were able to create not the best, but the ultimate approach! This game not only made me constantly criticize my own work, but made me a better researcher in general. I am glad to have had you as my office mate, my colleague, my mentor and my friend.

Lot(jes), my bra, my partner in crime, my Olympic weightlifting buddy. We knew a little of each other prior to the DB, but during our time at the DB we became inseparable! I always like your motivation and efforts to do good for others. Your selfless acts of kindness constantly warms my heart. Together we are the dynamic duo! Lot & Wen! No matter what we did, we'd always have fun together. I look forward to all of our future CrossFit workouts, weightlifting competitions, and pancake days together.

Rob, thanks for being our technician in BacGen and also for being a good friend. Not only did you teach me a lot about of the FPLC, you have also helped me on multiple projects. I enjoy hanging out with you either during lunch or when we have a short a break where you tell me all about your latest miniature creations. I wish you, **Aniek** and **Evy** all the best in the future.

Thanks you all the great office mates that had the pleasure of sharing an office with over the last few years. **Irene** and **Diana**, you both were my first office mates in my first few weeks. You were two happy, enthusiastic and hard working women that take great joy in their work, who make great role models for any future female scientist.

Alex, thank you for all the fun times at Microbiology. I have especially enjoyed the pranks and the drone flying which turned out not to be so good for the a certain plant. You were always a great colleague to be around and happy to have shared Dreijen's biggest office with you. *takes chair, steps on chair, pats Alex's head* **Tom de W.**, we were always the first ones in early in the morning, which meant us hanging out in the lab. I loved that we can always just make fun of each other and crack jokes and joke around. Your great presence (though sometimes crazy) in the lab always made me happy. I wish you well in the future together with your wife **Juliette, Sophie and Koen.** **Tijn**, thanks for the great conversation in the office. **Ioannis**, though we might have had our downfall, I do appreciate your insane work ethic and your passion for science.

Marnix, thanks for starting all the philosophical discussion in the office. You are a very creative person and I am always so amazed of all the cool things you can make out of wood. **Jorrit**, thank you for all the fun, jokes and games in the office. I always love to hear about your new "cash cow" stories. **Yifan**, the man that enjoys good things in life, good drinks and food. Because of you, I appreciate whiskey a lot more.

Despiona, I love how enthusiastic you can get about science. Your motivation, hard work and perseverance is something to admire. May you continue to do the science you love and follow your dreams. **Belen**, thanks for being the sweet nice person that you are. You are really a top notch researcher, an amazing teacher, an excellent collaborator and a great friend. I am glad that you have joined our office, It is always nice to have our short coffee breaks together with some nice chats in the office. **Thomas**, it has been great to have you in the office and I am very happy to always have someone to talk to when it comes to the soothing sound of mechanical keyboard, making figures or coding a python script. Your notoriously dry and awkward jokes often make me facepalm, but do make us all have a good laugh in the office. **Maartje**, though you have started your postdoc during the more "restricted" times of Microbiology. It has been nice to share an office with you. You seem to have a knack for giving the best names to your cats. Thank you for the advice on the layout of my thesis

To my BacGen colleagues, whom I have gotten to know over the past few years. **Joyshree**, thanks for the nice chats in the corridor and introducing me to delicious Indian food. Wish you all the best in your future job at Jansen, can't wait to hang out and have some nice Indian food again. It was great when **Prarthana, Belen**, you and I went to the Chainsmokers concert. Best thing about that day? Those amazing Indian Dosas. **Mihris**, I love how calm and down to earth of a person you are. It is always a pleasure to encounter you either in the corridor or in the lab. The way you greet me and say "goooooood" always makes my day. **EriC** with a C, it has been great to see you grow from a Msc student in our lab, to my colleague. Thank you for getting me into the hobby of longboarding and introducing me to one of the best things you can do in the summer.

Ca**o! **Lorenzo**, that is meant in a good way. I enjoyed all the laughs and good times in the lab. May you one day obtain all the records your heart desires. I also wish to one day taste this amazing pizza you've been talking about! **Thijs N.**, thank you for all the nice chats and insights on possible get rich scheme. I can see you owning a very successful business the future. Also, thanks for the help with troubleshooting my script and giving tips and advice on python. **Joep**, it was great to have you as my across the bench neighbor. Having nice conversations while both of us are pipetting. **Max**, thank you for all the good times inside and outside of the lab. You are an amazing and one of a kind human being, never change who you are. Thanks for being a fun conference buddy in Quebec. I can't wait to try all your other fermented beverages. **Jurre**, Thanks for all the nice lab chats and maybe I one day I will be able to try your famous fried chicken. I wish you, **Stijn**, **Niek** and **Bart** all the best in the future with Scope Biosciences. **Janneke**, I enjoy your efforts to always bring in either horizontal and vertical integration in the HELIX. I am sad that we never got to try out your elevator pitch idea.

Eugenios, thank you for working with me on the base editor project. I always enjoy seeing you present your fluorescent droplets. Best of luck with your *in vitro* Cas protein selection in droplets. **Carina**, you are a lady with many talents. I am so impressed that apart from your PhD you have many "side-hustles". I am also really happy that you have joined CrossFit. **Stijn**, thanks for us with running some assays for the Mmu project. It is always nice to see you going from an iGEM student to a researcher in Scope Biosciences. **Catarina**, thank you for helping me out with some NGS data and for the fun longboarding sessions in the parking garage. **James**, thank you for always being so nice and king whenever we encounter each other. I am always confused whether I should speak English or Dutch to you. Sometimes we do a mix of both and I love it.

Guus, thank you helping me with the FPLC when I needed it and ordering my precious primers. You always keep my reflexes in check by throwing random objects at me to catch. **Nico**, thank you for being my supervisor during my MSc thesis. It is always nice to talk to you in the lab and now in the corridor. Fun fact: I actually don't think you talk as much as people say you do. Glad to have you back. **Servé**, thank for nice conversations on the lunch table. **Adini**, thank you for the cactus plant, it's still alive, surprisingly. **Mark Young**, it was a great having your energetic self, walking around the lab. Happy that you came to Wageningen for your sabbatical. **Hanne**, I always enjoy our half Dutch half English conversation. Best of luck in the future with your new job. **Teunke**, thank you for all the nice lunches and just general conversation together. It has always been fun having you and **Mark** for our fun pub quizzes. Wish you, **Mark** and **Aiven** here a wonderful future together. **Gioavanni**, thank you for the amazingly delicious Italian lunches that brought me a taste of Italy.

To my other past BacGen colleagues, **Brenda**, **Elleke**, **Lione**, **Mamou**, **Ismael**, **Tim**, **Melvin**, **Jie**, **Stijn de V.**, **Jie**, thank you all for being part of BacGen and making BacGen such an amazing group to be in.

Carrie, thanks for kicking my butt during our bike ride and indulging me in my unicorn obsessions. Always nice to have a chat with you in the corridor. **Emmy**, you are a sweet and amazing soul, I always love how you try your absolute best to just do good in the world. Thank you for showing me how amazing vegan dishes can taste. **Giannis**, thank you for helping me out and giving me tips on wrapping up my thesis. It is always a great pleasure to see you at the box and having a quick chat together. **Fons**, thank you for all the coffee corner chats and the mug. I love it and use it every day!

Thank you to all my students that aided me in my research, whom I had the pleasure to teach and help develop their scientific skills. Without you guys, I don't think this thesis would be what it is today, thank you for all the work and the good times in the lab together. **Meral** and **Alex B.**, thanks for starting the adaptation work with me. You both have shown then and there that spacers were being acquired, just it never had the right PAM. **Meral**, I have thoroughly enjoyed our conversations on music and dance, your knowledge on hip hop is astounding. **Alex B.**, I really loved your enthusiasm for the project and all your questions really improved me as a supervisor. **Jorik** and **Bernard**, thank you for working on the cut & paste genome editing tool with me. Thank you for all the nice conversation in the lab. **Pilar**, thank you for your hard work on the adaptation detection system and the type V anti-CRISPR. No matter how tough things got in the lab, you always kept on going, something I really admired. Thank you for introducing me to these very addictive Mexican mangoes. **Laura**, thank for starting the ground work of the V-B adaptation project. Thank you for always being so kind and sweet. **Maaïke**, thank you for working on the Pdi Cas12u3 with me. You have shown me that miniprepping plasmids is not just a chore, but that it can be very therapeutic. **Suzan**, thank you for your help in bringing all adaptation modules (both V-A and V-B) to a whole new expression system. It is great to have you back in BacGen as a colleague. **Karlijn**, thank you for characterizing the Mmu protein with me. Countless hours spent at the cell sorter at BPE, that in the end led us to finding a PAM for Mmu. In addition, together we were able to test out the first Mmu Base editor. Thank you for surprising me with a game & food care package on a Saturday morning, when I was working with the cell sorter at BPE. **Suzan** and **Karlijn**, thank you for all the yummy desserts and the nice boardgames, may we continue to have those in the future. **Laurens**, thank you for all your hard work in testing all the different spacer lengths for Mmu. **Timon**, thanks for characterizing the different C to T Mmu base editors. It was hard getting a couple of constructs done, but it went very well in the end. It is also incredible that you were able to follow me and **Sjoerd**, when we have our back and forth discussion moments! **Efthymios**, thank you for the enthusiasm in the A to G base editor project and laying the ground works for the project. These were hard fusion proteins to make with tricky linker sequences, but you managed to construct and test them. **Daan**, thank you for helping me further characterize and validate the Mmu Base editors. Even with restricted access and various uncertainties, I am happy that you were able to finish your project. Ola **Ricardo**, thanks for helping with characterizing the A to G base editors. I always like your enthusiasm to create and engineer new proteins. During the time when the lab was closed, I had lots of fun staring at protein structures with you

and thinking on strategies on how to engineer them. **Ezra**, thank you for setting up the initial plasmids for the single molecule project. You are one of the most quick thinking students I know. **Cleo**, thank you continuing the single molecule project. Glad to have you on board the project.

Apart from my own work, being part iGEM will always remain something very dear to my heart. I am grateful to have supervised the iGEM teams of 2017 MANTIS and 2019 XYLENCER. Thank you to the 2017 supervision team: **Raymond, Christian F., Rob, Marta, Emma, Prarthana** and **Rik**. Thank you to very talented team **Niek, Stijn, Sabine, Linda, Jurre, Bart, José, Mark, Tom** and **Natalie**. You have done amazingly well in the jamboree! The Mantis costume was such a crowd pleaser. Thank you for the 2019 supervision team: **Raymond, Jasper, Carina, Jurre, Rik, Despiona, Maria, Enrique, Lyon** and **Joep**. Thank you to the also very talented team: **Cleo, Alex, Marijn, Alba, Robert, Ben, Santi, Hetty, Niels, Sebastian** and **Dannie**. Your achievements at the jamboree was astounding! I really enjoyed watching the voice of iGEM (on replay). To my iGEM students, **Tom, Stijn, Jurre** and **Marijn**. Thank you for being the great students that you were. Supervising you guys has been a blast. To both **iGEM teams**, achievements aside, it is your creativity, perseverance and enthusiasm that has made me so proud to have been your supervisor.

During my time at Microbiology I had the opportunity to organize different activities together with some great people. To the BBQ committee, **Hugo, Daan** and **Wim**, it was nice organizing such a delicious and fun activity together. To the drinking group! **James** and **Carrie**, thanks for making sure that the people at MIB always have snacks and (non)alcoholic beverages for the Friday afternoon drinks. **Ioannis M., Johanna, Jeroen, Aleks, Joyshree, Prarthana** and **Erikal** had a great time organizing the 2017 MIB SSB PhD trip together with all of you. **Caifang, Tika, Jorrit, Ran, Nong, Emma, Martijn, Ruben, Hugo, Erica, Emmy, Aleks, Giannis, Indra, Jeroen, Joyshree** and **Daan** for the great company during the MIB SSB PhD trip 2017. Our accommodation and time in Hamburg was, an experience. Thank you **Irene** and **Richard** for being such nice supervisors. Thank you **Costas, Ivette, Lot, Giannis, Despiona, Carrie, Max, Thijs, Christos, Enrique, Maria, Menia, Rik, Janneke, Catarina, Caifang, Ran, Joep, Wasin, Mamou, Martha, Nong, Diana** and **Raymond** for being great travel companions for our great city adventure, MIB SSB 2019 PhD trip in Boston and New York.

To the PhD board: **Patrick, Peter, Nico V., Carrie, Marie-Luise, Catarina, Thijs, Max, Lot, Jolanda** and **Marina** thank you for setting up the PhD board together. May the PhD board continue to hear, transmit and initiate.

To the daily board: **Diana, Hauke, John, Thijs E., Willem, Caroline, Phillipe, Steven, Erwin, Servé, Gerben, Anja, Heidi, Sjon**, and **Lot**, thank for all the productive meetings on Wednesday morning. Thank for making it such a nice atmosphere, with some chuckles here and there. To **Anja, Hannie, Detmer, Nico C., Nico V.** and **Patricia**, it has been very nice organizing the online MIB get together 2020 with all of you.

A really big thanks to the technician team at Microbiology, **Philippe, Rob, Guus, Hans, Ineke, Steven, Ton, Iame, Sjon Tom de W., and Wim**, the lab would be nothing without you guys keeping everything safe and in check. I would also like to thank **Tom S.**, who is always a helpful neighbor whenever you need him. Thank you for constantly helping me out, starting from when I was a bachelor student learning how to set-up a bioreactor, to a PhD student asking where things are and how they work.

Thank you **Hannie, Anja** and **Heidi** for all your help with all bureaucratic related tasks. It is always a pleasure stepping into your office and be greeted with you smiles. I also enjoy our short chats afterwards, of which always ends in laughter. I would say “Het is altijd gezellig bij jullie!”

I would also like to thank **Gosse, Philippe, Wim, Corline, Servé** and **John** for teaching about microbiology during my time as a BSc or MSc student. Your enthusiasm for the subject has inspired me to continue to learn and explore the microbial world. Thank you for being great teachers and hope you continue to inspire future students like myself.

To the rest of my colleagues at Microbiology, **Ran, Caifang, Yuan, Sudarshan, Johanna, Menia, Taojun Chen, Martijn, Reinier, Ruth, Will, Felix, Detmer, Guillaume, Ivette, Yuan, Nam, Aleks, Indra, Peter, Daan, Hikmah, Martha, Jolanda, Laura, Prokopis, Ineke, Steven, Gerben, Jie, Peng, Hauke, Tika, Nam, Iame, Hans** thank you all for the nice corridor and coffee corner conversations that we’ve had over the years.

To the great group of people at Systems and Synthetic Biology group. You guys have taken me in as one of your own. Thank you for all the nice lunches, coffee breaks and cakes! **Bastiaan**, there is a reason why you were the voted the most popular PhD student of SSB! Some say it was a conspiracy, but I think you are worthy of that title. You are one of the most social and inclusive person of SSB in my opinion =). **Niru, Nhung** and **Erica**, thank for all the great times hanging out together, either in your office, canteen or outside of work. Thanks for all nice Indian, Italian and Vietnamese cuisines you have shared with me during our lunches. **Linde**, thank you for all the cute drawings you have made for me. Your drawings always put a smile on my face. **Rik**, thank you for the most amazing cookie recipe ever. I will never forget our blue tosti extravaganza during iGEM. Odd, but delicious. Thanks for always being so helpful and giving some tips on learning Python. **Rob** and **Maarten**, thank you for all the entertaining banter between the two of you. I would always say Rob is the fitter one :p. **Emma**, I love the humor that you bring on the lunch table. To the rest of SSB, **Stamati, Nong, Michael, Maria, Niels, Ruben, Dorett, Benoit, Wasim, Jesse, Marta, Rita, Anna, Jasper, Peter, Tom. S, Lyon, Maria, Enrique** and **Luis**, thank you for all the good times and nice conversations during each of our encounter!

To the Delft CRISPR/phage group, thank you for taking me in as one of your own and included me in your meetings. **Franklin**, thank you my dear friend and mentor for teaching me so much about phage biology and also about the world of science

in general. I was so much fun having you around from the moment I started. **Becca**, thank you for teaching me so much on adaptation! I have learned so much in my first year from you. You always treat everyone with open arms and are always inclusive. Thank you for being a great colleague while in Wageningen and thank you for all the nice catching up times during the CRISPR-meetings or in Delft. **Jochem**, thanks again to organizing basically the whole Israel trip and teaching me how to properly say Rehovot. I had such a blast! **Sebastian**, you started a couple of months before me, so thanks for all the advice and tips of things I had to do in the beginning of my PhD. Also, thanks you for all the nice conversations in the lab. **Patrick**, the master of pub quiz, it is insane how much fun facts you know. Thank you for always randomly telling me fun facts throughout the day.

Apart from the people I interacted with during work, there are many people that I would like to thank that have brought me so much joy outside of work. **Nico V, Jeroen, Rob, Ismael, Max, Daan, Peter, Thijs**, thanks for all the fun DnD adventures together. By day we were scientist, by night, we were either a creepy dwarf, a half human half warlock, a half-naked monk, a bard, a tiny cleric, a skilled archer and a wizard that was obsessed with creating owl owl-bears. What started as a couple of after-work sessions in the meetings rooms, later on escalated into many different campaigns, which also included **Jarret** and **Thomas** as well. From exploring dungeons, to creepy (play) houses, to pirates ships, we let our imaginations go wild. May we always keep our imagination going and conquer new adventures together.

Thank you to the Ice-scream group, **Ruben, Dorett, Joyshree, Prarthana, Teunke, Mark, Alex, Nico, Emmy** and **Carrie** thank you for the fun games, pub quiz and conversation. Of course accompanied by ice scream and other enjoyable food items.

Around the time I started my PhD, I also fell in love with a sport that I enjoy doing and still continue to do, a.k.a. CrossFit. Thank you **Ricardo** and **Manos** for being my first CrossFit buddies and mentors in the sport. **Claudia, Judith, Erik, Costas, Kutay, Paul, Ioannis K., Paula, Jolanda, Maria** and **Annette**, thanks for making CrossFit a fun place to be even when we have to do gruesome workouts. Also, **Milène, Andrea** and **Whitney** for the fun Thursday night performance training with **Sven**. **Milène**, for the home gym training during lockdown. **Sven, Jip, Rutger, Iris, Luuk and Willem Jan**, thank you for all the coaching at CrossFit Ede, the PhD trajectory made me grow mentally, but you guys made me grow physically.

Apart from CrossFit, I also enjoy climbing a lot, so to all my climbing/boulder buddies: **Margo, Ruben, Dorett, Michael, David, Thijs, Koen, Thomas, Sanne, Livia, Adi, Tim, Alyana, and Costas**. Thanks for all the fun bouldering sessions both indoor and outdoor!

Thanks **Claudia, Costas** and **Kutay** for the fun dinner parties and the best banana bread ever. **Thomas** and **Sanne** the nice board games throughout the years, always a joy playing Battlestar Galactica with the whole group, together with **Ruben, Margo**,

Dorett and **Paul**. **Erik-Jan**, **Irene** en **Paul**, bedankt voor alle leuke avonturen op woensdag avonden. Maak niet uit als we als pandemie specialisten, time-agents of de worst-clan speelde, het is altijd een leuke avond vol gezelligheid! **Maritza**, **Lisette** en **Irene**, bedankt voor alle leuke, gekke, luide, lullige en lollige (dank EJ) tijden! Op naar meer gekkigheid and slechte series.

Michael, thank you for being that sarcastic friend that I never knew I needed. It has been great having you around during the starting phases of my PhD. **Ruben**, **Dorett**, you and I at one point basically did everything together all the time! I am happy that we all still get to see each other in our annual let's visit Michael weekends. A weekend full with fun activities and ESCAPE (the boardgame). Thank you for teaching me the art of sarcasm, it is an art that I am only now slowly comprehending. I wish you a wonderful future together with **Sarah** and **Samu**.

Naar mijn jaarclub (BonBom) en vriendinnen van meer dan 10 jaar. **Anke**, **Selina**, **Roos**, **Daniëlle**, **Sabine**, **Maritza**, **Lisette** en **Irene**. Bedankt voor alle leuke bijeenkomsten, zoals onze nieuw jaar borrel, alle verjaardagen en (dictator)weekendjes weg samen. Dagen vol met gezelligheid, uren lang praten, koken, thee drinken, en lekker veel eten. We zijn er altijd voor elkaar, maak niet uit wanneer en waar. Toch als we elkaar pesten in spelletjes, blijven wij altijd vriendinnen! Bedankt voor alle steun en liefde.

Tian, danki bo tur e chilletjes i hang outs na Aruba. Danki pa tur bo ayudo cu mi ouders ora mi no tawata tei. Bo ta un bon brother **Tian**, i mi ta desaya abo i **Laura** un bon futuro hunto. **Manfen**, danki pa tur e leukjes comiendo hot pot i hugando mahjong na Tilburg cu **Quiting**. **Seiten** i **Quiting**, mi dushinan! No ta tur dia den e aña nos ta den mesun pais, pero ora nos ta hunto, nos lo ta chilling i bibando e bida dushi. Endless Starbucks na Marriott, papiando te madrugada laat over di tur cos den bida. Nos luxurious brunch na Hyatt bebendo mimosa cu un poquito champagna. Nos fin dia aña, first sunrise of the year a bira gewoon tradition. Toch si nos tur ta mal drumi, boso ta lanta hunto cumi, i dreanta den auto pa wak e solo subi na San Nicolas! **Quiting** i **Seiten**, mi ta desea boso dos un bon Futuro. **Quiting**, un bon futuro unda bo tin tur cos bo ta desea den bida. **Seiten**, un futuro bunita hunto cu bo famia, **Wuyun** i **Hailey**.

Dorett, **Ruben** and **Margo**, thank you for being my great friends and my dearest sunshines. Apart from being my friends, you are also people that I greatly admire and constantly look up to. Some of your principles in both life and science are ones that I also follow. I have enjoyed all the fun times together (sometime also with **Paul**). Memories such as playing all the boardgames!, building a spaceship and hiking. Oh let us not forget overcooked, the game where yelling at friends is ok. Most importantly, thanks for always being there, for the good times and the bad times. **Margo**, thank you for introducing me to the world of podcast and showing me that you can keep plants alive and well. I can't wait to for you to make me your delicious pancakes again. **Ruben**, thank you for introducing me to two of my greatest hobbies, boardgames and climbing. In addition, you also always keep my sarcasm detection meter in check and know when I need to practice more. Thank you for all the great vacation togethers

with **Dorett, Paul** and **I**. Our group can always have a nice relaxing vacation with zero fuss, it's amazing! **Dorett** *hugs*, never would have I known then when I asked you for the coffee card that we would end up with such a beautiful friendship. A friendship that understands and accepts each other no matter what. I always love that we share similar food obsessions, such as oatmeal, Korean food, sushi, noodles and broccoli to name a few. You are the kindest, sweetest and most sincere human that I know of and I am glad to have you in my life. May we continue to eat our spaghetti drenched in soy sauce together!

Michèle and **Anton**, thank you for taking me into your family as one of your own. I am happy to be part of the family. I always enjoy our conversation and discussions whenever we come over for dinner. To celebrate Sinterklaas with the whole family is something I like forward to the most every year. Thank you for all the help with almost anything I ever needed, that be either house renovations, gardening tips or a good history fun fact, you guys are always there. Lastly, thank you for bringing me to your cycling vacations, I have seen parts of the world that I would never imagine I would see. Wishing you an amazing future to come. **Lucas**, it is always great having you around for dinners, it is always a conversation full of excitements.

Vincen, 哥哥. Thank you for being most coolest brother in the world. I love your constant curiosity to learn new things and has allowed you to accumulate a vast amount of knowledge on various topics. You are always there to make sure I go on the right path, giving me advice on what is the best to do based on your in depth research on things. Also, thank you for always bringing me along for surfing or snowboarding and constantly letting me to try new things. I can't wait to hang out with you again and fun things like surfing, snowboarding, cycling (only in London), cooking or playing games together with you and **Yvonne**.

Mom and dad, thank you for all that you have done through your life for Vincen and I. You worked from the ground up, with long hours doing labor intensive jobs with little to no rest. All this for the sake of a better life for the family. While other children had to work for at the shop after school, you instead motivated us to pursue a path of academics. I am forever gratefully to the both of you. Thank you for always believing in me and always push me to do better and greater things. You are the reason why this whole PhD was ever even an option in my life. I love you Mom and Dad.

Chinese translation: 亲爱的爸爸妈妈，感谢您们为我和哥哥奉献了自己的一生。您们起早贪黑地工作，长时间进行密集式的劳动工作，几乎没有休息过，所有的付出只为了让家人过上更好的生活。别人家的父母会让自己的孩子放学后，到商店工作以帮补家计，而您们却激励我们走上了一条学术之路。我一直对你们俩深怀感谢，感谢您一直以来对我的信任，并一直推动我去做更好，更大的事业。您们的鼓励就是我选择博士学位作为我人生方向的原因。我爱你，爸爸妈妈。Thanks to my amazing cousin **Siuyu**, for translating this message for me.

Paul, you have been my biggest support system during these past few years, and that is why you are getting your full page after all. Thank you for constantly supporting me in everything that I do, rather it be my job, my sports or my new hobbies, which constantly changes with the season. Thank you for always believing in me and constantly tell me I can do anything I want, as long as a I put my mind into it. Thank you joining me on this rollercoaster ride we call life, that have mostly ups but also some downs. You are everything that I am not, you are calm and quiet, I am chaotic and loud, you can reach the top shelf and I the bottom shelf and together I think we make the perfect team. I look forward our future ahead together, because with you by my side, I know that everything is going to be fine. Love you muchos **Paul**.

About the cover

On the cover are DNA nucleotides/letters consisting of A, G, T or C. Highlighted is the title of the thesis to show that using CRISPR-Cas proteins, you can find a specific DNA sequence in a vast amount of nucleotides. You can find what you need to find, only if you use the right tool for the right job. The DNA sequence on contains the Cas12a CRISPR repeat (GTCTAA GAACCTTTAAATAATTTCTACTGTTGTAGAT) and the mmuCas12u1 CRISPR repeat (GTGTCATAGCCCAGCTTGGCGGGCGAAGGCCAAGAC) that were studied in this thesis. Can you find them all?

Moreover, each chapter starts with a DNA Sequence, of which within the chapter is indicated. On the 3rd line of that sequence, you can find a piece of sequence that is related to the chapter. Chapter 1, the E. coli CRISPR repeat used in our lab to show crRNA processing by Cascade. Chapter 2, a piece of the sgRNA from Cas9. Chapter 3, the N-terminal sequence of Cas4 (V-A) that was missing but now restored. Chapter 4, Cas12a/Cpf1 CRISPR repeat. Chapter 5, spacer 1 and 3 used to prove the concept of “cut & paste”. Chapter 6, MmuCas12u1 CRISPR repeat. Chapter 7, C-tile spacers used to test our Mmu base editor. Chapter 8, A motif spacers used to test our Mmu adenosine’s base editors.

The research described in this thesis was financially supported by
The Dutch Research Council (NWO)
TOP grant (714.015.001) to Prof. Dr John van der Oost.

Financial support from Wageningen University
for printing this thesis is gratefully acknowledged.

Cover and layout design by Univorm
Printed by GVO drukkers & vormgevers B.V.

1. *What is the main purpose of this document?*

2. *What are the key findings or conclusions?*

3. *What are the main challenges or obstacles?*

4. *What are the next steps or recommendations?*

5. *What are the main stakeholders or participants?*

6. *What are the main risks or uncertainties?*

7. *What are the main sources of information or data?*

8. *What are the main assumptions or premises?*

9. *What are the main limitations or constraints?*

10. *What are the main conclusions or findings?*

11. *What are the main recommendations or suggestions?*

12. *What are the main conclusions or findings?*

13. *What are the main recommendations or suggestions?*

14. *What are the main conclusions or findings?*

15. *What are the main recommendations or suggestions?*

16. *What are the main conclusions or findings?*

17. *What are the main recommendations or suggestions?*

18. *What are the main conclusions or findings?*

19. *What are the main recommendations or suggestions?*

20. *What are the main conclusions or findings?*

G T C T A A G A A C T T T A
T T G T A G A T G A G A A G
C A C T G T T A A A A G T C
T A A A T A A T T T C T A C
C T A T T C C T G T G C C T
C A G T G T C A T A G C C C
A G G C C A A G A C G T C G
C G C G G T A T G G C A G T
C C A G C T T G G C G G G C
C T A A G A A C T T T A A A
G T A G A T G A G A A G T C
G G C C A C T G T T A A A A
A A T A A T T T C T A C T G
A T T C C T G T G C C T T C
T G T C A T A G C C C A G C