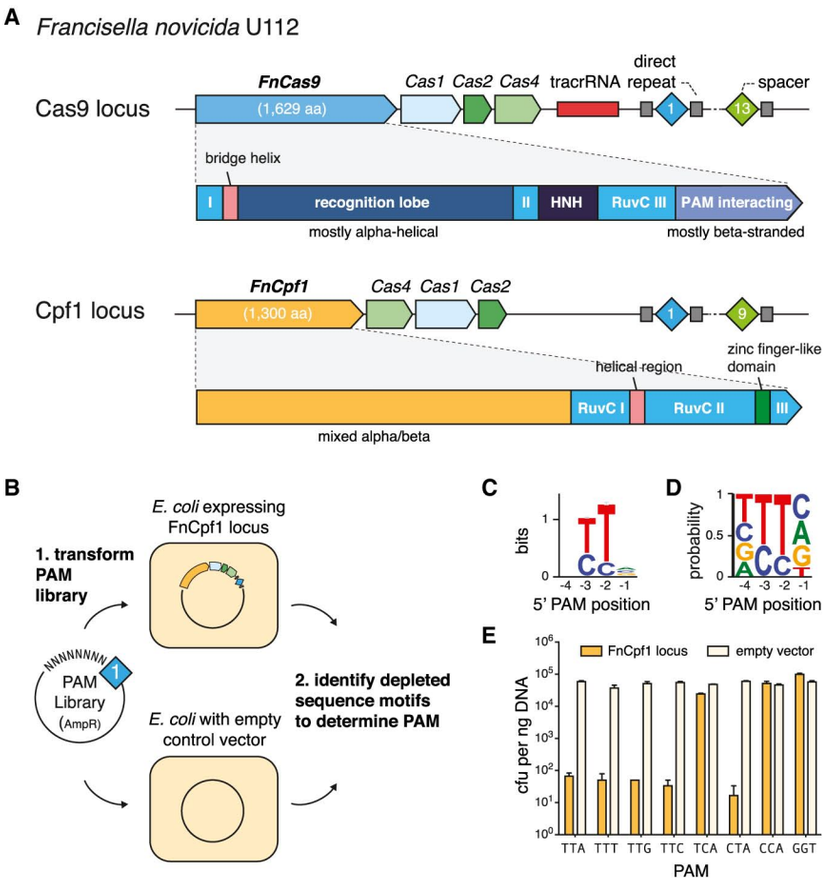

Figure 1. The *Francisella novicida U112* Cpf1 CRISPR Locus Provides Immunity against Transformation of Plasmids Containing Protospacers Flanked by a 5′-TTN PAM

(A) Organization of two CRISPR loci found in *Francisella 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 and D) Sequence logo for the FnCpf1 PAM as determined by the plasmid depletion assay. Letter height at each position is measured by information content (C) or frequency (D); error bars show 95% Bayesian confidence interval.

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

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, whether 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 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, and we identified 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 Cpf1 as 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

Subheading states conclusion

Section content corresponds to Fig. 1 (above)

If you think your data really prove something definitively, and it's important for your reader to know it, you can use subjective terms like "clearly."

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 that 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.

Experimental rationale + methods description without too much detail

To simplify experimentation, we cloned the *Francisella 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: (1) the target sequence has to match one of the spacers present in the respective CRISPR array, and (2) 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, 1D 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.

Quick description of how assay was interpreted

Statement of findings

Conclusion

Transition + rationale + methods

Findings

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.

Conclusion

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 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 with that of type II CRISPR-Cas systems in which 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).

Section content corresponds to Fig. 2 (and so on for following sections)

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

In this paper, comparison with previously discovered systems is important for establishing what the findings mean.

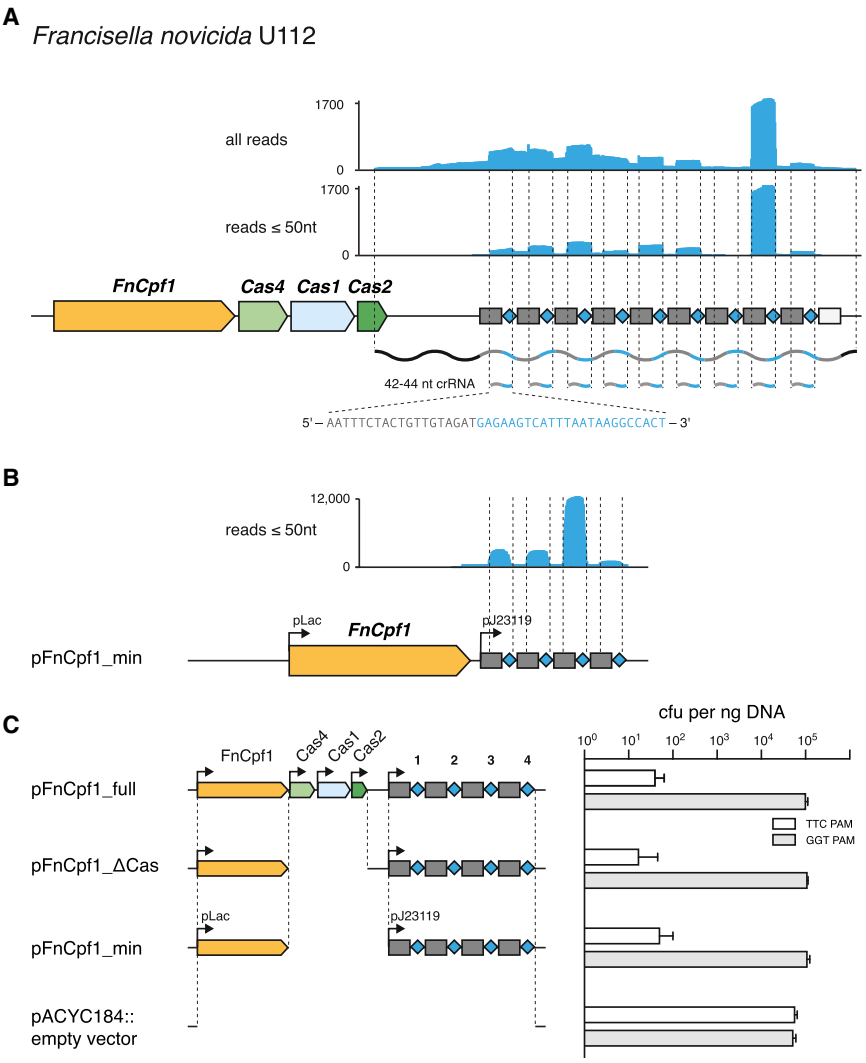

Figure 2. Heterologous Expression of FnCpf1 and CRISPR Array in *E. coli* Is Sufficient to Mediate Plasmid DNA Interference and crRNA Maturation

(A) Small RNA-seq of *Francisella 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 ± SEM).

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–S3D), 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–S3D). 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 of 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 ~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

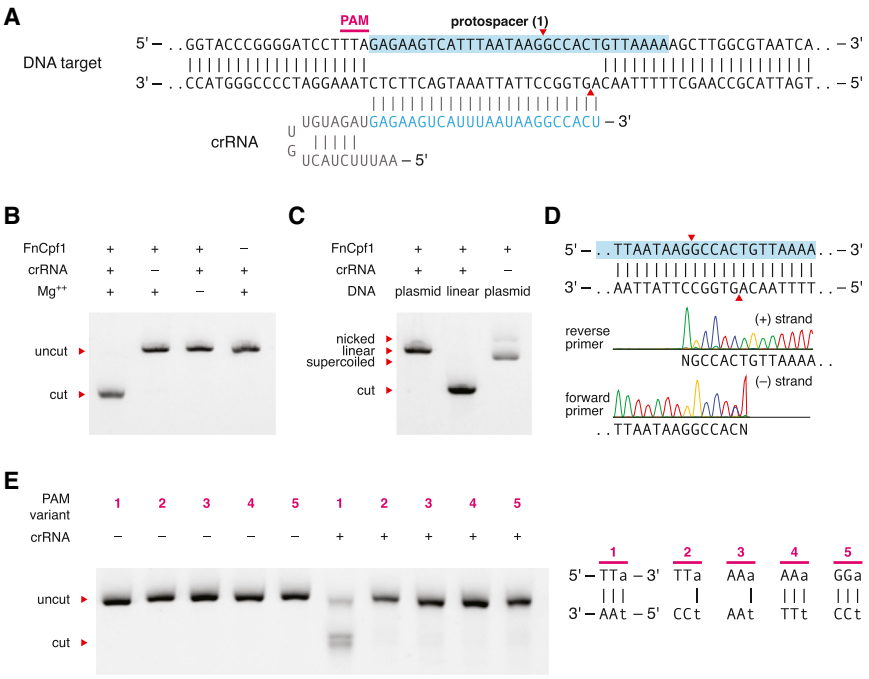

Figure 3. FnCpf1 Is Guided by crRNA to Cleave DNA In Vitro

(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.

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, in which a minimum of 16–17 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 uracil base immediately proceeding 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 ~1,200 and ~1,500 amino acids.

The direct repeat sequences for each of these Cpf1-family proteins show strong conservation in the 19 nt 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 candidates 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 seven

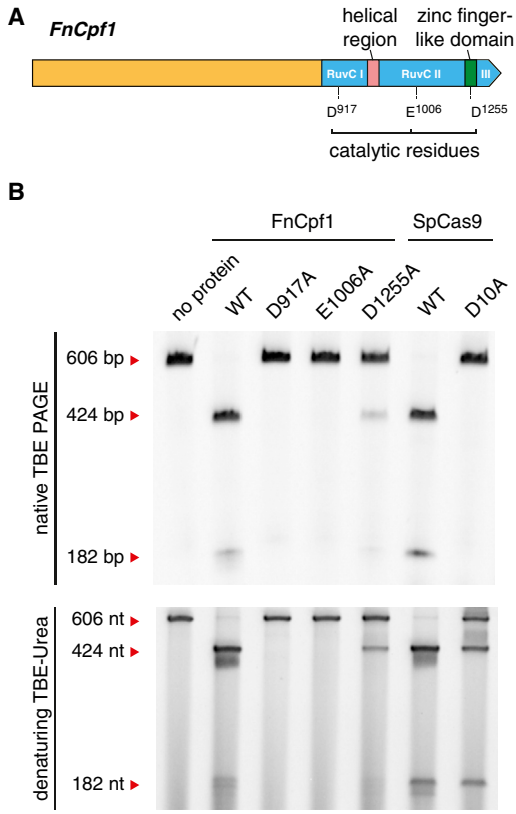

Figure 4. Catalytic Residues in the C-Terminal RuvC Domain of FnCpf1 Are Required for DNA Cleavage

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

new Cpf1-family proteins (Figures 6E, S6B, and S6C), and the screen confirmed the PAM for FnCpf1 as 5′-TTN. The remaining eight 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 (Figures 6E, S6B, and S6C).

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 (HEK293FT) cells, only two out of the eight 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 in which 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, we cannot rule out 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 elucidation of 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

