Systematically investigating the key features of the DNase deactivated Cpf1 for tunable transcription regulation in prokaryotic cells

Chensi Miao\textsuperscript{a,b,1}, Huiwei Zhao\textsuperscript{b,1}, Long Qian\textsuperscript{c,*,1}, Chunbo Lou\textsuperscript{a,d,*}

\textsuperscript{a} CAS Key Laboratory of Microbial Physiological and Metabolic Engineering, State Key Laboratory of Microbial Resources, Institute of Microbiology Chinese Academy of Sciences, Beijing, 100101, China
\textsuperscript{b} College of Life Science, University of Science and Technology of China, Hefei, 230027, China
\textsuperscript{c} Center for Quantitative Biology, Peking University, Beijing, 100871, China
\textsuperscript{d} University of Chinese Academy of Science, Beijing, 100149, China

\textbf{ABSTRACT}

With a unique crRNA processing capability, the CRISPR associated Cpf1 protein holds great potential for multiplex gene regulation. Unlike the well-studied Cas9 protein, however, conversion of Cpf1 to a transcription regulator and its related properties have not been systematically explored yet. In this study, we investigated the mutation schemes and crRNA requirements for the DNase deactivated Cpf1 (dCpf1). By shortening the direct repeat sequence, we obtained genetically stable crRNA co-transcripts and improved gene repression with multiplex targeting. A screen of diversity-enriched PAM library was designed to investigate the PAM-dependency of gene regulation by dCpf1 from Francisella novicida and Lachnospiraceae bacterium. We found novel PAM patterns that elicited strong or medium gene repressions. Using a computational algorithm, we predicted regulatory outputs for all possible PAM sequences, which spanned a large dynamic range that could be leveraged for regulatory purposes. These newly identified features will facilitate the efficient design of CRISPR-dCpf1 based systems for tunable multiplex gene regulation.

1. Introduction

Ever since the discovery of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) mechanism, its DNA-targeting strategy has been extensively characterized and masterfully adapted to a biotechnological tool for sequence-specific DNA manipulation that has rapidly revolutionized the fields of genome editing and engineering [1–4]. This simple yet elegant system consists of the Cas9 endonuclease from Streptococcus pyogenes and a guide RNA (gRNA) that directs Cas9 to the complementary DNA target in the presence of a protospacer adjacent motif (PAM) [2,4]. The programmability, achieved through the guide sequence, has been further leveraged in variants of the system utilizing the engineered nuclease-deactivated Cas9 (dCas9) on its own or linked to diverse effector protein domains [5]. These dCas9-based CRISPR toolkits have proven extremely powerful for the systematic perturbation of single genes in regulatory and metabolic networks, advancing our knowledge in synthetic and systems biology at an unprecedented speed [6,7].

To push forward the CRISPR technology to the systems level, the ability to simultaneously manipulate multiple genes is highly demanded. Multiplex gene targeting, ideally through co-expressing multiple gRNAs in the same cell, enables the interrogation of much more complex interactions in genome-scale networks [8,9], as well as the combinatorial optimization of large heterologous pathways for metabolic engineering [5,8–12]. However, expressing gRNAs from independent plasmids suffers from a scalability issue, while encoding multiple gRNAs on the same expression cassette requires subsequent co-transcript processing, which relies on either endogenous RNase III activity, or in many systems, the introduction of sequence specific RNA endonuclease such as Csy4, the self-cleaving ribozyme sequences, or tRNA sequences that invoke the tRNA processing machinery [13,14]. For the purpose of application, these solutions either impose some level of cytotoxicity, or require lengthy additions to the gRNA sequence, causing greater genetic instability on a repeat-laden structure. This conundrum may now be solved thanks to the discovery of Cpf1, a Class II CRISPR endonuclease of Type V-A, which displays endoribonuclease activity and was shown to process CRISPR RNA (crRNA) co-transcripts into independent mature crRNAs, in addition to its DNA cleavage

https://doi.org/10.1016/j.synbio.2018.11.002

Received 18 August 2018; Received in revised form 6 October 2018; Accepted 6 November 2018

© 2018 Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
activity [15–17]. Besides functional duality, the Cpf1 system displays some enticing features – a concise crRNA, ~40nt in its natural form, is more compatible with current DNA oligomer synthesis techniques and more resistant to homologous recombination-deactivated cassette disruption in a co-transcript context, and a thymine-rich PAM preference extends the targetable regions especially in AT-rich genomes. We thus believe in the great potential of a DNase deactivated Cpf1 (dCpf1) as an efficient tool for multiplex gene regulation.

Although aspects of the CRISPR-Cpf1 system as DNA endonuclease has been characterized, there have been only first attempts in using CRISPR-dCpf1 as transcriptional regulators. These studies proved its applicability in bacterial, plant, and mammalian cells [18–21]. To harness and streamline the system for multiplex gene regulation, three specific aspects need addressing or systematic characterization: 1. a mutational scheme that abolishes Cpf1’s DNA activity and yet minimally affects its DNA binding and RNase activities; 2. the requirements for pre-crRNA that contains multiple direct repeat-guide sequence units for efficient crRNA processing and DNA targeting [15,22]; and 3. the dependence of DNA binding strength on the PAM sequence [23–26].

In this study, we designed a negative reporter assay for transcriptional repression by the CRISPR-dCpf1 system in Escherichia coli. The reporter assay was used to systematically quantify the functional effects of dCpf1 mutations and crRNA variants. We evaluated the dependence of gene repression on crRNA processing, lengths of direct repeats and guide sequences, as well as the number of target sequences randomly located within the target gene. Most importantly, we investigated the PAM sequence preference for dCpf1 from Francisella novicida and Lachnospiraceae bacterium in a randomized 6nt PAM library. We found a broad range of repression strengths that did not conform to the previously identified PAM preferences. Therefore, we built an interpolation algorithm to predict gene repression activity for any PAM sequence based on a much limited number of sampled weak and strong PAMs. Without assuming context independency, the algorithm generated reliable estimates of PAM strengths, which could in principle lend great controllability to the CRISPR-dCpf1 system in synthetic biological applications.

2. Materials and methods

2.1. Strains and media

The E. coli DH5α was used as the host strain for all experiments. Luria-Bertani (LB) media (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) was used as the growth media. Cells for flow cytometric fluorescence analysis were cultured in M9 media (12.8 g/L Na2HPO4·7H2O, 3 g/L KH2PO4, 0.5 g/L NaCl, 1.67 g/L NH4Cl, 1 mM thiamine hydrochloride, 0.4% glucose, 0.2% casamino acids, 2 mM MgSO4, 0.1 mM CaCl2). Ampicillin, Kanamycin and Chloramphenicol concentrations for all experiments were 100 μg/ml, 50 μg/ml and 20 μg/ml, respectively.

2.2. Plasmid construction

The FncPf1 gene were synthesized by Genscript Inc. Then it was mutated into dFncPf1 and inserted into a vector containing a pTact inducible promoter, an ampicillin-selectable marker, and a p15A replication origin. The crRNA plasmid backbone contained a synthetic constitutive promoter (J23119), a chloramphenicol-selectable marker, and a ColE1 replication origin. Various guide sequences were inserted by the Golden Gate method. The reporter plasmid contained sf-gfp as the reporter gene under the control of a synthetic constitutive promoter (J23100), a KanR-selectable marker, and a pSC101 replication origin. The crRNA sequences used in this study was summarized in Tables S3–S5.

2.3. Flow cytometry and analysis

Overnight culture of E. coli DH5α containing test plasmids was diluted 196 times into M9 medium with corresponding antibiotics, followed by shaking at 37 °C for 3 h. Cells were then serially diluted 1000 times into M9 medium with antibiotics and appropriate concentrations of IPTG cultured at 37 °C. The levels of fluorescence protein were analyzed by BD™ LSR II flow cytometer (Becton Dickinson, San Jose, CA, USA) with appropriate voltage settings (FSC:440, SSC:260, FITC:480) after further dilution into PBS with 20 μg/ml Kanamycin. Each sample was collected at least 50,000 events. The mean fluorescence of each sample was calculated with Flowjo software (Treestar, Inc., San Carlos, CA, USA) and analyzed with GraphPad Prism software (GraphPad Software, La Jolla, CA, USA).

2.4. PAM screen and analysis

Randomized PAM library was constructed by reverse PCR and Gibson ligation, using Random_F/Random_R consisting of six randomized nucleotides as primers and plasmid P_R_PAM as the backbone (Fig. S3). The PAM plasmid library was then transformed into competent E. coli DH5α harboring dFncPf1 and crRNA plasmids. After transformation, cells were plated on LB agar supplemented with antibiotics of ampicillin, chloramphenicol and Kanamycin. After ~16 h of growth, >107 cells were collected and pooled, diluted into fresh LB medium with antibiotics, and cultured overnight (~16 h). The overnight culture was diluted ∼500 times into M9 medium with required antibiotics and appropriate concentrations of IPTG, followed by shaking at 37 °C for 3 h. Cultures were then diluted into PBS buffer to sort the cells with lowered fluorescence on a BD Influx Cell Sorter (Becton Dickinson, San Jose, CA, USA). From the sorted cells, random samples were collected, diluted and coated, and the remaining cells were cultured for the next round of sorting. After three rounds of sorting, colonies on the coated plates from all rounds were picked and subject to fluorescence measurements by flow cytometry and Sanger sequencing for their respective PAM sequences (Fig. S4).

2.5. PAM strength prediction and algorithm evaluation

The computation algorithm used to predict PAM strength was explained in detail in Supplementary Information. The code was written in Matlab®. Cross-validation was done by randomly selecting samples from measured mean values to generate training sets. Testing was done on measured mean values for unselected PAMs (testing sets). For original selection, samples were selected randomly from the original data set. For uniform selection, samples were selected with equal numbers from equally placed bins in the entire fluorescence range of the original data set. At each training-testing set splitting ratio, 100 independent runs were conducted. Sequence logos in Fig. 6A and Fig. S7A were generated on http://weblogo.berkeley.edu/logo.cgi.

3. Results

3.1. Single mutation dCpf1 elicits stronger gene repression than double mutation dCpf1

A previous study identified key amino acids in the RuvC-like domain of Cpf1 and proposed a double mutation scheme (D917A and E1006A) for deactivating the DNase activity of Cpf1 from F. novicida (FnCpf1), in much the same way as the design of dCas9 [25]. However, unlike Cas9, single mutations of either amino acid in Cpf1 was able to abolish cleavage of both DNA strands [18–21]. As Cpf1 has a more complex domain structure than Cas9 [15,16], we suspected that double mutations may interfere with the RNA processing and DNA binding abilities of Cpf1 and thereby affect its regulatory activity. Therefore, we constructed single mutation forms of Cpf1 from F. novicida (dFncPf1) and
3.2. Minimal crRNA length requirements for dCpf1’s regulatory activity

A unique function of Cpf1 is crRNA processing, where pre-crRNA containing multiple units of a 36nt direct repeat (DR) followed by gRNA is cleaved and truncated to mature crRNAs of a 19nt DR-gRNA structure containing multiple units of a 36nt direct repeat (DR) following pre-crRNA processing by Cas6-family of endoribonucleases is a prerequisite for the subsequent assembling of a functional Cas complex on crRNA [27]. To find out if crRNA processing is essential for the gene regulatory function of dCpf1, we expressed crRNAs of various DR lengths ranging from 16nt to 36nt in the reporter system (Fig. 2A). All crRNAs with DR length > 19nt showed the same repression activity as the crRNA with DR length of exactly 19nt (Fig. 2B&C). Since the latter did not undergo processing, we concluded that dCpf1 can load onto mature crRNA in the absence of extra processing signals, and thus its regulatory activity is independent of its crRNA processing activity. A previous in vitro experiment showed for Cpf1, crRNA with DR lengths of 16–18nt were still able to induce target DNA cleavage [16]. We found, however, no regulatory activity of dCpf1 with crRNAs having shorter than 19nt DRs (Fig. 2B&C).

Another functional element in crRNA is the guide sequence whose length is believed to be a crucial parameter for the DNA cleaving efficiency of the Cpf1 nuclease. Cpf1 generates mature crRNAs with guide sequence of typically 24nt long. We examined how the extension and truncation of the guide sequence affect the regulatory efficiency of dCpf1 by constructing a number of guide sequences with lengths from 16nt to 31nt (Fig. 2D). The results showed a guide sequence ≥ 16nt was able to elicit 200-fold gene repression. Repression was drastically weakened with further guide sequence truncation (Fig. 2E&F). For Cpf1, previous study suggested a threshold guide sequence length of 18nt below which DNA targeting and cleavage was not observed [16]. These results together suggested a 16-18nt minimal guide sequence length required for DNA targeting, depending possibly on the specific guide sequence used.

3.3. Enhanced gene repression through multiplex targeting of dCpf1

As the targeting of multiple genes has been demonstrated in several recent studies [18,19,21], and a single bound dCpf1, without dedicated inactivation domains, was not sufficient in suppressing gene expression in human HEK293T cells [21], we studied gene repression by tandemly positioned dCpf1 roadblocks. 24nt guide sequences targeting three independent segments within the coding region of the sf-gfp gene were
Fig. 2. The effect of repeat and guide sequence lengths on gene repression by dFnCpf1-crRNA. (A) Aligned repeat sequences of different lengths used in crRNAs. The dashed line indicates the cleavage site on crRNA during crRNA processing by dFnCpf1. Red colored sequences remain in the mature crRNA, while the rest of the sequences are cleaved off. (B) Gene repression curves of dFnCpf1 with truncated repeat sequences. (C) Maximal repression folds of dFnCpf1 with the same set of truncated repeat sequence as in (B). (D) Aligned guide sequences of different lengths used in crRNAs. (E) The repression curves for different lengths of guide sequences in the dFnCpf1-crRNA system. (F) Maximal repression folds of dFnCpf1 with the same set of truncated guide sequences as in (E). Error bars represent standard deviation of fluorescence for three independent experiments on different days. Positive and negative controls are the same as in Fig. 1.

Fig. 3. Repression by dFnCpf1 with co-transcribed crRNAs. (A) Schematic representation of target sequences for each single crRNA, as well as the design of individual and combined multiple crRNAs. (B) Different lengths of repeat sequence in the triply-combined crRNA co-transcript. (C) Repression curves of dFnCpf1 with single or multiple crRNAs. (D) Repression curves of dFnCpf1 with varied repeat sequence lengths in the same triply-combined crRNA co-transcript. Error bars represent standard deviation of fluorescence for three independent experiments on different days. Positive and negative controls are the same as in Fig. 1. For crRNA sequences see Table S3.
connected by the 36nt DR sequences and co-expressed under a constitutive promoter (Fig. 3A). We found that crRNAs targeting any one of the three segments resulted in varied but significant gene repression (10–100-fold). Repression was further augmented by doubly or triply combined crRNAs, presumably through a stronger blockage of transcription elongation (Fig. 3C). Strikingly, the triply combined crRNAs completely abolished gfp expression (> 300-fold reduction). The fold reduction by multiplex targeting, relative to individual targeting, was between additive and multiplicative. These results suggested that co-transcribed crRNAs targeting multiple DNA segments can be utilized by dCpf1 to combinatorially augment gene repression.

Co-transcription of multiple crRNAs ensures uniform expression among all gRNAs, and reduces the genetic instability associated with repeated expression cassettes. Yet, in the crRNA coding region, a repeat structure conferred by the DR sequences could also lead to genetic instability through an increased chance of homologous recombination as the length of DR increases [28]. We further optimized the system by truncating the interspersed DR sequences, and identified the minimal DR length essential for multiple DNA targeting (Fig. 3B). In consistence with the condition for single crRNAs, we found a 19nt DR is required for dCpf1-mediated multiplex repression (Fig. 3D).

3.4. dFnCpf1’s regulatory activity strongly depends on the PAM sequence

Previous studies have shown a strong dependence of CRISPR activity on the PAM sequence. For FnCpf1, CTN and TTN were identified as the preferred PAM sequences for DNA cleavage [16]. We selected two sets of targets on both the template and non-template strands of the sfgp gene based on these motifs, and tested the gene repression activity of dFnCpf1 (Fig. S1A). We observed that none of the non-template strand targets generated significant repression (Fig. S1B) – a strain bias also reported in other studies [21] – while the template strand targets showed a broad range of repression strengths (Fig. S1C). Unlike the case for dCas9 (Fig. S2A), for dCpf1, repression strengths were not correlated with the targets’ locations within the coding region (Fig. S2B), suggesting factors other than transcript length significantly influenced dCpf1’s regulatory activities. We further selected three sets of targets, each containing three targets starting from a T-rich region, but shifted by 1- or 2-nt relative to each other. Targets selected this way had similar distances from the transcription start site (TSS) and similar base/subword compositions, and all had TTN as the PAM sequence. However, within each set, repression activities were still drastically different (Fig. 4). These results were strongly indicative of TTN as a semi-smooth regulatory strength landscape in the PAM sequence space, in other words, the regulatory strength of a PAM sequence of length k is computed as the average strengths of all PAMs that are different by one nucleotide at only one of the k locations, except at non-degenerate locations (see below). Strength information at location i (i = 1 ... k) is weighted by the degeneracy of the location. A non-degenerate location is a location where variations in base identity have exhibited very different regulatory outputs in the sample set, and thus all information at this location is discarded for predictive purposes. Whenever possible, context dependency is considered in evaluating location degeneracy. A detailed explanation of the algorithm can be found in Supplementary Information. Unlike the conventional PWM model or sequence logo methods, the algorithm does not assume positional independence between bases, and therefore, it automatically captures all sequence patterns and features contained by the sample pool.

We predicted PAM strengths for all 6nt words based on data from 200 samples for dFnCpf1 and 133 samples for dLbCpf1 (Fig. 6A and Fig. S7A & Table S2). Conversely, we used the predicted values for unmeasured words to back-predict strengths of measured PAMs. This yielded a > 0.99 correlation with measured values, indicating a minimal loss of information through the course of interpolation (data not shown). The results indicated that in general, for both dFnCpf1 and dLbCpf1, positions 1 and 2 did not had significant effects on PAM strengths. For dFnCpf1, PAM strength was most sensitive to the 4- and 5-th location, while position 3 contributed to PAM strength diversity more than position 6. For dLbCpf1, positions 3–6 all affected PAM strength strongly (Fig. 6B and Fig. S7B). When ranking samples based on repression activity, we found that for dFnCpf1, the strongest PAMs were (T)TTTVT and (T)TTIV, whereas T was strongly disfavored at the last position. The other previously identified CTN motif generated only moderate repression activities (Fig. 6A). For dLbCpf1, the strongest repessions were elicited by (T) TTTV PAMs, followed by CTTV. Like dFnCpf1, there was a strong preference against T at the last position in both strong and moderate PAMs. However, for dLbCpf1, TTTV was able to induce medium repression, with a 5’- T further enhancing its activity (Figs. S5A and S7A).

To evaluate the predictive power of our algorithm, cross-validation was done by splitting the sample pool for dFnCpf1 into training and testing sets, at proportions from 20% to 90% (for the training set).
When using 50% randomly selected samples as the training set, the predictions for the testing set were > 0.90 correlated with measured values, and back-prediction showed > 0.95 correlations with data in the training sets (Fig. 6C). Even with only 20% (n = 40) of the sampled PAMs as the training data, a correlation > 0.8 could be obtained with the testing set (Fig. 6D). These numbers decreased mildly for dLbCpf1, whose sample pool were smaller and less biased toward high and low repression ranges (Figs. S7C and D). When we applied a strictly uniform selection method in the low, medium, and high repression ranges, irrespective of the repression strength distribution of the original sample pools, correlation between predictions and the testing sets were around 0.75 (at 53% data as training set) for dFnCpf1 and 0.4 (at 14% data as training set) for dLbCpf1 (Fig. S8). These results underscored the importance of PAM sequences sampled at high and low repression ranges in generating sufficient information for the algorithm to successfully interpolate for any other PAM sequence. Successive shrinkage of the training set suggested a threshold sample size of n = 40 (~1% of the sequence space), below which > 50% of the prediction attempts ended up with unpredictable sequences that were not covered by available information (Fig. S9).

For the dFnCpf1 dataset, we also performed LASSO regression with a simple linear model assuming position independence (i.e. having $4 \times 6 = 24$ independent variables, see Supplementary Information). The best performing model came back with a (T)TT(V) preference which captures the PWM motif for dFnCpf1, but missed the fine features at positions 1-3 (Fig. S10B). The linear model had a back-prediction correlation of ~0.86 (Fig. S10A). Cross-validation using randomly sampled training sets, as described above, generated models that predicted testing sets with 0.8–0.9 correlations (Fig. S10D). The inferiority compared to our algorithm was presumably the result of interdependency between positions which, upon close examination, had a greater impact on medium and low strength PAMs than on strong PAMs (Fig. S10A&C).

4. Discussion

In this article, we systematically investigated the key constraints and properties of the CRISPR-dCpf1 system as transcriptional repressors in E. coli cells. In comparison to the dCas9 based CRISPR systems, dCpf1 offers the unique potential of multiplex gene regulation with its ability to autonomously process crRNA co-transcripts and subsequently target multiple independent DNA sequences. This ability minimizes the uncertainty in crRNA relative dosages and genetic stabilities, as previously seen in systems with dCas9 and independently transcribed crRNAs. This
is key to large scale standardized perturbation experiments such as whole transcription network engineering. There have recently been multiple reports on dCpf1’s gene regulation applications in bacteria, plants, and human cells. Although repression in bacteria was attained, repression in Arabidopsis plants, and human cells. Although repression in bacteria was attained, repression in Arabidopsis and activation in human HEK293T cells was quantitatively unstable and somewhat idiosyncratic [19–21]. A systematic characterization of the CRISPR-dCpf1 system with respect to its DNA binding properties is obviously in need to further enhance performance in these experimental systems.

We compared the repression activities of dCpf1 mutant forms including single and double mutations at the two previously identified catalytic residues for Cpf1’s DNase activity. For both dFnCpf1 and dLbCpf1, double mutations compromised regulatory activities. Between the two single mutation variants, D917A/D832A generated consistently strong regulatory activity, whereas E1006A in dFnCpf1 was much less efficient in DNA binding than D917A. While it is possible that the single mutation variants exhibit residual DNase activities that went undetected in our growth rate measurements, based on the crystal structure of dCpf1 in complex with crRNA and DNA, we speculate that the subdued repression may reflect a genuine destabilization of dCpf1-crRNA-DNA complex, as E1006 in the RuvC-II domain of FnCpf1 is spatially close to the WED domain and the bridge helix that interact closely with the 5′ crRNA handle [30,31].

We found that for dCpf1, crRNA cleavage was not essential for subsequent DNA targeting. The wild type crRNAs adopt a 19nt DR-24nt gRNA form. In our studies, the minimal length requirements were 19nt for the direct repeat and 16nt for the guide sequence. A previous biochemical study revealed the importance of a 5′-AAU-3′ sequence at the −19 location of the processed crRNA [15]. This tri-base region may functionally be crucial for RNA processing as well as stabilizing the dCpf1-crRNA complex. With shorter crRNAs, Cpf1 may still form transient complexes with DNA and produce strand breaks in vitro [16]. However, tight binding of dCpf1-crRNA to the DNA target demanded an intact 19nt direct repeat sequence according to our results.

We further demonstrated enhanced gene repression by co-transcribed crRNAs targeting DNA sequences located in tandem in the coding region. Again, a ≥19nt DR length is required in the crRNA co-transcript for crRNA processing, which sets a lower limit of 35–40nt repetitive crRNA structure for the precursor crRNA expression cassette. We found the PAM sequence to be a major factor determining gene repression activity. The previously identified TTN and CTN motifs for FnCpf1 in DNA cleavage assays did not explain the PAM preference in terms of gene regulation by dFnCpf1. Although a T-rich PAM for Cpf1 greatly expands the genomic regions that could be targeted for cleavage, gene regulatory response was sensitively dependent on the exact PAM sequences used. On the same target sequence, a wide range of repression folds were observed when different 6nt preceding sequences were used. We designed a negative reporter screen to identify PAM sequences eliciting strong, medium and weak repressions. We further developed an interpolation algorithm based on context-dependent sequence similarities, using which, we predicted regulatory strengths for all 6nt sequences as PAMs based on measurements of 200 and 133 PAMs for dFnCpf1 and dLbCpf1, respectively. Compared to motif analysis by next-generation sequencing, the algorithm provides a fast and economic way of assessing PAM preferences, and is especially suited for revealing moderate and weak PAMs, which might be masked by biases produced through DNA amplification. The algorithm also showed superiority over context independent linear models, revealing the significance of higher order PAM features in Cpf1-target recognition.

Our analyses suggested for both enzymes a general 4nt core sequence dependence, with T strongly disfavored in the last position, and slightly favored at the proceeding 2nt positions. Specifically, dFnCpf1 and dLbCpf1 both displayed a preference for TTTV PAMs; while for dLbCpf1, other PAMs also emerged as mediating strong regulatory
TTTV was previously identified for LbCpf1 [23], and recently identified in a study on genome editing by FnCpf1 in Baker’s yeast [32] while we were preparing this manuscript. This suggests that differences in the strengths of extended PAMs may also be relevant when cleavage is concerned, especially for improving CRISPR DNases that did not function well in certain systems. In Ref. [32], the authors found that targets with TTTA and (CT) TTTC PAMs did not lead to genome editing, despite conforming to the TTTV motif. Our data suggest a range of 50–300 fluorescence for NNTTTA PAMs and a ∼110 fluorescence for CTTTTC. Although these are all strong repressions in the 6nt library, the six-fold difference might still significantly affect reaction outcome.

For Cas9-based CRISPR applications, the most common strategies for tuning activity include coding/non-coding strand targeting, target distance from the TSS, protein concentration, and gRNA-target sequence complementarity. While the first two do not apply to Cpf1-based systems, our results mapped out a quantitative relationship between the PAM sequence and dCpf1’s regulatory activity. Modulation in cis, such as by the PAM sequence or by target complementarity, allows for the orthogonal regulation of multiple targets at a single dCpf1 induction level. This would grant much flexibility for quantitative assessment of complex transcription networks. Moreover, the screening method we developed could be utilized to introduce a control element in arbitrary genes. Compared to targets in the upstream promoter regions, insertions within the 5’UTR region followed by an insulator could minimize the interference on background gene expression levels. Compared to targets in the coding sequences, PAM sequences and target sequences in the inserted fragment can be designed separately to achieve desired repression outputs with high specificity. Besides dFnCpf1 and dLbCpf1, dCpf1s from Acidaminococcus sp. and Eubacterium eligens have also been tested in bacterial and eukaryotic cells [18–21]. Our screening and prediction methods could serve a pipeline for rapid characterization of the natural diversity of dCpf1 proteins. When coupled with technologies already developed for dCas9 [33,34], dCpf1 may be transformed into powerful tools for sophisticated applications of multiplex gene interrogation.

**Funding**

National Natural Science Foundation of China (No. 31470818 and 31722002); the 973 projects of Ministry of Science and Technology of China (No. 2015CB910300); the Key Research Program of the Chinese Academy of Sciences (No. QYZDB-SSW-SMC050); and the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB29040000).

**Acknowledgements**

We thank Lili Ji (Peking University), Tingting Li (Peking University) and Junying Jia (Institute of Biophysics, CAS) for technical assistance of flow cytometry.

**Appendix A. Supplementary data**

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2018.11.002.

**References**


