HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2016 October 22.

Published in final edited form as:

Cell. 2015 October 22; 163(3): 759–771. doi:10.1016/j.cell.2015.09.038.

Cpf1 is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system

Bernd Zetsche[1,2,3,4,5,*], **Jonathan S. Gootenberg**[1,2,3,4,6,*], **Omar O. Abudayyeh**[1,2,3,4], **Ian M. Slaymaker**[1,2,3,4], **Kira S. Makarova**[7], **Patrick Essletzbichler**[1,2,3,4], **Sara Volz**[1,2,3,4], **Julia Joung**[1,2,3,4], **John van der Oost**[8], **Aviv Regev**[1,9], **Eugene V. Koonin**[7], and **Feng Zhang**[1,2,3,4,†]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142 [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139 [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 [5]Department of Developmental Pathology, Institute of Pathology, Bonn Medical School, Sigmund Freud Street 25, 53127 Bonn, Germany [6]Department of Systems Biology, Harvard Medical School, Boston, MA 02115 [7]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894 [8]Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, Netherlands [9]Department of Biology, Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

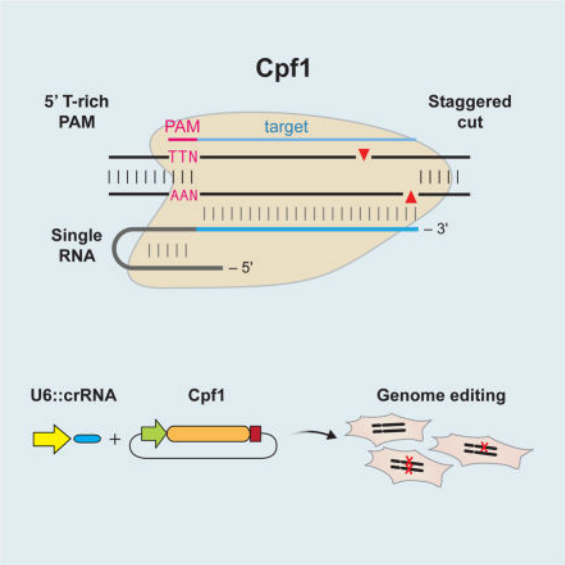
The microbial adaptive immune system CRISPR mediates defense against foreign genetic elements through two classes of RNA-guided nuclease effectors. Class 1 effectors utilize multi-protein complexes, whereas Class 2 effectors rely on single-component effector proteins such as the well-characterized Cas9. Here we report characterization of Cpf1, a putative Class 2 CRISPR effector. We demonstrate that Cpf1 mediates robust DNA interference with features distinct from Cas9. Cpf1 is a single RNA-guided endonuclease lacking tracrRNA, and it utilizes a T-rich protospacer adjacent motif. Moreover, Cpf1 cleaves DNA via a staggered DNA double stranded break. Out of 16 Cpf1-family proteins, we identified two candidate enzymes, from Acidominococcus and Lachnospiraceae, with efficient genome editing activity in human cells. Identifying this mechanism of interference broadens our understanding of CRISPR-Cas systems and advances their genome editing applications.

[†]To whom correspondence should be addressed: zhang@broadinstitute.org (F.Z.).
[*]These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

F.Z., B.Z., and J.S.G. conceived this study. B.Z., J.S.G., O.A., and F.Z. designed the experiments. B.Z., J.S.G., O.A., J.J., S.V., P.E., and F.Z. performed the experiments and analyzed the data. I.S. conducted FnCpf1 purification. A.R. assisted with RNA sequencing and analysis. K.S.M., E.V.K., and F.Z. performed the computational sequence analysis. F.Z., E.V.K., B.Z., J.S.G., O.A., K.S.M., and J.V.O wrote the manuscript that was read and approved by all authors.

Graphical Abstract



INTRODUCTION

Almost all archaea and many bacteria achieve adaptive immunity through a diverse set of CRISPR-Cas (Clustered Regularly-Interspaced Short Palindromic Repeats and CRISPR-ASsociated proteins) systems, each of which consists of a combination of Cas effector proteins and CRISPR RNAs (crRNAs) (Makarova et al., 2011; Makarova et al., 2015). The defense activity of the CRISPR-Cas systems includes three stages: (i) adaptation, when a complex of Cas proteins excises a segment of the target DNA (known as a protospacer) and inserts it into the CRISPR array (where this sequence becomes a spacer); (ii) expression and processing of the precursor CRISPR (pre-cr) RNA resulting in the formation of mature crRNAs; and (iii) interference, when the effector module – either another Cas protein complex or a single large protein – is guided by a crRNA to recognize and cleave target DNA (or in some cases, RNA) (Horvath and Barrangou, 2010; Sorek et al., 2013; Barrangou and Marraffini, 2014). The adaptation stage is mediated by the complex of the Cas1 and Cas2 proteins, which are shared by all known CRISPR-Cas systems, and sometimes involves additional Cas proteins. Diversity is observed at the level of processing of the pre-crRNA to mature crRNA guides, proceeding via either a Cas6-related ribonuclease or a housekeeping RNaseIII that specifically cleaves double stranded RNA hybrids of pre-crRNA and tracrRNA. Moreover, the effector modules differ substantially among the CRISPR-Cas systems (Makarova et al., 2011; Makarova et al., 2015; Charpentier et al., 2015). In the latest classification, the diverse CRISPR-Cas systems are divided into two classes according to the configuration of their effector modules: Class 1 CRISPR systems utilize several Cas proteins and the crRNA to form an effector complex, whereas Class 2 CRISPR systems employ a large single-component Cas protein in conjunction with crRNAs to mediate interference (Makarova et al., 2015).

Multiple Class 1 CRISPR-Cas systems, which include the type I and type III systems, have been identified and functionally characterized in detail, revealing the complex architecture and dynamics of the effector complexes (Brouns et al., 2008; Marraffini and Sontheimer, 2008; Hale et al., 2009; Sinkunas et al., 2013; Jackson et al., 2014; Mulepati et al., 2014). Several Class 2 CRISPR-Cas systems have also been identified and experimentally characterized, but they are all type II and employ homologous RNA-guided endonucleases of the Cas9 family as effectors (Barrangou et al., 2007; Garneau et al., 2010; Deltcheva et al., 2011; Sapranauskas et al., 2011; Jinek et al., 2012; Gasiunas et al., 2012). A second, putative Class 2 CRISPR system, tentatively assigned to type V, has been recently identified in several bacterial genomes (http://www.jcvi.org/cgi-bin/tigrfams/HmmReportPage.cgi?acc=TIGR04330) (Schunder et al., 2013; Vestergaard et al., 2014; Makarova et al., 2015). The putative type V CRISPR-Cas systems contain a large, ~1,300 amino acid protein called Cpf1 (CRISPR from *Prevotella* and *Francisella* 1). It remains unknown, however, if Cpf1-containing CRISPR loci indeed represent functional CRISPR systems. Given the broad applications of Cas9 as a genome engineering tool (Hsu et al., 2014; Jiang and Marraffini, 2015), we sought to explore the function of Cpf1-based putative CRISPR systems.

Here we show that Cpf1-containing CRISPR-Cas loci of *Francisella tularensis subsp. novicida U112* encode functional defense systems capable of mediating plasmid interference in bacterial cells guided by the CRISPR spacers. Unlike Cas9 systems, Cpf1-containing CRISPR systems have three features: First, Cpf1-associated CRISPR arrays are processed into mature crRNAs without the requirement of an additional trans-activating crRNA (tracrRNA) (Deltcheva et al., 2011; Chylinski et al., 2013). Second, Cpf1-crRNA complexes efficiently cleave target DNA proceeded by a short T-rich protospacer adjacent motif (PAM), in contrast to the G-rich PAM following the target DNA for Cas9 systems. Third, Cpf1 introduces a staggered DNA double stranded break with a 4 or 5-nt 5′ overhang.

To explore the suitability of Cpf1 for genome editing applications, we characterized the RNA-guided DNA targeting requirements for 16 Cpf1-family proteins from diverse bacteria, we identify two Cpf1 enzymes, from *Acidaminococcus sp. BV3L6* and *Lachnospiraceae bacterium ND2006*, that are capable of mediating robust genome editing in human cells. Collectively, these results establish a Class 2 CRISPR-Cas system that includes an effective single RNA-guided endonuclease with distinct properties that has the potential to substantially advance our ability to manipulate eukaryotic genomes.

RESULTS

Cpf1-containing CRISPR loci are active bacterial immune systems

Cpf1 was first annotated as a CRISPR-associated gene in TIGRFAM (http://www.jcvi.org/cgi-bin/tigrfams/HmmReportPage.cgi?acc=TIGR04330) and has been hypothesized to be the effector of a CRISPR locus that is distinct from the Cas9-containing type II CRISPR-Cas loci that are also present in the genomes of some of the same bacteria, such as multiple strains of *Francisella* and *Prevotella* (Schunder et al., 2013; Vestergaard et al., 2014; Makarova et al., 2015) (Figure 1A). The Cpf1 protein contains a predicted RuvC-like endonuclease domain that is distantly related to the respective nuclease domain of Cas9. However, Cpf1 differs from Cas9 in that it lacks a second, HNH endonuclease domain,

which is inserted within the RuvC-like domain of Cas9. Furthermore, the N-terminal portion of Cpf1 is predicted to adopt a mixed α/β structure and appears to be unrelated to the N-terminal, α-helical recognition lobe of Cas9 (Figure 1A). It has been shown that the nuclease moieties of Cas9 and Cpf1 are homologous to distinct groups of transposon-encoded TnpB proteins, the first one containing both RuvC and HNH nuclease domains and the second one containing the RuvC-like domain only (Makarova and Koonin, 2015). Apart from these distinctions between the effector proteins, the Cpf1-carrying loci encode Cas1, Cas2, and Cas4 proteins that are more closely related to orthologs from types I and III than to those from type II CRISPR systems (Makarova et al., 2015). Taken together, these differences from type II have prompted the classification of Cpf1-encoding CRISPR-Cas loci as the putative type V within Class 2 (Makarova et al., 2015). The features of the putative type V loci, especially the domain architecture of Cpf1, suggest not only that type II and type V systems independently evolved through the association of different adaptation modules (*cas1*, *cas2*, and *cas4* genes) with different TnpB genes, but also that type V systems are functionally unique. The notion that Cpf1-carrying loci are *bona fide* CRISPR systems is further buttressed by the search of microbial genome sequences for similarity to the type V spacers which produced several significant hits to prophage genes, in particular those from *Francisella* (Schunder et al., 2013). Given these observations and the prevalence of Cpf1-family proteins in diverse bacterial species, we sought to test the hypothesis that Cpf1-encoding CRISPR-Cas loci are biologically active and can mediate targeted DNA interference, one of the primary functions of CRISPR systems.

To simplify experimentation, we cloned the *Francisella tularensis subsp. novicida U112* Cpf1 (FnCpf1) locus (Figure 1A) into low-copy plasmids (pFnCpf1) to allow heterologous reconstitution in *Escherichia coli.* Typically, in currently characterized CRISPR-Cas systems, there are two requirements for DNA interference: (i) the target sequence has to match one of the spacers present in the respective CRISPR array, and (ii) the target sequence complementary to the spacer (hereinafter protospacer) has to be flanked by the appropriate protospacer adjacent motif (PAM). Given the completely uncharacterized functionality of the FnCpf1 CRISPR locus, we adapted a previously described plasmid depletion assay (Jiang et al., 2013) to ascertain the activity of Cpf1 and identify the requirement for a PAM sequence and its respective location relative to the protospacer (5′ or 3′) (Figure 1B). We constructed two libraries of plasmids carrying a protospacer matching the first spacer in the FnCpf1 CRISPR array with the 5′ or 3′ 7 bp sequences randomized. Each plasmid library was transformed into *E. coli* that heterologously expressed the FnCpf1 locus or into a control *E. coli* strain carrying the empty vector. Using this assay, we determined the PAM sequence and location by identifying nucleotide motifs that are preferentially depleted in cells heterologously expressing the FnCpf1 locus. We found that the PAM for FnCpf1 is located upstream of the 5′ end of the displaced strand of the protospacer and has the sequence 5′-TTN (Figures 1C–D and S1). The 5′ location of the PAM is also observed in type I CRISPR systems, but not in type II systems, where Cas9 employs PAM sequences that are located on the 3′ end of the protospacer (Mojica et al., 2009; Garneau et al., 2010). Beyond the identification of the PAM, the results of the depletion assay clearly indicate that heterologously expressed Cpf1 loci are capable of efficient interference with plasmid DNA.

To further characterize the PAM requirements, we analyzed plasmid interference activity by transforming *cpf1*-locus expressing cells with plasmids carrying protospacer 1 flanked by 5′-TTN PAMs. We found that all 5′-TTN PAMs were efficiently targeted (Figure 1E). In addition, 5′-CTA but not 5′-TCA was also efficiently targeted (Figure 1E), suggesting that the middle T is more critical for PAM recognition than the first T and that, in agreement with the sequence motifs depleted in the PAM discovery assay (Figure S1D), the PAM might be more relaxed than 5′-TTN.

The Cpf1-associated CRISPR array is processed independent of tracrRNA

After showing that *cpf1*-based CRISPR loci are able to mediate robust DNA interference, we performed small RNA sequencing to determine the exact identity of the crRNA produced by these loci. By sequencing small RNAs extracted from a *Francisella tularensis subsp. novicida U112* culture, we found that the CRISPR array is processed into short mature crRNAs of 42–44 nt in length. Each mature crRNA begins with 19 nt of the direct repeat followed by 23–25 nt of the spacer sequence (Figure 2A). This crRNA arrangement contrasts that in type II CRISPR-Cas systems where the mature crRNA starts with 20–24 nt of spacer sequence followed by ~22 nt of direct repeat (Deltcheva et al., 2011; Chylinski et al., 2013). Unexpectedly, apart from the crRNAs, we did not observe any robustly expressed small transcripts near the *Francisella cpf1* locus that might correspond to tracrRNAs, which are associated with Cas9-based systems.

To confirm that no additional RNAs are required for crRNA maturation and DNA interference, we constructed an expression plasmid using synthetic promoters to drive the expression of *Francisella cpf1* (FnCpf1) and the CRISPR array (pFnCpf1_min). Small RNAseq of *E. coli* expressing this plasmid still showed robust processing of the CRISPR array into mature crRNA (Figure 2B), indicating that FnCpf1 and its CRISPR array are the only elements required from the FnCpf1 locus to achieve crRNA processing. Furthermore, *E. coli* expressing pFnCpf1_min as well as pFnCpf1_ΔCas, a plasmid with all of the *cas* genes removed but retaining native promoters driving the expression of FnCpf1 and the CRISPR array, also exhibited robust DNA interference, demonstrating that FnCpf1 and crRNA are sufficient for mediating DNA targeting (Figure 2C). By contrast, Cas9 requires both crRNA and tracrRNA to mediate targeted DNA interference (Deltcheva et al., 2011; Zhang et al., 2013).

Cpf1 is a single crRNA-guided endonuclease

The finding that FnCpf1 can mediate DNA interference with crRNA alone is highly surprising given that Cas9 recognizes crRNA through the duplex structure between crRNA and tracrRNA (Jinek et al., 2012; Nishimasu et al., 2014), as well as the 3′ secondary structure of the tracrRNA (Hsu et al., 2013; Nishimasu et al., 2014). To ensure that crRNA is indeed sufficient for forming an active complex with FnCpf1 and mediating RNA-guided DNA cleavage, we investigated whether FnCpf1 supplied only with crRNA can cleave target DNA *in vitro*. We purified FnCpf1 (Figure S2) and assayed its ability to cleave the same protospacer 1-containing plasmid used in the bacterial DNA interference experiments (Figure 3A). We found that FnCpf1 along with an *in vitro* transcribed mature crRNA targeting protospacer 1 was able to efficiently cleave the target plasmid in a Mg^{2+}- and

crRNA-dependent manner (Figure 3B). Moreover, FnCpf1 was able to cleave both supercoiled and linear target DNA (Figure 3C). These results clearly demonstrate the sufficiency of FnCpf1 and crRNA for RNA-guided DNA cleavage.

We also mapped the cleavage site of FnCpf1 using Sanger sequencing of the cleaved DNA ends. We found that FnCpf1-mediated cleavage results in a 5-nt 5′ overhang (Figures 3A, 3D, and S3A–D), which is different from the blunt cleavage product generated by Cas9 (Garneau et al., 2010; Jinek et al., 2012; Gasiunas et al., 2012). The staggered cleavage site of FnCpf1 is distant from the PAM: cleavage occurs after the 18th base on the non-targeted (+) strand and after the 23rd base on the targeted (−) strand (Figures 3A, 3D, and S3A–D). Using double-stranded oligo substrates with different PAM sequences, we also found that FnCpf1 requires the 5′-TTN PAM to be in a duplex form in order to cleave the target DNA (Figure 3E).

The RuvC-like domain of Cpf1 mediates RNA-guided DNA cleavage

The RuvC-like domain of Cpf1 retains all the catalytic residues of this family of endonucleases (Figures 4A and S4) and is thus predicted to be an active nuclease. Therefore we generated three mutants, FnCpf1(D917A), FnCpf1(E1006A), and FnCpf1(D1225A) (Figure 4A) to test whether the conserved catalytic residues are essential for the nuclease activity of FnCpf1. We found that the D917A and E1006A mutations completely inactivated the DNA cleavage activity of FnCpf1, and D1255A significantly reduced nucleolytic activity (Figure 4B). These results are in contrast to the mutagenesis results for *Streptococcus pyogenes* Cas9 (SpCas9), where mutation of the RuvC (D10A) and HNH (N863A) nuclease domains converts SpCas9 into a DNA nickase (i.e. inactivation of each of the two nuclease domains abolished the cleavage of one of the DNA strands) (Jinek et al., 2012; Gasiunas et al., 2012) (Figure 4B). These findings suggest that the RuvC-like domain of FnCpf1 cleaves both strands of the target DNA, perhaps in a dimeric configuration. Interestingly, size exclusion gel filtration of FnCpf1 shows that the protein is eluted at a size of approximately 300 kD, twice the molecular weight of a FnCpf1 monomer (Figure S2B).

Sequence and structural requirements for the Cpf1 crRNA

Compared with the guide RNA for Cas9, which has elaborate RNA secondary structure features that interact with Cas9 (Nishimasu et al., 2014), the guide RNA for FnCpf1 is notably simpler and only consists of a single stem loop in the direct repeat sequence (Figure 3A). We explored the sequence and structural requirements of crRNA for mediating DNA cleavage with FnCpf1.

We first examined the length requirement for the guide sequence and found that FnCpf1 requires at least 16 nt of guide sequence to achieve detectable DNA cleavage and a minimum of 18 nt of guide sequence to achieve efficient DNA cleavage *in vitro* (Figure 5A). These requirements are similar to those demonstrated for SpCas9 where a minimum of 16 to 17 nt of spacer sequence is required for DNA cleavage (Cencic et al., 2014; Fu et al., 2014). We also found that the seed region of the FnCpf1 guide RNA is approximately within the first 5 nt on the 5′ end of the spacer sequence (Figures 5B and S3E).

Next, we studied the effect of direct repeat mutations on the RNA-guided DNA cleavage activity. The direct repeat portion of mature crRNA is 19 nt long (Figure 2A). Truncation of the direct repeat revealed that at least 16, but optimally more than 17 nt, of the direct repeat is required for cleavage. Mutations in the stem loop that preserved the RNA duplex did not affect the cleavage activity, whereas mutations that disrupted the stem loop duplex structure completely abolished cleavage (Figure 5D). Finally, base substitutions in the loop region did not affect nuclease activity, whereas the U immediately 5′ of the spacer sequence could not be substituted (Figure 5E). Collectively, these results suggest that FnCpf1 recognizes the crRNA through a combination of sequence-specific and structural features of the stem loop.

Cpf1-family proteins from diverse bacteria share common crRNA structures and PAMs

Based on our previous experience in harnessing Cas9 for genome editing in mammalian cells, only a small fraction of bacterial nucleases can function efficiently when heterologously expressed in mammalian cells (Cong et al., 2013; Ran et al., 2015). Therefore, in order to assess the feasibility of harnessing Cpf1 as a genome editing tool, we exploited the diversity of Cpf1-family proteins available in the public sequences databases. A BLAST search of the WGS database at the NCBI revealed 46 non-redundant Cpf1-family proteins (Figure S5A), from which we chose 16 candidates that, based on our phylogenetic reconstruction (Figure S5A), represented the entire Cpf1 diversity (Figures 6A and S5). These Cpf1-family proteins span a range of lengths between ~1200 and ~1500 amino acids.

The direct repeat sequences for each of these Cpf1-family proteins show strong conservation in the 19 nucleotides at the 3′ of the direct repeat, the portion of the repeat that is included in the processed crRNA (Figure 6B). The 5′ sequence of the direct repeat is much more diverse. Of the 16 Cpf1-family proteins chosen for analysis, three (2 - *Lachnospiraceae bacterium MC2017*, Lb3Cpf1; 3 - *Butyrivibrio proteoclasticus*, BpCpf1; and 6 - *Smithella sp. SC_K08D17*, SsCpf1) were associated with direct repeat sequences that are notably divergent from the FnCpf1 direct repeat (Figure 6B). However, even these direct repeat sequences preserved stem loop structures that were identical or nearly-identical to the FnCpf1 direct repeat (Figure 6C).

Given the strong structural conservation of the direct repeats that are associated with many of the Cpf1-family proteins, we first tested whether the orthologous direct repeat sequences are able to support FnCpf1 nuclease activity *in vitro*. As expected, the direct repeats that contained conserved stem sequences were able to function interchangeably with FnCpf1. By contrast, the direct repeats from candidate 2 (Lb3Cpf1) and 6 (SsCpf1) were unable to support FnCpf1 cleavage activity (Figure 6D). The direct repeat from candidate 3 (BpCpf1) supported only a low level of FnCpf1 nuclease activity (Figure 6D), possibly due to the conservation of the 3′-most U.

Next, we applied the *in vitro* PAM identification assay (Figure S6A) to determine the PAM sequence for each Cpf1-family protein. We were able to identify the PAM sequence for 7 new Cpf1-family proteins (Figures 6E and S6B–C), and the screen confirmed the PAM for FnCpf1 as 5′-TTN. The remaining 8 tested Cpf1 proteins did not show efficient cleavage during *in vitro* reconstitution. The PAM sequences for the Cpf1-family proteins were

predominantly T-rich, only varying in the number of Ts constituting each PAM (Figure 6E and S6B–C).

Cpf1 can be harnessed to facilitate genome editing in human cells

We tested each Cpf1-family protein, for which we were able to identify a PAM, for nuclease activity in mammalian cells. We codon optimized each of these genes and attached a C-terminal nuclear localization signal (NLS) for optimal expression and nuclear targeting in human cells (Figure 7A). To test the activity of each Cpf1-family protein, we selected a guide RNA target site within the *DNMT1* gene (Figure 7B). We first found that each of the Cpf1-family proteins along with its respective crRNA designed to target *DNMT1* was able to cleave a PCR amplicon of the *DNMT1* genomic region *in vitro* (Figure 7C). However, when tested in human embryonic kidney 293FT (HEK 293FT) cells, only 2 out of the 8 Cpf1-family proteins (7 – AsCpf1 and 13 – LbCpf1) exhibited detectable levels of nuclease-induced indels (Figures 7C and 7D). This result is consistent with previous experiments with Cas9 where only a small number of Cas9 orthologs were successfully harnessed for genome editing in mammalian cells (Ran et al., 2015).

We further tested each Cpf1-family protein with additional genomic targets and found that AsCpf1 and LbCpf1 consistently mediated robust genome editing in HEK293FT cells, whereas the remaining Cpf1 proteins showed either no detectable activity or only sporadic activity (Figures 7E and S7), despite robust expression (Figure S6D). The only Cpf1 candidate that expressed poorly was PdCpf1 (Figure S6D). When compared to Cas9, AsCpf1 and LbCpf1 mediated comparable levels of indel formation (Figure 7E). Additionally, we used *in vitro* cleavage followed by Sanger sequencing of the cleaved DNA ends and found that 7 - AsCpf1 and 13 - LbCpf1 also generated staggered cleavage sites (Figures S6E and S6F, respectively).

DISCUSSION

In this work, we characterize Cpf1-containing Class 2 CRISPR systems, classified as type V, and show that its effector protein, Cpf1, is a single RNA-guided endonuclease. Cpf1 substantially differs from Cas9, to date the only other experimentally characterized Class 2 effector, in terms of structure and function and might provide important advantages for genome editing applications. Specifically, Cpf1 contains a single identified nuclease domain, in contrast to the two nuclease domains present in Cas9. The results presented here show that in FnCpf1, inactivation of RuvC-like domain abolishes cleavage of both DNA strands. Conceivably, FnCpf1 forms a homodimer (Figure S2B), with the RuvC-like domains of each of the two subunits cleaving one DNA strand. However, it is also likely that FnCpf1 contains a second yet-to-be-identified nuclease domain. Structural characterization of Cpf1-RNA-DNA complexes will allow testing of these hypotheses and elucidating the cleavage mechanism.

Perhaps the most notable feature of Cpf1 is that it is a single crRNA-guided endonuclease. Unlike Cas9, which requires tracrRNA to process crRNA arrays and both crRNA and tracrRNA to mediate interference (Deltcheva et al., 2011), Cpf1 processes crRNA arrays independent of tracrRNA and Cpf1-crRNA complexes alone cleave target DNA molecules,

without the requirement for any additional RNA species. This feature could simplify the design and delivery of genome editing tools. For example, the shorter (~42 nt) crRNA employed by Cpf1 has practical advantages over the long (~100 nt) guide RNA in Cas9-based systems, because shorter RNA oligos are significantly easier and cheaper to synthesize. In addition, these findings raise more fundamental questions regarding the guide processing mechanism of the type V CRISPR-Cas systems. In the case of type II, processing of the pre-crRNA is catalyzed by the bacterial RNAse III, which recognizes the long duplex formed by the tracrRNA and the complementary portion of the direct repeat (Deltcheva et al., 2011). Such long duplexes are not present in the pre-crRNA of type V systems, making it unlikely that RNAse III is responsible for processing. Further experiments aimed at elucidating the processing mechanism of type V systems will shed light on the functional diversity of different CRISPR-Cas systems.

Cpf1 generates a staggered cut with a 5′ overhang, in contrast to the blunt ends generated by Cas9 (Garneau et al., 2010; Jinek et al., 2012; Gasiunas et al., 2012). This structure of the cleavage product could be particularly advantageous for facilitating non-homologous end joining (NHEJ)-based gene insertion into the mammalian genome (Maresca et al., 2013). Being able to program the exact sequence of a sticky end would allow researchers to design the DNA insert so that it integrates into the genome in the proper orientation. Specifically, in non-dividing cells, where genome editing via homology-directed repair (HDR) mechanisms are especially challenging (Chan et al., 2011), Cpf1 could provide an effective way to precisely introduce DNA into the genome via non-HDR mechanisms.

Another potentially useful feature of Cpf1 that might aid the introduction of new DNA sequences is that Cpf1 cleaves target DNA at the distal end of the protospacer, far away from the seed region. Therefore, Cpf1-induced indels will be located far from the target site, which is thus preserved for subsequent rounds of Cpf1 cleavage. With Cas9, any indel resulting from the dominant NHEJ repair pathway will disrupt the target site, effectively eliminating the possibility of inserting new DNA at that site in that particular cell. In the case of Cpf1, it appears possible that, if the first round of targeting results in an indel, a subsequent round of targeting could yet be repaired via HDR. Future exploration of these and other strategies using Cpf1 and other Class 2 effectors is expected to bring solutions for some of the biggest challenges facing genome editing.

The T-rich PAMs of the Cpf1 family also allow for applications in genome editing in organisms with particularly AT-rich genomes, such as *Plasmodium falciparum* (Gardner et al., 2002), or areas of interest with AT-enrichment, such as scaffold/matrix attachment regions. To date, all characterized mammalian genome editing proteins require the presence of at least one G (Hsu et al., 2014; Jiang et al., 2015), so the T and T/C-dependent PAMs of Cpf1-family proteins expand the targeting range of RNA-guided genome editing nucleases.

The natural diversity of CRISPR systems provides a wealth of opportunities for understanding the origin and evolution of prokaryotic adaptive immunity, as well as for harnessing potentially transformative biotechnological tools. There is little doubt that, beyond the already classified and characterized diversity of the CRISPR-Cas types, there are additional systems with distinctive characteristics that await exploration and could further

enhance genome editing and other areas of biotechnology as well as shed further light on the evolution of these defense systems.

EXPERIMENTAL PROCEDURES

Generation of heterologous plasmids

To generate the FnCpf1 locus for heterologous expression, genomic DNA from *Francisella Novicida* (generous gift from Wayne Conlan) was PCR amplified using Herculase II polymerase (Agilent Technologies) and cloned into pACYC-184 using Gibson cloning (New England Biolabs). Cells harboring plasmids were made competent using the Z-competent kit (Zymo). Sequences of all bacterial expression plasmids can be found in Table S1.

Bacterial RNA-sequencing

RNA was isolated from stationary phase bacteria by first resuspending *F. novicida* (generous gift from David Weiss) or *E. coli* in TRIzol and then homogenizing the bacteria with zirconia/silica beads (BioSpec Products) in a BeadBeater (BioSpec Products) for 3 one-minute cycles. Total RNA was purified from homogenized samples with the Direct-Zol RNA miniprep protocol (Zymo), DNase treated with TURBO DNase (Life Technologies), and 3′ dephosphorylated with T4 Polynucleotide Kinase (New England Biolabs). rRNA was removed with the bacterial Ribo-Zero rRNA removal kit (Illumina). RNA libraries were prepared from rRNA-depleted RNA using NEBNext® Small RNA Library Prep Set for Illumina (New England Biolabs) and size selected using the Pippin Prep (Sage Science)

For heterologous *E. coli* expression of the FnCpf1 locus, RNA sequencing libraries were prepared from rRNA-depleted RNA using a derivative of the previously described CRISPR RNA sequencing method (Heidrich et al., 2015). Briefly, transcripts were poly-A tailed with *E. coli* Poly(A) Polymerase (New England Biolabs), ligated with 5′ RNA adapters using T4 RNA Ligase 1 (ssRNA Ligase) High Concentration (New England Biolabs), and reverse transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent Technologies). cDNA was PCR amplified with barcoded primers using Herculase II polymerase (Agilent Technologies).

RNA-sequencing analysis

The prepared cDNA libraries were sequenced on a MiSeq (Illumina). Reads from each sample were identified on the basis of their associated barcode and aligned to the appropriate RefSeq reference genome using BWA (Li and Durbin, 2009). Paired-end alignments were used to extract entire transcript sequences using Picard tools (http://broadinstitute.github.io/picard), and these sequences were analyzed using Geneious 8.1.5.

In vivo FnCpf1 PAM Screen

Randomized PAM plasmid libraries were constructed using synthesized oligonucleotides (IDT) consisting of 7 randomized nucleotides either upstream or downstream of the FnCpf1 spacer 1. The randomized ssDNA oligos (Table S1) were made double stranded by annealing to a short primer and using the large Klenow fragment (New England Biolabs) for

second strand synthesis. The dsDNA product was assembled into a linearized pUC19 using Gibson cloning (New England Biolabs). Competent Stbl3 *E. coli* (Invitrogen) were transformed with the cloned products, and more than 10^7 cells were collected and pooled. Plasmid DNA was harvested using a Maxi-prep kit (Qiagen). We transformed 360 ng of the pooled library into *E. coli* cells carrying the FnCpf1 locus or pACYC184 control. After transformation, cells were plated on ampicillin. After 16 hours of growth, $>4*10^6$ cells were harvested and plasmid DNA was extracted using a Maxi-prep kit (Qiagen). The target PAM region was amplified and sequenced using a MiSeq (Illumina) with single-end 150 cycles.

Computational PAM discovery pipeline

PAM regions were extracted, counted, and normalized to total reads for each sample. For a given PAM, enrichment was measured as the log ratio compared to pACYC184 control, with a 0.01 psuedocount adjustment. PAMs above a 3.5 enrichment threshold were collected and used to generate sequence logos (Crooks et al., 2004).

PAM validation

Sequences corresponding to both PAMs non-PAMs were cloned into digested pUC19 and ligated with T4 ligase (Enzymatics). Competent *E. coli* with either the FnCpf1 locus plasmid or pACYC184 control plasmid were transformed with 20ng of PAM plasmid and plated on LB agar plates supplemented with ampicillin and chloramphenicol. Colonies were counted after 18 hours.

Synthesis of crRNAs and sgRNAs

All crRNA and sgRNAs used *in vitro* were synthesized using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). ssDNA oligos (Table S1) corresponding to the reverse complement of the target RNA sequence were synthesized from IDT and annealed to a short T7 priming sequence. T7 transcription was performed for 4 hours and then RNA was purified using the MEGAclear™ Transcription Clean-Up Kit (Ambion).

Purification of Cpf1 Protein

FnCpf1 protein was cloned into a bacterial expression vector (6-His-MBP-TEV-Cpf1, a pET based vector kindly given to us by Doug Daniels). Two liters of Terrific Broth growth media with 100 μg/mL ampicillin was inoculated with 10 mL overnight culture 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 hours 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, 2M NaCl, 5 mM $MgCl_2$, 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 hour to clear the lysate. The lysate was filtered through 0.22 micron 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 MgCl, 2 mM DTT). After dialysis, TEV cleavage was confirmed by SDS-PAGE, and the sample was concentrated to 500 μL prior to loading on a gel filtration column (HiLoad 16/600 Superdex 200) via FPLC (AKTA Pure). Fractions from gel filtration were analyzed by SDS-PAGE; fractions containing Cpf1 were pooled and concentrated to 200 μL and either used directly for biochemical assays or frozen at −80°C for storage. Gel filtration standards were run on the same column equilibrated in 2M NaCl, Hepes pH 7.0 to calculate the approximate size of FnCpf1.

Generation of Cpf1 Protein Lysate

Cpf1 proteins codon optimized for human expression were synthesized with an N-terminal nuclear localization tag and cloned into the pcDNA3.1 expression plasmid by Genscript. 2000ng of Cpf1 expression plasmids were transfected into 6-well plates of HEK293FT cells at 90% confluency using Lipofectamine 2000 reagent (Life Technologies). 48 hours later, cells were harvested by washing once with DPBS (Life Technologies) and scraping in lysis buffer [20mM Hepes pH 7.5, 100mM KCl, 5mM $MgCl_2$, 1 mM DTT, 5% glycerol, 0.1% Triton X-100, 1X cOmplete Protease Inhibitor Cocktail Tablets (Roche)]. Lysate was sonicated for 10 minutes in a Biorupter sonicator (Diagenode) and then centrifuged. Supernatant was frozen for subsequent use in *in vitro* cleavage assays.

In vitro cleavage assay

Cleavage *in vitro* was performed either with purified protein or mammalian lysate with protein at 37°C in cleavage buffer (NEBuffer 3, 5mM DTT) for 20 minutes. The cleavage reaction used 500ng of synthesized crRNA or sgRNA and 200ng of target DNA. Target DNA involved either protospacers cloned into pUC19 or PCR amplicons of gene regions from genomic DNA isolated from HEK293 cells. Reactions were cleaned up using PCR purification columns (Qiagen) and run on 2% agarose E-gels (Life Technologies). For native and denaturing gels to analyze cleavage by nuclease mutants, cleaned-up reactions were run on TBE 6% polyacrylamide or TBE-Urea 6% polyacrylamide gels (Life Technologies)

In vitro Cpf1-family protein PAM Screen

In vitro cleavage reactions with Cpf1-family proteins were run on 2% agarose E-gels (Life Technologies). Bands corresponding to un-cleaved target were gel extracted using QIAquick Gel Extraction Kit (Qiagen) and the target PAM region was amplified and sequenced using a MiSeq (Illumina) with single-end 150 cycles. Sequencing results were entered into the PAM discovery pipeline.

Western blot analysis

Cells were lysed in 1xRIPA buffer (Cell Signaling Technology) supplemented with protease inhibitor cocktail (Roche). Equal volumes cell lysate were run on BOLT 4–12% Bis-Tris gradient gels (Invitrogen) and transferred to PVDF membranes (Millipore). Non-specific antigen binding was blocked with TBS-T (50mM Tris, 150mM NaCl and 0.05% Tween-20) with 5% BLOT-QuickBlocker Reagent (Millipore) for 1 hour. Membranes were incubated

with primary antibodies (anti-HA-tag (Cell Signaling Technology C29F4) or HRP-conjugated GAPDH (Cell Signaling Technology 14C10)) for 1 hour in 1% BLOT-QuickBlocker. Membranes were washed for 3 10 minute washes and anti-HA-tag membranes were further incubated with anti-rabbit antibody (Cell Signaling Technology 7074) for 1h followed by 6 10 minute washes in TBS-T. Proteins were visualized with West Pico Chemiluminescent Substrate (Life Technology) and imaged using the ChemiDoc MP Imaging System (Bio-Rad) and processed with ImageLab software (Bio-Rad).

Activity of Cpf1 cleavage in 293FT cells

Cpf1 proteins codon optimized for human expression were synthesized with an N-terminal nuclear localization tag and cloned into the pcDNA3.1 CMV expression plasmid by Genscript (Table S1). PCR amplicons comprised of a U6 promoter driving expression of the crRNA sequence were generated using Herculase II (Agilent Technologies) and appropriate U6 reverse primers (Table S2). 400ng of Cpf1 expression plasmids and 100ng of the crRNA PCR products were transfected into 24-well plates of HEK293FT cells at 75–90% confluency using Lipofectamine 2000 reagent (Life Technologies). Genomic DNA was harvested using QuickExtract™ DNA Extraction Solution (Epicentre).

SURVEYOR nuclease assay for genome modification

293FT cells were transfected with 400ng Cpf1 expression plasmid and 100ng U6::crRNA PCR-fragments using Lipofectamin 2000 reagent (Life Technologies). Cells were incubated at 37 °C for 72 h post-transfection before genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. The genomic region flanking the CRISPR target site for each gene was PCR amplified, and products were purified using QiaQuick Spin Column (Qiagen) following the manufacturer's protocol. 200 – 500 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 (Integrated DNA Technologies) following the manufacturer's recommended protocol, and analyzed on 4–20% 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.

Deep sequencing to characterize Cpf1 indel patterns in 293FT cells

HEK293FT cells were transfected and harvested as described for assessing activity of Cpf1 cleavage. The genomic region flanking DNMT1 targets were amplified using a two-round PCR region to add Illumina P5 adapters as well as unique sample-specific barcodes to the target amplicons. PCR products were ran on 2% E-gel (Invitrogen) and gel-extracted using QiaQuick Spin Column (Qiagen) as per the manufacturer's recommended protocol. Samples were pooled and quantified by Qubit 2.0 Fluorometer (Life Technologies). The prepared

cDNA libraries were sequenced on a MiSeq (Illumina). Indels were mapped using a Python implementation of the Geneious 6.0.3 Read Mapper.

Computational Analysis of Cpf1 loci

PSI-BLAST program (Altschul et al., 1997) was used to identify Cpf1 homologs in the NCBI NR database using several known Cpf1 sequences as queries with the Cpf1 with the E-value cutoff of 0.01 and low complexity filtering and composition based statistics turned off. The TBLASTN program with the E-value cut-off of 0.01 and low complexity filtering turned off parameters was used to search the NCBI WGS database using the Cpf1 profile (Makarova et al., 2015) as the query. Results of all searches were combined (Table S3). The HHpred program was used with default parameters (Soding et al., 2006) to identify remote sequence similarity using a subset of representative Cpf1 sequences queries. Multiple sequence alignment were constructed using MUSCLE (Edgar, 2004) with manual correction based on pairwise alignments obtained using PSI-BLAST and HHpred programs. Phylogenetic analysis was performed using the FastTree program with the WAG evolutionary model and the discrete gamma model with 20 rate categories (Price et al., 2010). Protein secondary structure was predicted using Jpred 4 (Drozdetskiy et al., 2015). CRISPR repeats were identified using PILER-CR (Edgar, 2007) and CRISPRfinder (Grissa et al., 2007).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank R. Macrae for critical reading of the manuscript. We would like to thank Wayne Conlan for *F. novicida* genomic DNA, David S. Weiss for generously providing *F. novicida* total RNA, Doug Daniels for kindly providing us with the bacterial expression vector, and Sergei Shmakov and Yuri Wolf for help with sequence analysis. E.V.K. and K.S.M is supported by the intramural program of the US Department of Health and Human Services (to the National Library of Medicine). J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. F.Z. is supported by the NIMH (1DP1-MH100706), the Poitras, Vallee, Simons, Paul G. Allen, and New York Stem Cell Foundations, David R. Cheng, and Bob Metcalfe. A patent application has been filed related to this work, and the authors plan to make the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab web site (www.genome-engineering.org).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315:1709–1712. [PubMed: 17379808]

Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol Cell. 2014; 54:234–244. [PubMed: 24766887]

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]

Cencic R, Miura H, Malina A, Robert F, Ethier S, Schmeing TM, Dostie J, Pelletier J. Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. PLoS One. 2014; 9:e109213. [PubMed: 25275497]

Chan F, Hauswirth WW, Wensel TG, Wilson JH. Efficient mutagenesis of the rhodopsin gene in rod photoreceptor neurons in mice. Nucleic acids research. 2011; 39:5955–5966. [PubMed: 21478169]

Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev. 2015; 39:428–441. [PubMed: 25994611]

Chylinski K, Le Rhun A, Charpentier E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol. 2013; 10:726–737. [PubMed: 23563642]

Clark JM. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. Nucleic Acids Res. 1988; 16:9677–9686. [PubMed: 2460825]

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14:1188–1190. [PubMed: 15173120]

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; 471:602–607. [PubMed: 21455174]

Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007; 8:18. [PubMed: 17239253]

Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol. 2014; 32:279–284. [PubMed: 24463574]

Gardner MJ, Shallom SJ, Carlton JM, Salzberg SL, Nene V, Shoaibi A, Ciecko A, Lynn J, Rizzo M, Weaver B, et al. Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature. 2002; 419:531–534. [PubMed: 12368868]

Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010; 468:67–71. [PubMed: 21048762]

Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012; 109:E2579–2586. [PubMed: 22949671]

Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 2007; 35:W52–57. [PubMed: 17537822]

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell. 2009; 139:945–956. [PubMed: 19945378]

Heidrich, N.; Dugar, G.; Vogel, J.; Sharma, C. Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. In: Lundgren, M.; Charpentier, E.; Fineran, PC., editors. CRISPR. Springer; New York: 2015. p. 1-21.

Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010; 327:167–170. [PubMed: 20056882]

Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014; 157:1262–1278. [PubMed: 24906146]

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013; 31:827–832. [PubMed: 23873081]

Jackson RN, Golden SM, van Erp PB, Carter J, Westra ER, Brouns SJ, van der Oost J, Terwilliger TC, Read RJ, Wiedenheft B. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science. 2014; 345:1473–1479. [PubMed: 25103409]

Jiang F, Zhou K, Ma L, Gressel S, Doudna JA. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science. 2015; 348:1477–1481. [PubMed: 26113724]

Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013; 31:233–239. [PubMed: 23360965]

Jiang W, Marraffini LA. CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. Annu Rev Microbiol. 2015

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337:816–821. [PubMed: 22745249]

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011; 6:26. [PubMed: 22115189]

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011; 9:467–477. [PubMed: 21552286]

Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015; 1311:47–75. [PubMed: 25981466]

Makarova KS, Wolf YI, Alkhnbashi O, Costa FSS, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. Updated evolutionary classification of CRISPR-Cas systems and cas genes. Nature Rev Microbiol. 2015 in press.

Maresca M, Lin VG, Guo N, Yang Y. Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. Genome research. 2013; 23:539–546. [PubMed: 23152450]

Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008; 322:1843–1845. [PubMed: 19095942]

Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009; 155:733–740. [PubMed: 19246744]

Mulepati S, Heroux A, Bailey S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science. 2014; 345:1479–1484. [PubMed: 25123481]

Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014; 156:935–949. [PubMed: 24529477]

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010; 5:e9490. [PubMed: 20224823]

Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015; 520:186–191. [PubMed: 25830891]

Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. 2011; 39:9275–9282. [PubMed: 21813460]

Schunder E, Rydzewski K, Grunow R, Heuner K. First indication for a functional CRISPR/Cas system in Francisella tularensis. Int J Med Microbiol. 2013; 303:51–60. [PubMed: 23333731]

Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. EMBO J. 2013; 32:385–394. [PubMed: 23334296]

Soding J, Remmert M, Biegert A, Lupas AN. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res. 2006; 34:W374–378. [PubMed: 16845029]

Sorek R, Lawrence CM, Wiedenheft B. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem. 2013; 82:237–266. [PubMed: 23495939]

Vestergaard G, Garrett RA, Shah SA. CRISPR adaptive immune systems of Archaea. RNA Biol. 2014; 11:156–167. [PubMed: 24531374]

Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol Cell. 2013; 50:488–503. [PubMed: 23706818]

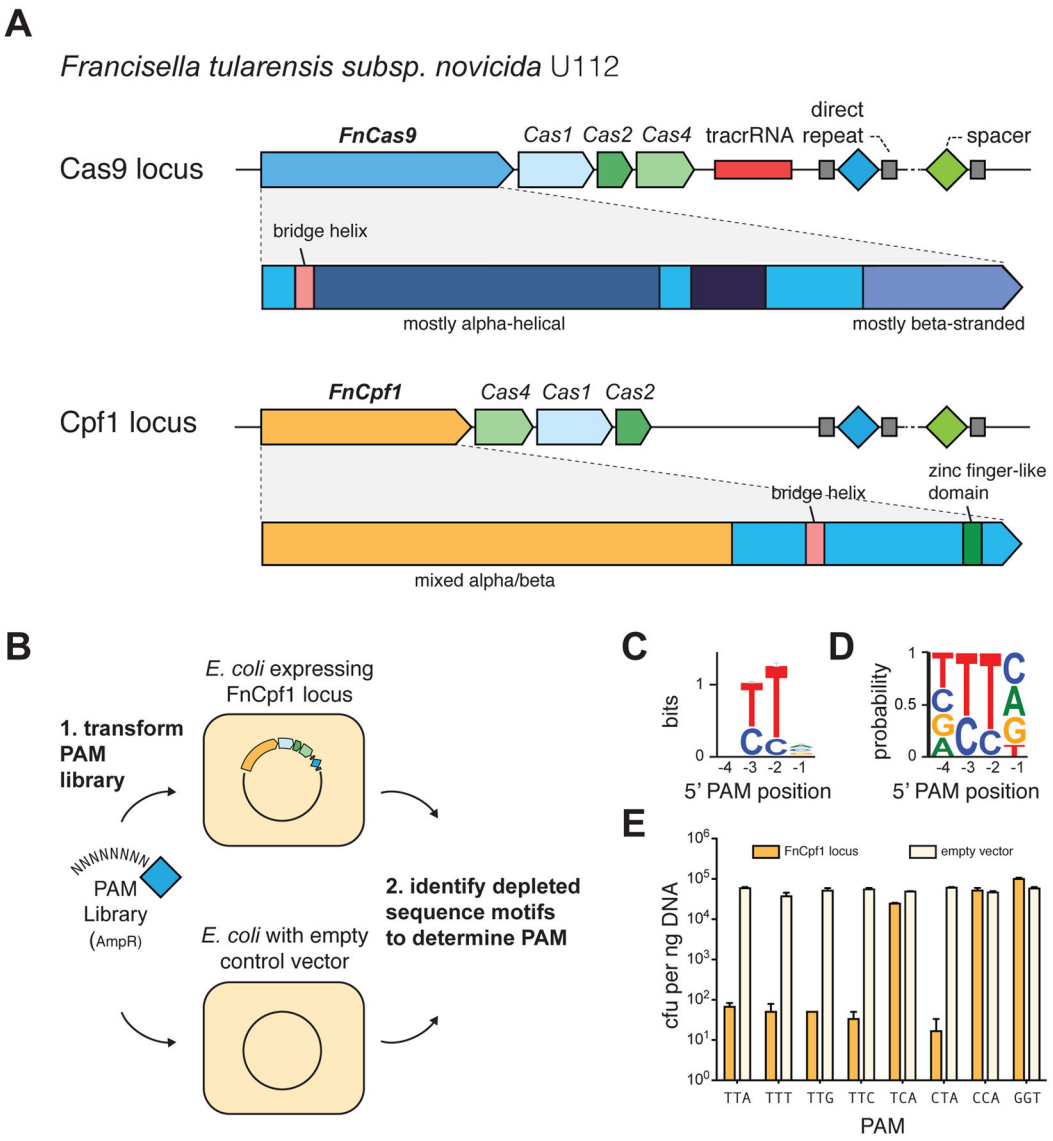


Figure 1. The *Francisella tularensis subsp. novicida U112* Cpf1 CRISPR locus provides immunity against transformation of plasmids containing protospacers flanked by a 5′-TTN PAM

(A) Organization of two CRISPR loci found in *Francisella tularensis subsp. novicida U112* (NC_008601). The domain architectures of FnCas9 and FnCpf1 are compared.

(B) Schematic illustrating the plasmid depletion assay for discovering the PAM position and identity. Competent *E. coli* harboring either the heterologous FnCpf1 locus plasmid (pFnCpf1) or the empty vector control were transformed with a library of plasmids containing the matching protospacer flanked by randomized 5′ or 3′ PAM sequences and selected with antibiotic to deplete plasmids carrying successfully-targeted PAM. Plasmids from surviving colonies were extracted and sequenced to determine depleted PAM sequences.

(C) Sequence logo for the FnCpf1 PAM as determined by the plasmid depletion assay. Letter height at each position is measured by information content; error bars show 95% Bayesian confidence interval.

(D) *E. coli* harboring pFnCpf1 provides robust interference against plasmids carrying 5′-TTN PAMs (n = 3, error bars represent mean ± S.E.M.).
See also Figure S1.

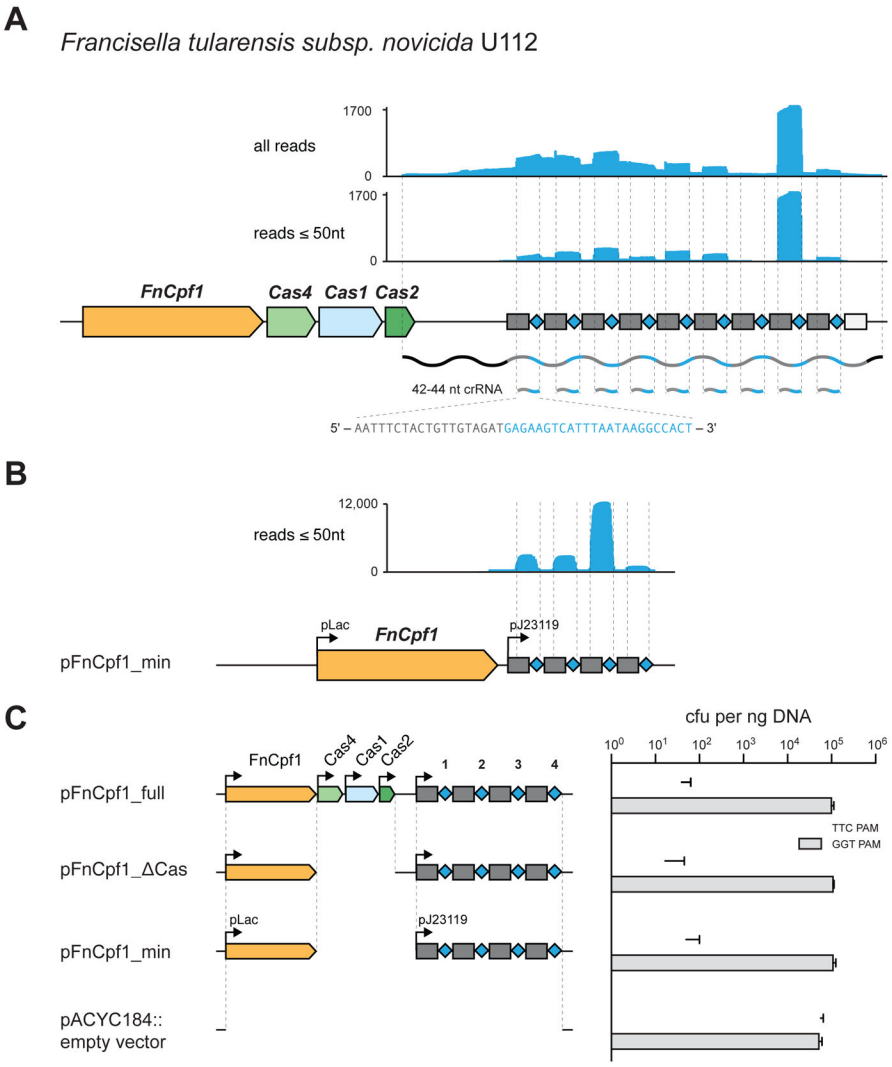


Figure 2. Heterologous expression of FnCpf1 and CRISPR array in *E. coli* is sufficient to mediate plasmid DNA interference and crRNA maturation

(A) Small RNA-seq of *Francisella tularensis subsp. novicida U112* reveals transcription and processing of the FnCpf1 CRISPR array. The mature crRNA begins with a 19 nt partial direct repeat followed by 23–25 nt of spacer sequence.

(B) Small RNA-seq of *E. coli* transformed with a plasmid carrying synthetic promoter-driven FnCpf1 and CRISPR array shows crRNA processing independent of Cas genes and other sequence elements in the FnCpf1 locus.

(C) *E. coli* harboring different truncations of the FnCpf1 CRISPR locus shows that only FnCpf1 and the CRISPR array are required for plasmid DNA interference (n = 3, error bars show mean ± S.E.M.).

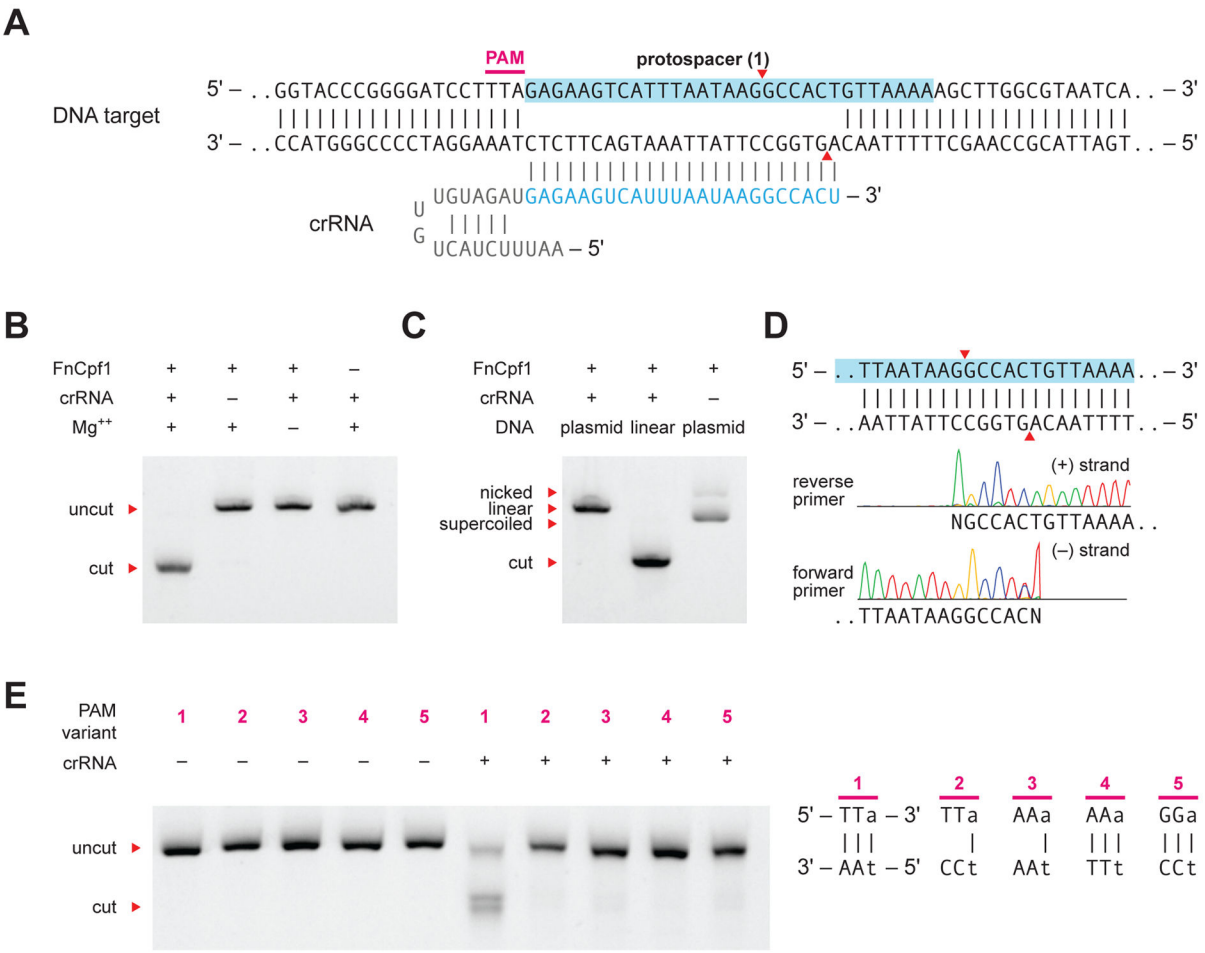


Figure 3. FnCpf1 is guided by crRNA to cleave DNA *in vitro*

(A) Schematic of the FnCpf1 crRNA-DNA targeting complex. Cleavage sites are indicated by red arrows.

(B) FnCpf1 and crRNA alone mediated RNA-guided cleavage of target DNA in a crRNA- and Mg^{2+}-dependent manner.

(C) FnCpf1 cleaves both linear and supercoiled DNA.

(D) Sanger sequencing traces from FnCpf1-digested target show staggered overhangs. The non-templated addition of an additional adenine, denoted as N, is an artifact of the polymerase used in sequencing (Clark, 1988). Reverse primer read represented as reverse complement to aid visualization. See also Figure S3.

(E) Dependency of cleavage on base-pairing at the 5′ PAM. FnCpf1 can only recognize the PAM in correctly Watson-Crick paired DNA.

See also Figures S2 and S3.

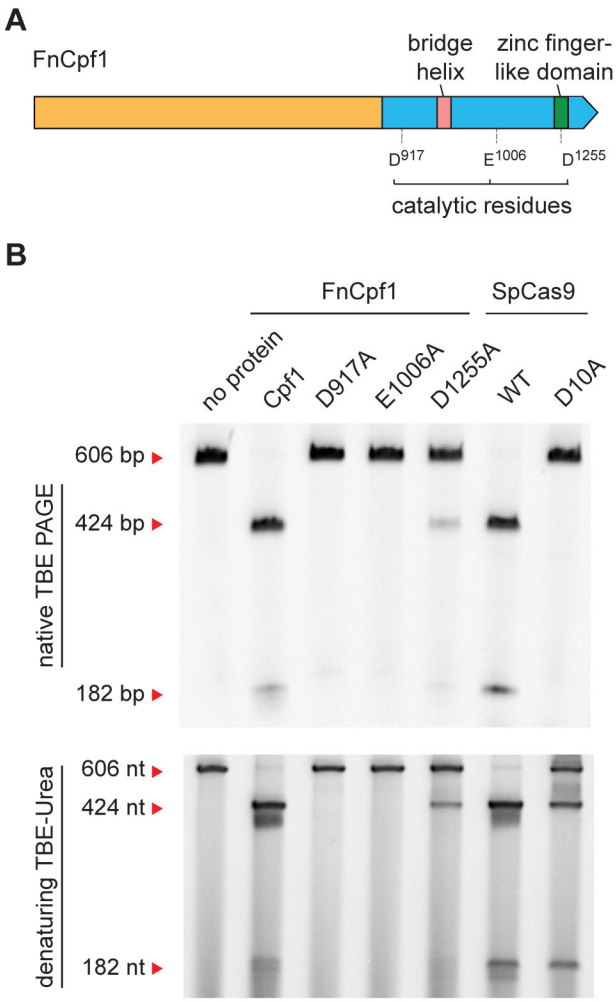


Figure 4. Catalytic residues in the C-terminal RuvC domain of FnCpf1 are required for DNA cleavage

(A) Domain structure of FnCpf1 with RuvC catalytic residues highlighted. The catalytic residues were identified based on sequence homology to *Thermus thermophilus* RuvC (PDB ID: 4EP5).

(B) Native TBE PAGE gel showing that mutation of the RuvC catalytic residues of FnCpf1 (D917A and E1006A) and mutation of the RuvC (D10A) catalytic residue of SpCas9 prevents double stranded DNA cleavage. Denaturing TBE-Urea PAGE gel showing that mutation of the RuvC catalytic residues of FnCpf1 (D917A and E1006A) prevents DNA nicking activity, whereas mutation of the RuvC (D10A) catalytic residue of SpCas9 results in nicking of the target site.

See also Figure S4.

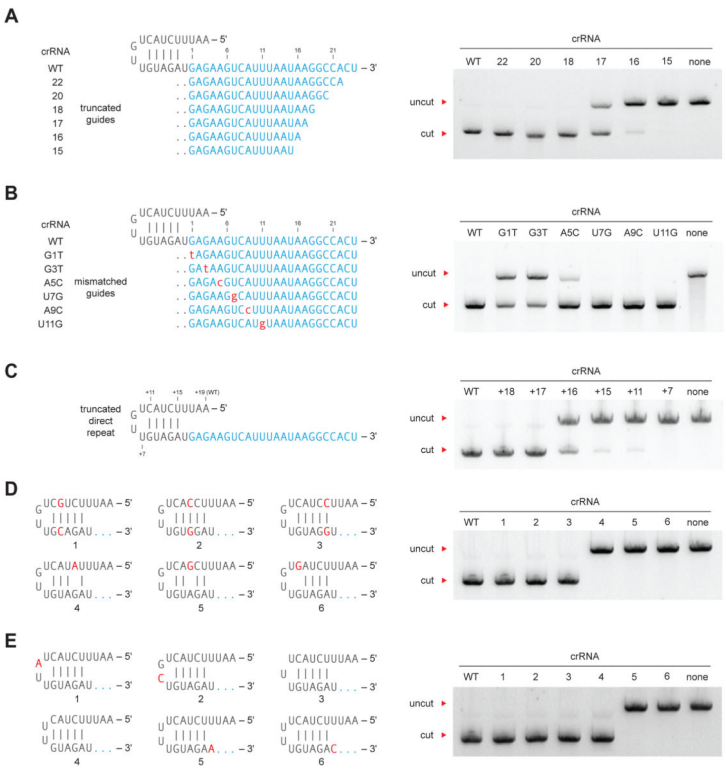


Figure 5. crRNA requirements for FnCpf1 nuclease activity *in vitro*

(A) Effect of spacer length on FnCpf1 cleavage activity.

(B) Effect of crRNA-target DNA mismatch on FnCpf1 cleavage activity. See also Figure S3E.

(C) Effect of direct repeat length on FnCpf1 cleavage activity.

(D) FnCpf1 cleavage activity depends on secondary structure in the stem of the direct repeat RNA structure.

(E) FnCpf1 cleavage activity is unaffected by loop mutations but is sensitive to mutation in the 3′-most base of the direct repeat.

See also Figure S4.

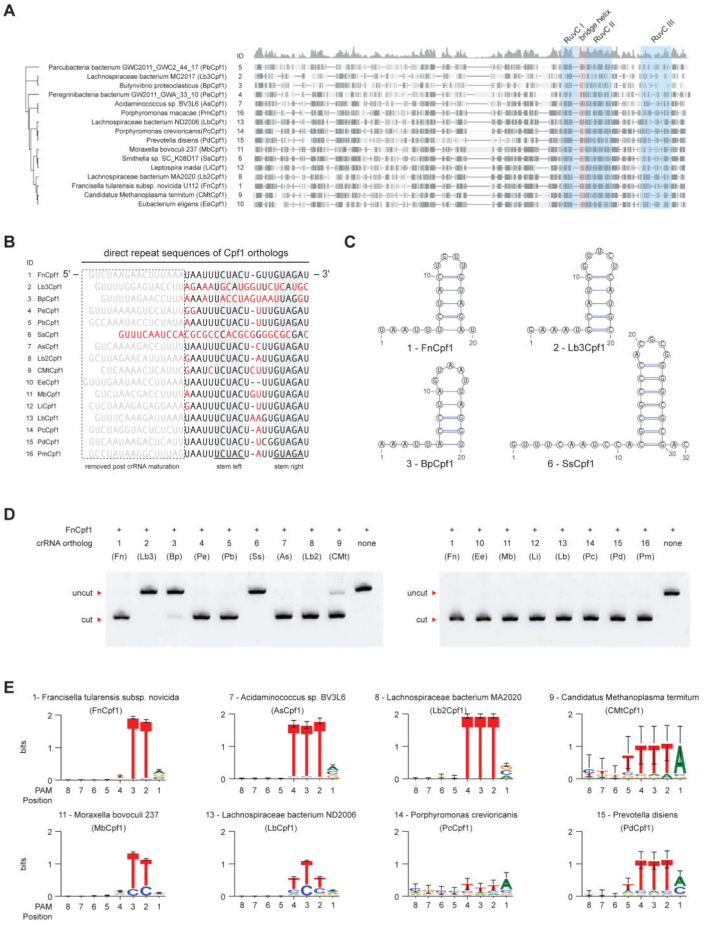


Figure 6. Analysis of Cpf1-family protein diversity and function

(A) Phylogenetic tree of 16 Cpf1 orthologs selected for functional analysis. Conserved sequences are shown in dark gray. The RuvC domain, bridge helix, and zinc finger are highlighted.

(B) Alignment of direct repeats from the 16 Cpf1-family proteins. Sequences that are removed post crRNA maturation are colored gray. Non-conserved bases are colored red. The stem duplex is highlighted in gray.

(C) RNAfold (Lorenz et al., 2011) prediction of the direct repeat sequence in the mature crRNA. Predictions for FnCpf1 along with three diverged type V loci are shown.

(D) Type V crRNAs from different bacteria with similar direct repeat sequences are able to function with FnCpf1 to mediate target DNA cleavage.

(E) PAM sequences for 8 Cpf1-family proteins identified using *in vitro* cleavage of a plasmid library containing randomized PAMs flanking the protospacer.

See also Figures S5 and S6.

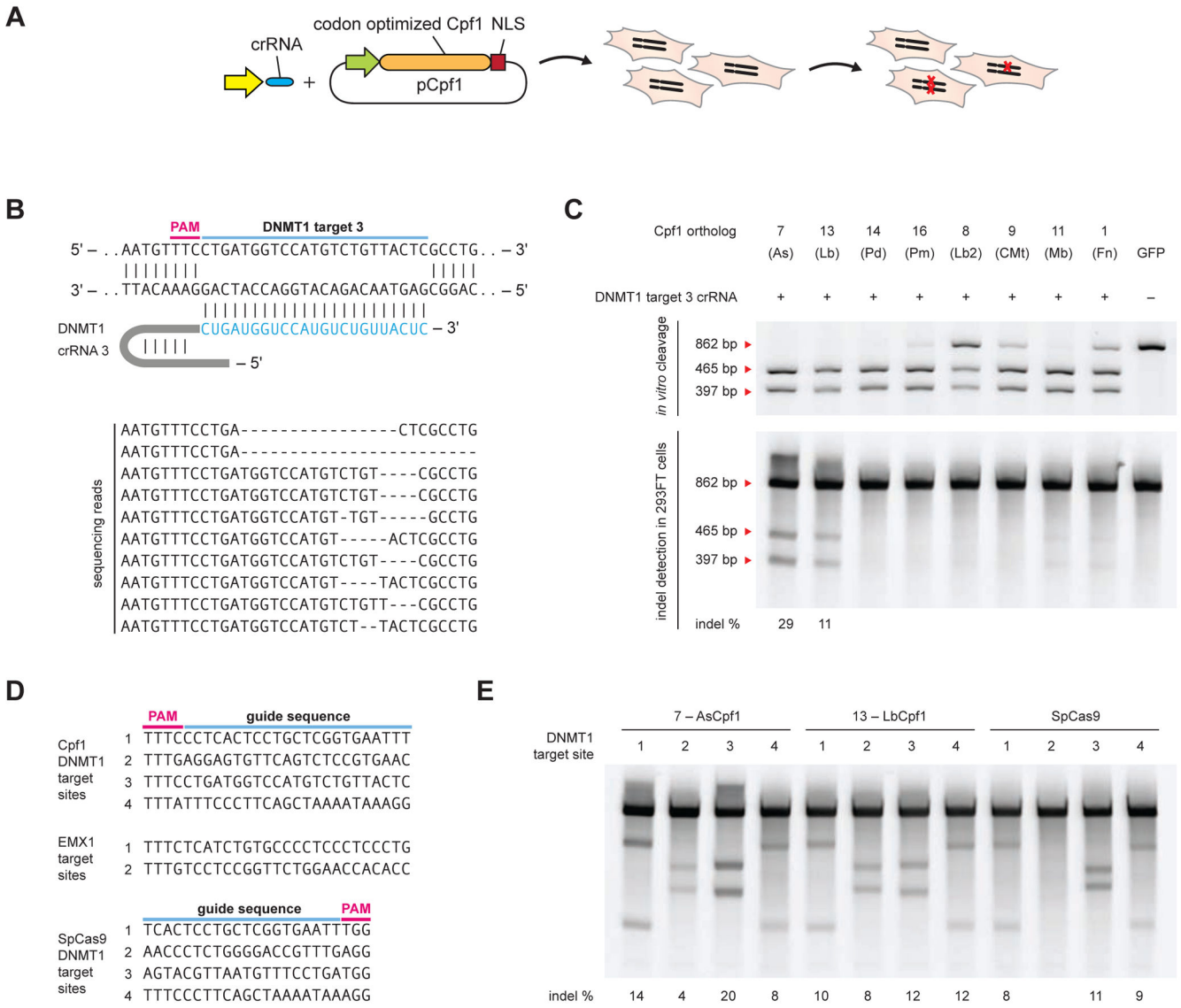


Figure 7. Cpf1 mediates robust genome editing in human cell lines

(A) Eight Cpf1-family proteins are individually expressed in HEK 293FT cells using CMV-driven expression vectors. The corresponding crRNA is expressed using a PCR fragment containing a U6 promoter fused to the crRNA sequence. Transfected cells were analyzed using either Surveyor nuclease assay or targeted deep sequencing.

(B) Schematic showing the sequence of DNMT1-targeting crRNA 3. Sequencing reads show representative indels.

(C) Comparison of *in vitro* and *in vivo* cleavage activity. The DNMT1 target region was PCR amplified and the genomic fragment was used to test Cpf1-mediated cleavage. All 8 Cpf1-family proteins showed DNA cleavage *in vitro* (top), but only candidates 7 – AsCpf1 and 13 – Lb3Cpf1 facilitated robust indel formation in human cells.

(D) Cpf1 and SpCas9 target sequences in the human DNMT1 locus.

(E) Comparison of Cpf1 and SpCas9 genome editing efficiency. Target sites correspond to sequences shown in Figure 7D.

See also Figure S7.