Directed Evolution of CRISPR/Cas Systems for Precise Gene Editing

Liu, Rongming; Liang, Liya; Freed, Emily F.; Gill, Ryan T.

Published in:
Trends in Biotechnology

Link to article, DOI:
10.1016/j.tibtech.2020.07.005

Publication date:
2021

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Directed Evolution of CRISPR–Cas systems for precise gene editing

Rongming Liu¹#,*; Liya Liang¹#; Emily F. Freed¹; Ryan T. Gill¹,²,*

¹Renewable and Sustainable Energy Institute (RASEI), University of Colorado Boulder, Boulder, CO, USA, ²NNF-Center for Biosustainability, Danish Technical University, Lyngby, Denmark

*Correspondence:
Rongming.liu@colorado.edu (R. Liu)
rtg@biosustain.dtu.dk (R. T. Gill).

#The authors contributed equally to this work.

Link to the lab: https://www.colorado.edu/rasei/ryan-gill

Abstract

CRISPR technology is a universal tool for genome engineering that has revolutionized biotechnology. Recently identified unique CRISPR-Cas systems, as well as re-engineered Cas proteins, have rapidly expanded the functions and applications of CRISPR-Cas systems. The structures of Cas proteins are complex, containing multiple functional domains. These protein domains are evolutionarily conserved polypeptide units that generally show independent structural or functional properties. In this review, we propose using protein domains as a new way to classify protein engineering strategies for these proteins, and discuss common ways to engineer key domains to modify the functions of CRISPR-Cas systems.

Keywords: CRISPR; Cas9; Cas12a; Cas protein engineering; gene editing
**Highlights**

Engineering of CRISPR nucleases, such as Cas9 and Cas12a, has proven to be a good strategy for improving CRISPR technology as a genome engineering toolset.

Introducing exogenous protein domains can generate new functions in CRISPR-Cas systems, including highly specific genome editing, base editing, gene expression regulation, and genome imaging.

Deletion of certain protein domains can generate smaller CRISPR nucleases with reduced activity.

Engineering specific protein domains in Cas9 or Cas12a can result in altered PAM specificity and on-target specificity for improved CRISPR-Cas performance.

**CRISPR–Cas Systems for Gene Editing and Gene Regulation**

The development of precise, efficient, and versatile technologies for customized manipulation of genetic information is an important objective for gene therapy, biological engineering, and synthetic biology. However, developing modular DNA-binding proteins that bind at specific targets often requires expertise in protein engineering. In the past decade, the CRISPR-Cas system has emerged as a powerful technology for genome editing due to its simplicity of design and high efficiency, and is now widely used in the fields of gene editing and base editing [1–4], gene regulation [5–7], epigenetic engineering [8,9], and imaging [10,11], among others.

The most commonly used CRISPR nucleases are Cas9 and Cas12a (BOX 1). Both of these nucleases are part of the class 2 family of Cas proteins, which means that a single Cas protein is responsible for target DNA recognition, binding, and cleavage. Protein engineering of CRISPR nucleases is commonly used to improve CRISPR as a tool for genetic modification and manipulation. For example, nuclease-dead Cas molecules, also known as dCas9 and dCas12a, are Cas9 or Cas12a mutants whose
nuclease activity is eliminated by point mutations in the endonuclease domain [12]. dCas9 and dCas12a can be used for transcriptional inhibition (CRISPRi), and by fusing functional domains to dCas9, it can also be used for transcriptional activation (CRISPRa) [5–7]. Moreover, Cas protein variants created by rational design or directed evolution can enable high-fidelity gene targeting [4,13].

In the last few years, much effort has been spent on gaining a deeper understanding of CRISPR-Cas systems. Many new CRISPR-Cas systems have been discovered in bacterial and archeal species, and the structures and functions of the Cas proteins in these systems are being elucidated [14–17]. Protein engineering strategies, particularly with the more commonly used Cas9 and Cas12a proteins, have given further insight into the mechanisms of action of these proteins [4,18–20], but CRISPR-Cas systems still are not fully understood. Therefore, we present recent developments in protein engineering of Cas proteins from the perspective of a new classification system linking specific functions to specific protein domains (Figure 1, Key Figure). We will discuss introducing exogenous protein domains for new functions; deleting protein domains for the development of smaller nucleases; and engineering specific domains for high-precision editing and/or low off-target effects. By understanding which exogenous and/or endogenous protein domains are responsible for specific functions of CRISPR-Cas systems, it will be possible to specifically engineer Cas proteins for precise clinical and industrial applications.

Introducing Exogenous Protein Domains for New Functions

CRISPR-/Cas9 technology uses a short guide RNA (gRNA) with a defined 20-27
bp sequence complementary to the DNA sequence to be targeted. The gRNA is used as an efficient targeting tool that results in the binding of a Cas protein to the desired genomic locus for generating site-specific genome alterations. To take advantage of the high DNA-binding capability of CRISPR-Cas systems for new functions, Cas proteins can be dimerized by end-to-end fusion of a Cas protein with active domains. A protein domain is a conserved part of a protein sequence that can also exist and function independently. Many proteins consist of multiple domains, and some domains may appear in a group of proteins which have similar functions or structures. Introducing an exogenous protein domain is a popular strategy to generate new protein architectures. This approach has resulted in significant improvement to functional diversity and efficiency of CRISPR-based systems [4,7,21–23].

Some examples of this strategy include domain fusions of dCas9/dCas12a protein with non-Cas nucleases (such as FokI) for highly specific genome editing [21,24], cytidine deaminase or adenosine deaminase for base editing [2,25–27], transcriptional repressor/activation domains for gene expression regulation [28,29], histone modification enzymes (such as methyltransferases) for epigenetic regulation [23,30], and fluorescent proteins for genome imaging [10,31]. In addition to dCas9/dCas12a, Cas12k, a naturally nuclease–deficient Cas protein, has also been used for targeting the desired locus [17]. Strecker and coworkers characterized a CRISPR-associated transposase from the cyanobacteria, Scytonema hofmanni, that contains Tn7-like transposase subunits and Cas12k. This system, called ShCAST, can insert DNA segments 60 to 66 bp downstream of the protospacer. In all of these applications, the Cas protein is only used for locating the targeted position, while the main function is achieved by the fused domain(s).
However, there are also limitations in these methods, generally deriving from two issues: the large size of the dimerized protein limits the target scope, and the fused domain may not perform efficiently or may have other limitations. For example, a FokI-dCas9 fusion increased the specificity of genome cutting in human cells compared to wild-type Cas9 [21,24], but this method restricts the target scope due to the limited space (~15 to 25 bp) of binding sites between two FokI-dCas9 monomers [21,24]. As an example of the second issue, fused base editing domains can readily mutate a G·C base pair to an A·T base pair without an editing template, but off-target effects of base editors have recently been reported [32–34].

In the above examples, a nuclease-dead Cas protein is used to target the desired locus in the genome; however, the Cas9 nickase (Cas9n) can also be used to target the genome and enable new functions. Cas9n has the D10A or H840A mutation which results in a single-strand rather than double-strand break [35]. Recently, Halperin and coworkers developed a targeted mutagenesis tool named EvolvR, consisting of Cas9n fused to a DNA polymerase that performs error-prone nick translation to generate mutations. The range of targeted mutation rates are up to 7,770,000-fold greater than rates seen in wild-type cells, and editing windows are up to 350 nucleotides. Cas9n has also been used to develop a prime editor tool, which can edit bases in mammalian cells without donor DNA or double-strand breaks [36]. In this system, Cas9n is fused to an engineered reverse transcriptase, programmed with a prime editing guide RNA (pegRNA) that both targets the specific genomic locus and encodes the desired edit. Recently, this method was also adapted for plants, such as rice and wheat [37].
Fusing other protein domains to a Cas protein can introduce new genome alteration capabilities. Fusing protein domains can also be used to control the expression of the Cas protein, which can limit off-target activity by spatially and/or temporally limiting where and/or when the CRISPR-Cas system is active. To this end, efforts have focused on the fine control of CRISPR-Cas systems by inserting or fusing active domains to Cas proteins, including inducible Cas proteins (or split Cas proteins) by intervening protein (intein) insertion [38,39], photoregulation of Cas proteins [40,41], and light-induced or chemically induced CRISPR-Cas systems [42,43]. Furthermore, these systems can be combined for synergetic applications. Nihongaki and coworkers engineered a split Cas12a enzyme that consists of N-terminal and C-terminal fragments, and also identified split Cas12a pairs that can provide chemical- and light-inducible Cas12a activation by using chemically-inducible dimerization domains and light-inducible dimerization domains, respectively [38]. Similarly, Zhou and coworkers engineered “photoswitchable” Cas9 (ps-Cas9) proteins. In this system, a dimeric fluorescent protein is fused to Cas9 (or dCas9), blocking Cas9 access to the target DNA sequence and therefore turning off Cas9. However, in the presence of cyan light, the fluorescent protein dissociates, resulting in an active Cas9 protein [40]. These strategies regulate the activity of CRISPR-Cas systems by the recognition of signals by a sensor domain, so they can reduce the off-target activity and allow for fine control of the activity of Cas protein. However, background activity caused by incomplete alteration of the Cas protein structure and weakened activity after induction due to insufficient restoration are areas to further improve conditionally controlled CRISPR-Cas systems [43–45].
Deleting Protein Domains to Develop Smaller CRISPR Nucleases

Cas9 is the most commonly used and best studied of CRISPR nucleases. Generally, Cas9 endonucleases contain a recognition (REC) lobe and a nuclease (NUC) lobe, which can be further separated into subdomains [46] (BOX 1). However, the size (~4.3 kb) of Cas9 can be limiting, particularly in gene therapy applications involving delivery through rAAVs (recombinant adeno-associated viruses) [47]. To identify the essential domains in Cas9, various studies deleted REC subdomains individually. SpCas9ΔREC2 retained ~50% activity compared to native SpCas9, while the ΔREC3 mutant had a 1000-fold reduction in activity and the ΔREC1 mutant lost all activity [48–50].

Cas12a is distantly related to Cas9 evolutionarily and structurally. However, Cas12a domains are named based on Cas9 counterparts [51], so Cas12a endonucleases contain a REC lobe and a NUC lobe, which can also be divided into subdomains (BOX 1). A recently constructed smaller dCas12a, dCas12a(ΔNuc), has a much lower target binding affinity than that of wild-type dCas12a [52]. Thus, while deleting domains can reduce the size of CRISPR nucleases, this approach also leads to reduced function, in most cases to an extent that this does not represent a suitable solution for engineering of smaller-size CRISPR nucleases.

Protein Engineering of Cas Protein Domains for Altered PAM Specificity

Although CRISPR-Cas systems are widely used for genome editing and other genome modification, the scope of sequences that Cas proteins can recognize is
dependent on PAM specificity [53,54]. The SpCas9 nuclease recognizes the sequence “NGG” as its PAM. An NGG occurs on average in 1 in 16 random genomic sequences, which limits the targeting scope of SpCas9, for example with base editing that requires precise Cas protein positioning [2,25–27]. When a Cas protein binds to the target DNA, the PAM-containing DNA duplex is positioned between the WED and PI domains in the NUC lobe, where the PAM nucleotides are recognized by a specific combination of amino-acid residues in the PI domain [18]. Thus, the WED and PI domains are the main regions that are targeted for engineering PAM specificity (Table 1) [55–57].

Sequence databases contain over 1000 Cas9 orthologs with differing PAM specificities. The PI domain, due to its direct interaction with the PAM, is the primary region for engineering PAM specificity. Generally, protein domains can exist and function independently, so the PI domain or its key region can be replaced to expand PAM compatibility [56,57]. For example, Ma and coworkers swapped the conserved key region in the PI domain of SaCas9-KKH with the corresponding region of SaCas9 orthologs. They generated a group of chimeric Cas9 variants, and identified several variants exhibiting broadened PAM recognition [56]. These SaCas9 variants recognize up to 1/4 of all possible PAM sequences (compared to 1/64 with SaCas9) in mammalian cells, allowing for expanded targeting.

Direct evolution approaches have also been applied for engineering PAM specificity (Table 1) [15,55,58–61]. The prototypical strategy has employed random mutagenesis to generate a Cas protein library, then applied a bacterial selection/screen to identify positive mutations, followed by construction of combinatorial mutations for further engineering of PAM recognition [15,55,58,59,61]. Using this method, Kleinstiver
and coworkers found a series of triple or quadruple mutations in the PI domain of SpCas9 and SaCas9 that had different PAM specificities [55,59,61]. Similarly, Gao and coworkers used a random mutagenesis approach to alter all 60 amino acids in AsCas12a that are in proximity to the PAM, and found two variants, each with combinatorial mutations in the WED and PI domains, that led to recognition of different PAM sites [15]. When the corresponding mutations were made in LbCas12a, similar changes in PAM specificity were observed [15]. In yet another demonstration, Hu and coworkers used phage-assisted continuous evolution (PACE) to generate a group of Cas9 variants (xCas9s) that recognized an expanded range of PAM sequences and also had reduced off-target activity, as discussed below. The majority of these variants had mutations in the PI, REC1, and REC3 domains [58]. Miller and coworkers further used phage-assisted evolution to identify three new SpCas9 variants capable of recognizing non-G PAM sequences. These variants had mutations located in the WED and PI domains [62]. The above approaches resulted in a 1.4- to 4-fold increase in the number of sites that can be targeted in the human or Arabidopsis genomes. In addition, some engineered Cas proteins with expanded PAM recognition can be used to edit disease causing mutations that can not be edited by the non-engineered Cas protein [62,63].

As a complementary approach to random mutagenesis based methods, structure-guide engineering methods have also been used to expand PAM specificity (Table 1) [60,64,65]. For example, it has been shown that mutations in residues that directly interact with the third G in the NGG PAM (e.g. the R1556 residue of FnCas9 and R1335 of SpCas9, both of which are in the PI domain) can reduce PAM constraints [60,64]. Other targeted mutations in the WED and PI domains have also led to Cas9 variants with more
relaxed PAM recognition. For example, the RHA FnCas9 variant and SpCas9-NG variant can recognize YG and NG PAM sites, respectively. Overall, mutations in the WED and PI domains are the primary way to change PAM compatibility in different Cas proteins [15,55,58–61,64]. However, further investigations into the mechanisms that control PAM specificity are required to develop a model for forward engineering of PAM recognition in order to engineer nucleases that can precisely target any desired genome sequence.

**Protein Engineering of REC Domains for On-target Specificity**

A main limitation of CRISPR-Cas systems is unwanted and unpredictable off-target activity. Therefore, much effort has been focused on directly engineering Cas proteins for improved on-target specificity (Table 2). REC domains, in conjunction with WED domains, are responsible for gRNA recognition [66,67]. Furthermore, crystal structures have revealed multiple amino acids within the REC domains of both Cas9 and Cas12a proteins that make non-specific contact with the target DNA [50,66,67]. Therefore, many of the strategies to decrease off-target DNA-cleavage while preserving on-target activity have focused on REC domains [13,68,69].

Multiple Cas9 and Cas12a nucleases have been engineered that have reduced off-target activity while maintaining wild-type levels or near wild-type levels of on-target activity (Table 2). These high specificity nucleases were engineered by one of two approaches: structure-guided mutagenesis [4,19,68,69] or screening of libraries created by random mutagenesis [13,70,71]. The first approach begins with examining the crystal structures of SpCas9, SaCas9, or AsCas12a to identify amino acids that directly interact with, or that are in close proximity to, the target DNA. These amino acids, which are located primarily in the REC domains, are then mutated singly and combinatorially to
identify variants with high on-target activity (>70% of the wild-type activity) and reduced off-target activity [4,19,68,69]. For the latter approach, libraries of randomly generated mutations, either in a single domain (REC3) or across the entire protein, were screened in yeast [13] or E. coli [70,71]. In these systems, dual reporters are used to simultaneously screen/select/counter-select for both on- and off-target activity. All of these approaches have resulted in nucleases that have no off-target activity with the vast majority of target sites (gRNAs) that are tested, allowing these nucleases to be used for precision genome editing.

**Protein Engineering of the nt-groove for On-target Specificity**

The majority of engineered nucleases with enhanced specificity have mutations in their REC domains. However, mutations in other domains can also lead to decreased off-target activity (Table 2). The crystal structure of SpCas9 shows that there is a positively charged groove (nt-groove) in the NUC lobe that interacts with the non-target strand of the target DNA [20]. The non-target strand needs to be unwound from the target strand and then displaced prior to DNA cleavage; therefore, mutations in the nt-groove could affect unwinding of the target DNA resulting in changes in on- and/or off-target activity [20,48,72].

Slaymaker and coworkers [20] tested this hypothesis by mutating 31 amino acids in the SpCas9 nt-groove to alanine and testing their activity. They then made combinatorial mutants with the top single mutants. Their best variant, named eSpCas9(1.1), has the same on-target activity as wild-type SpCas9 for the vast majority of gRNAs that were tested and shows greatly reduced genome-wide off-target activity for the two tested
gRNAs. Similarly, Gao and coworkers made a number of mutations in AsCas12a in residues that were predicted to interact with either the target or non-target strand of the target DNA [15]. They identified a few mutations that reduced off-target activity while still maintaining on-target activity. Their top mutation is predicted to contact the non-target DNA strand, similar to the mutations in eSpCas9(1.1).

For both categories of SpCas9s with increased specificity (engineered REC domain or nt-groove), several mechanisms have been proposed to explain the increased specificity. In order for SpCas9 to cleave the target DNA, Cas9 must recognize, bind, and unwind the target DNA. Studies with two Cas9s that have increased on-target specificity, SpCas9-HF1 and eSpCas9(1.1), have shown that these Cas9s bind to the target with the same frequency and affinity as wild-type SpCas9 regardless of whether the target sequence is a match or contains any mismatches [69,72]. However, the engineered Cas9s do not bind as stably to the target in the presence of a mismatch [72]. This less stable binding reduces the ability of these engineered Cas9s to unwind the target DNA [72]. These engineered Cas9s also block a conformational change when bound to mismatched targets, locking the enzymes in an inactive state [69]. Overall, cleavage rates for matched targets are not affected with these engineered Cas9s, but cleavage rates for mismatched targets are significantly lower [69,72]. Similarly, a crystal structure of another high-specificity Cas9, xCas9-3.7, shows large conformational changes in the REC2 and REC3 domains that eliminate contacts between the engineered Cas9 and the target DNA [58], which could affect the stability of xCas9-3.7 binding to the target DNA, target DNA unwinding or the conformational change that is required for DNA cleavage. To fully understand the mechanisms of Cas nuclease activity and specificity, interactions between
the Cas protein, the gRNA, and both the target and the non-target strand of the target DNA must be further investigated.

**Concluding Remarks and Future Perspectives**

The demand for precision genome editing, particularly in the gene therapy field, requires efficient and accurate CRISPR-Cas editing systems. Here, we described protein engineering strategies based on engineered protein domains in class 2 Cas proteins. In recent years, several advances have been made, including introducing exogenous protein domains to expand Cas protein functions, deleting a protein domain to generate a smaller CRISPR nuclease, and engineering specific domains for altered PAM specificities and/or decreased off-target activity. These developments show that modifications in certain Cas protein domains can result in specific characteristic alterations in activity or specificity. For example, mutations in the WED and PI domains can change the PAM compatibility, and mutations mainly in the REC domain can reduce off-target activity. Thus, Cas proteins can be engineered so that the new variants have different capabilities and also confer high specificity and efficiency. Furthermore, since these enhanced functions can be introduced in a modular fashion by engineering the appropriate protein domain, combinations of new functions can be engineered in a single nuclease. Many of these engineered nucleases have been used for clinically, agriculturally, or industrially relevant applications (BOX 2).

In addition to protein engineering of Cas nucleases, gRNA engineering also can enhance the functionality of CRISPR-Cas systems. For instance, on-target efficiency has
been improved both by engineering of Cas proteins [68,69,71,73] and of gRNA [74,75]. However, these two strategies are generally applied separately for engineering of the CRISPR-Cas system. The combination of both approaches can generate a synergistic effect to expand Cas protein applications [76]. Similarly, editing efficiency and off-target activity can also be improved by optimizing the delivery of both the Cas protein and the gRNA into the target cells [70]. Given the rapid development of diverse engineered Cas proteins, it is likely that a wide range of tailored Cas proteins, combined with other strategies to improve efficiency, such as gRNA engineering and the development of novel delivery methods, will be developed for a wide variety of genome-editing applications.

CRISPR-Cas systems have made gene therapy a plausible and versatile treatment modality [77]. Recently, the first CRISPR-based therapy trial began in the United States (ClinicalTrials.gov Identifier: NCT02793856), and similar studies, such as those involving ex vivo [92] and in vivo therapies [93] associated with CRISPR technology have become increasingly popular. It is likely more clinical trials will appear in the future using CRISPR technology. The specificity, efficiency, and safety of CRISPR-based genome editing technologies will determine if these trials are successful, and Cas proteins will be further investigated and engineered as needed to meet the requirements for gene therapy.

Despite successes with CRISPR-Cas systems in a wide variety of applications, new Cas-based tools are required to overcome the features that are still poorly explored (see Outstanding Questions). For example, exogenous functional domains still have off-target activities [32–34], and some inducible CRISPR-Cas systems require induction methods that are expensive and can have relatively high toxicity [78,79]. Structure-guided engineering of Cas proteins has identified groups of mutations that improve CRISPR-Cas
systems, but the mechanisms controlling PAM recognition and off-target activity are still not fully understood, which limits further refinement of Cas proteins. Therefore, there is still further research required to understand the mechanisms of action of CRISPR-Cas systems and to further engineer Cas-based systems for expanded applications. For example, more detailed models of Cas proteins and their interaction with target DNA can be built to further understand how to engineer both greater on-target activity and reduced off-target activity. More detailed models may also allow for the rational design of Cas proteins to target a specific PAM sequence (e.g. in order to differentiate between a WT sequence and a disease-causing mutation). Machine learning using data from high-throughput screening of Cas variants would aid in creating improved models. In addition to engineering the Cas proteins themselves, exogenous functional domains, such as base editors, can also be engineered to achieve the exact desired function. By gaining a greater understanding of how CRISPR-Cas systems function, it should be possible to create designer CRISPR-Cas systems for any clinical or industrial application. Furthermore, many CRISPR-Cas systems have only been tested in mammalian cells, but their application could be expanded by introducing these tools to other biological systems such as bacteria, yeast, and/or plants.

Acknowledgments

The project was sponsored by the US Department of Energy (Grant DE-SC0018368).

References

4 Kleinstiver, B.P. et al. (2019) Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. Nat. Biotechnol. 37, 276–282
13 Casini, A. et al. (2018) A highly specific SpCas9 variant is identified by in vivo screening in yeast. Nat. Biotechnol. DOI: 10.1038/nbt.4066
22 Hong, Y. et al. (2018) Comparison and optimization of CRISPR/dCas9/gRNA genome-labeling systems for live cell imaging. Genome Biol. 19, 39


28 Wang, X. et al. (2019) Transcriptional repression of endogenous genes in BmE cells using CRISPRi system. *Insect Biochem. Mol. Biol.* 111, 103172


Cui, Y. et al. Direct observation of the formation of a CRISPR–Cas12a R-loop complex at the single-molecule level. Chemical Communications, 56. (2020), 2123–2126

Chatterjee, P. et al. (2018) Minimal PAM specificity of a highly similar SpCas9 ortholog. Sci Adv 4, eaau0766

Tóth, E. et al. (2018) Mb- and FnCpf1 nuclease are active in mammalian cells: activities and PAM preferences of four wild-type Cpf1 nucleases and of their altered PAM specificity variants. Nucleic Acids Res. 46, 10272–10285

Yamamoto, A. et al. (2019) Developing Heritable Mutations in Arabidopsis thaliana Using a Modified CRISPR/Cas9 Toolkit Comprising PAM-Altered Cas9 Variants and gRNAs. Plant Cell Physiol. 60, 2255–2262

Ma, D. et al. (2019) Engineer chimeric Cas9 to expand PAM recognition based on evolutionary information. Nat. Commun. 10, 560


Walton, R.T. et al. (2020) Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science DOI: 10.1126/science.aba8853


Bratović, M. *et al.* (2020) Bridge helix arginines play a critical role in Cas9 sensitivity to mismatches. *Nat. Chem. Biol.* DOI: 10.1038/s41589-020-0490-4


Walker, J.E. *et al.* (2020) Development of both type I-B and type II CRISPR/Cas genome editing systems in the cellulolytic bacterium Clostridium thermocellum.

Jiang, Y. et al. (2017) CRISPR-Cpf1 assisted genome editing of Corynebacterium glutamicum. *Nat. Commun.* 8, 15179


Zetsche, B. et al. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 163, 759–771


Figure 1. Schematic Representations of Engineering of Cas Protein Domains. We use SpCas9 as an example to show the strategies for engineering Cas protein domains. (A) End-to-end gene fusion and domain insertion approach for introducing exogenous protein domains. (B) Deletion of protein domains for the development of smaller CRISPR nucleases. (C) Protein engineering of WED and PI domains for altered PAM specificity. (D) Protein engineering of REC domains or domains involving the nt-groove for on-target specificity.
BOX1 Structures of CRISPR-Cas9/Cas12a RNA-guided nuclease.

The most commonly used CRISPR-associated (Cas) nucleases are Cas9 and Cas12a (formerly called Cpf1). Due to the ease of programming these nucleases to recognize a specific sequence of DNA, CRISPR-Cas systems have become widely utilized for a variety of applications [80]. However, CRISPR-Cas systems have limitations due to their ability to only recognize particular sequences of DNA and also due to off-target activity. To overcome these limitations, and to gain insight into the mechanisms of DNA recognition, binding, and cleavage, the crystal structures of Cas9 and Cas12a have been determined (Figure I). These structures have contributed greatly to protein engineering efforts for these nucleases.

Figure I. (A) Cas9 from *Streptococcus pyogenes* (SpCas9) consists of nuclease (NUC) and recognition (REC) lobes. The REC lobe can be separated into four regions, the bridge helix (BH), as well as REC1, REC2, and REC3 domains [46]. The NUC lobe can be divided into RuvC, HNH and PAM-interacting (PI) domains [46]. (B) The gRNA (red) contains a constant structural region, which interacts with the SpCas9 protein, and a variable targeting region, which identifies a complementary region on the target DNA molecule (yellow). The **Protospacer-adjacent motif (PAM)** (purple) is recognized by the SpCas9 protein. The rest of the target DNA is shown in grey. (C) SpCas9 edits genomes by cleaving target dsDNA under the guidance of gRNA (PDB: 5U43). (D) Cas12a from *Acidaminococcus* sp. (AsCas12a) consists of NUC and REC lobes. The REC lobe can be separated into two regions, REC1 and REC2 domains [51]. The NUC lobe can be divided into WED, PI, RuvC, BH, and Nuc domains [51]. (E) The gRNA (red) contains a constant structural region, which interacts with the AsCas12a protein, and a variable
targeting region, which identifies a complementary region on the target DNA molecule (yellow). The PAM (purple) is recognized by AsCas12a protein. The rest of the target DNA is shown in grey. (F) The AsCas12a edits genomes by cleaving target dsDNA under the guidance of gRNA (PDB: 5B43).
BOX 2 Applications of engineered Cas nucleases

Mammalian systems

Engineered Cas nucleases offer great potential for translating CRISPR-Cas editing to therapeutic applications due to their ability to make precise genome edits with reduced off-target activity. For example, Vakulskas and coworkers showed that HiFi Cas9 was able to introduce edits into six therapeutically relevant loci in human hematopoietic stem and progenitor cells (HSPCs) [70]. The editing efficiency with HiFi Cas9 was the same or higher than WT SpCas9 and the off-target effects were significantly reduced. Engineered Cas9s have also been used to prevent disease in animal models. György and coworkers used SaCas9 KKH to prevent dominant progressive hearing loss in mice [63]. The deafness causing mutation is a single point mutation in the Tmc1 gene that changes a GGA TGT in the WT allele to GGA AGT in the mutant allele. SaCas9 KKH recognizes a PAM (NNNRRT) in the mutant allele, but not in the WT allele, allowing for specific editing of only the mutant copy of the gene.

Plants

Precise genome editing with minimal off-target activity is also important in plants, particularly in crops. Studies in rice have shown that by optimizing the gRNA, in addition to the Cas protein, high editing efficiencies can be achieved with the Cas9-VQR variant [76] and high specificity can be achieved with the SpCas9-HF1 and eSpCas9 variants [81]. Higher editing efficiencies have also been observed in rice using the Cas9-NG variant [82]. Similarly, the both the Cas9-VQR nickase and dCas9-NG, when fused to a base editor, result in high editing efficiencies at single bases without the need for a homology donor template [82,83].
Bacteria

CRISPR-Cas systems are also highly effective for engineering non-eukaryotic systems. The genomes of many species of bacteria can be edited with very high efficiency using homologous recombination combined with Cas9 or Cas12a to select against non-edited cells [84–88]. Although genome editing has been optimized in many bacteria, other CRISPR-based tools still need to be expanded for bacterial use. Recently, nuclease-dead xCas9-3.7 was used to create a CRISPRa system in E. coli [89]. Use of dxCas9-3.7 allowed for targeting more promoters and also improved the dynamic range of activation when compared to WT Cas9. This system will allow for better engineering of industrially-relevant bacterial strains.
Outstanding Questions

To what extent does engineering specific domains in Cas proteins effect on-/off-target efficiency and PAM compatibility?

What molecular and structural changes occur due to Cas protein engineering? Will understanding these changes help us design Cas nucleases with exactly the properties we want?

How can we decrease the off-target activities of exogenous functional domains such as base editing domains?

How can Cas protein engineering be combined with other modifications, such as improved gRNA design, optimization of expression levels of both the gRNA and the Cas protein, and optimization of gRNA/Cas delivery method, to achieve efficient and precise genome editing?

To what extent can Cas protein engineering contribute to the effectiveness and success rate of the CRISPR technology in genome editing-based gene therapy?

Some CRISPR-Cas systems have only been tested in mammalian cells. Can these tools be introduced to other biological systems such as viruses, bacteria, yeast, and/or plants?
<table>
<thead>
<tr>
<th>Name</th>
<th>Engineered domain(s)</th>
<th>Mutation(s)</th>
<th>Biological or technological significance</th>
<th>PAM</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpCas9*</td>
<td></td>
<td></td>
<td></td>
<td>NGG</td>
<td>[90]</td>
</tr>
<tr>
<td>SpCas9 VQR</td>
<td>WED, PI</td>
<td>D1135V/R1335Q/T1337R</td>
<td>2-fold increase in targeting potential in the human genome vs. SpCas9 (WT); SpCas9 VQR has higher editing efficiency in rice vs. SpCas9 (WT)</td>
<td>NGAN</td>
<td>[16,68,76]</td>
</tr>
<tr>
<td>SpCas9 EQR</td>
<td>WED, PI</td>
<td>D1135E/R1335Q/T1337R</td>
<td></td>
<td>NGAG</td>
<td>[16,68]</td>
</tr>
<tr>
<td>SpCas9 VRER</td>
<td>WED, PI</td>
<td>D1135V/G1218R/R1335E/T1337R</td>
<td></td>
<td>NGCG</td>
<td>[16,68]</td>
</tr>
<tr>
<td>Sp-Sl3Cas9</td>
<td>PI</td>
<td>SpCas9 with the PI domain of Sl3Cas9</td>
<td>Demonstrates the PI domain determines the PAM (exchanging the PI domain also exchanges the PAM recognition)</td>
<td>NGGNG</td>
<td>[51]</td>
</tr>
<tr>
<td>St3-SpCas9</td>
<td>PI</td>
<td>Sl3Cas9 with the PI domain of SpCas9</td>
<td></td>
<td>NGG</td>
<td>[51]</td>
</tr>
<tr>
<td>xCas9-3.7</td>
<td>REC1, REC2, REC3, PI</td>
<td>A262T/R324L/S409I/E480K/E543D/M694I/E1219V**</td>
<td>Allows ~3-fold more pathogenic SNPs to be corrected by base editing</td>
<td>NG, NNG, GAA, GAT and CAA</td>
<td>[58]</td>
</tr>
<tr>
<td>SpCas9-NG</td>
<td>WED, PI</td>
<td>R1335V/L1111R/D1135V/G1218R/E1219F/A1322R/T1337R</td>
<td>~3-fold increase in targeting potential vs. SpCas9 (WT) in human coding sequences; allows editing of 3 pathogenic mutations that cannot be targeted by SpCas9 (WT)</td>
<td>NG</td>
<td>[64]</td>
</tr>
<tr>
<td>SpG</td>
<td>WED, PI</td>
<td>D1135L/S1136W/G1218K/E1219Q/R1335Q/T1337R</td>
<td></td>
<td>NGN</td>
<td>[65]</td>
</tr>
<tr>
<td>SaCas9*</td>
<td></td>
<td></td>
<td></td>
<td>NNGRRT#</td>
<td>[84]</td>
</tr>
<tr>
<td>Cas9 Variant</td>
<td>PI Type</td>
<td>Mutations</td>
<td>Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
<td>----------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SaCas9 KKH</td>
<td>PI</td>
<td>E782K/N968K/R1015H</td>
<td>2 to 4-fold increase in targeting potential vs. SaCas9 (WT); allows editing of a mutation that causes deafness, but not the WT allele</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v42-wt</td>
<td>PI</td>
<td>13-aa region swap from v42 (Sediminibacillus albus) to the wild-type SaCas9</td>
<td>NNVRRN# [56]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V17-L</td>
<td>PI</td>
<td>13-aa region swap from v42 (Sediminibacillus albus) to the wild-type SaCas9, I991L</td>
<td>NNVRRN [56]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v21-R-wt and v21-L-wt</td>
<td>PI</td>
<td>13-aa region swap from v21 (Staphylococcus water) to the wild-type SaCas9; v21-R (v21, I991R) and v21-L (v21, I991L)</td>
<td>NNGACT, NNGATG, NNGATT, NNGGCT, NNGGTT, and NNGGTT [56]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FnCas9*</td>
<td>WED, PI</td>
<td>E1369R/E1449H/R1556A</td>
<td>NGG [91]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FnCas9 RHA</td>
<td>WED, PI</td>
<td>S542R/K607R</td>
<td>YG### [60]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AsCas12a/LbCas12a*</td>
<td></td>
<td>S542R/K548V/N552R</td>
<td>TTTV [92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AsCas12a</td>
<td>WED, PI</td>
<td>~3-fold increase in targeting potential in human coding sequences vs. SaCas9 (WT)</td>
<td>TYCV [15,93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AsCas12a</td>
<td>WED, PI</td>
<td>S542R/K548V/N552R</td>
<td>TATV [15]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LbCas12a  WED, PI  G532R/K595R  AsCas12a/LbCas12a (WT); TYCV PAM allows editing of a liver disease mutation that cannot be targeted by AsCas12a/LbCas12a (WT)  TYCV  [15, 93]

LbCas12a  WED, PI  G532R/K538V/Y542R  TATV  [15]

*The original PAM is shown for the wild-type Cas9 and Cas12a from each organism.

**some of the mutations in xCas9-3.7 are responsible for the enhanced on-target efficiency.

# R=A/G

## V=G/A/C

### Y=C/T
Table 2. Nucleases engineered for greater on-target specificity

<table>
<thead>
<tr>
<th>Name</th>
<th>Engineered domain(s)</th>
<th>Mutation(s)</th>
<th>PAM</th>
<th>Fold increase in specificity* (range shown is for different target sites)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpCas9-HF1</td>
<td>REC3, RuvC3</td>
<td>N497A/R661A/Q6 95A/Q926A</td>
<td>NGG</td>
<td>0.73-19</td>
<td>[68,69]</td>
</tr>
<tr>
<td>HypaCas9</td>
<td>REC3</td>
<td>N692A/M694A/Q6 95A/H698A</td>
<td>NGG</td>
<td>1.83-19</td>
<td>[69]</td>
</tr>
<tr>
<td>evoCas9</td>
<td>REC3</td>
<td>M495V/Y515N/K5 26E/H698Q</td>
<td>NGG</td>
<td>0.34-9</td>
<td>[13,94]</td>
</tr>
<tr>
<td>HiFi Cas9</td>
<td>REC3</td>
<td>R691A</td>
<td>NGG</td>
<td>0.22-2.46</td>
<td>[70]</td>
</tr>
<tr>
<td>xCas9-3.7</td>
<td>REC1, REC2, REC3, PI</td>
<td>A262T/R324L/S4 09I/E480K/E543D/M694I/E1219V**</td>
<td>NG, NNG, GAA, GAT and CAA</td>
<td>0.20-0.29</td>
<td>[58,95]</td>
</tr>
<tr>
<td>Opti-SpCas9</td>
<td>REC3, RuvC3 (nt-groove)</td>
<td>R661A/K1003H</td>
<td>NGG</td>
<td>0.05-5</td>
<td>[94]</td>
</tr>
<tr>
<td>Sniper-Cas9</td>
<td>REC3, RuvC2, HNH</td>
<td>F539S/M763I/K890N</td>
<td>NGG</td>
<td>0.03-8.47</td>
<td>[71]</td>
</tr>
<tr>
<td>SaCas9-HF</td>
<td>REC, RuvC3</td>
<td>R245A/N413A/N419A/R654A</td>
<td>NNGRT</td>
<td>0.37-2</td>
<td>[19]</td>
</tr>
<tr>
<td>Enzyme</td>
<td>Domain/Variant</td>
<td>Recognition Site</td>
<td>Specificity (vs. enAsCas12a)</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------</td>
<td>------------------</td>
<td>------------------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>enAsCas12a-HF1</td>
<td>REC1</td>
<td>N282A***</td>
<td>0.13-1.89</td>
<td>[4]</td>
<td></td>
</tr>
<tr>
<td>M44</td>
<td>REC1</td>
<td>REC1 domain from TX_Cas12a</td>
<td>TTTN</td>
<td>0.01-24</td>
<td>[96]</td>
</tr>
<tr>
<td>eSpCas9(1.1)</td>
<td>HNH, RuvC3 (nt-groove)</td>
<td>K848A/K1003A/R1060A</td>
<td>NGG</td>
<td>0.83-19</td>
<td>[20,69]</td>
</tr>
<tr>
<td>AsCas12a-K949A</td>
<td>nt-groove</td>
<td>K949A</td>
<td>TTTV</td>
<td>0.23</td>
<td>[15]</td>
</tr>
</tbody>
</table>

*specificity is calculated by dividing the number of cuts at the on-target site by the total number of cuts (off-target and on-target), as determined by GUIDE-seq or other next-generation sequencing approach

**some of the mutations in xCas9-3.7, particularly E1219V, are responsible for the expanded PAM compatibility

***enAsCas12a also contains the E174R/S542R/K548R mutations, which result in expanded PAM compatibility
Glossary

**Adenosine deaminase**: an enzyme that catalyzes the irreversible deamination of adenosine and deoxyadenosine to inosine and deoxyinosine, respectively. Used in base editing.

**Base editing**: a genome editing approach that introduces a point mutation in specific sites of DNA or RNA without DNA cleavage and a repair template. A base modifying enzyme is fused to a dCas protein.

**CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa)**: a technique to inducing target-specific inactivation or activation of gene expression, respectively. These methods use a gRNA and a dCas nuclease fused to a transcriptional repressor or activator protein, respectively.

**Cytidine deaminase**: an enzyme that catalyzes the irreversible hydrolytic deamination of cytidine and deoxycytidine to uridine and deoxyuridine, respectively. Used in base editing.

**Directed evolution**: an iterative strategy used for protein engineering involving: i) diversification of the targeted gene (via mutagenesis); ii) identification of desired variants from selection or screening; and iii) template amplification for the next round experiments.

**FokI**: a type IIS restriction endonuclease found in *Flavobacterium okeanokoites*. It comprises an N-terminal DNA-binding domain and a non-specific DNA cleavage domain at the C-terminus.

**gRNA**: a short nucleotide sequence that is used to direct a Cas protein to a particular locus in the genome.

**Intein**: a protein sequence that can excise itself from a longer protein and then fuse the remaining portions with a peptide bond in a process called “protein splicing”.

**Off-target**: nonspecific and unintended genetic modifications using programmable nuclease technologies. Occurs at DNA regions with similar sequence to the target DNA sequence.

**Phage-assisted continuous evolution (PACE)**: a method that uses bacteriophage to continuously evolve nucleic acids or proteins toward desired functions *in vivo*. The host cell is engineered so that cells possessing the desired function will produce the phage infectivity protein, pIII. Cells that have more of the desired activity will produce more pIII, resulting in propagation of variants with the desired activity.

**Protospacer-adjacent motif (PAM)**: a short DNA sequence of 2–6 nt following the DNA sequence targeted by the CRISPR-Cas system.