A Cas9 with PAM recognition for adenine dinucleotides

Pranam Chatterjee
Massachusetts Institute of Technology

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/oapubs

Part of the Bioinformatics Commons, Computational Biology Commons, Enzymes and Coenzymes Commons, Genetic Phenomena Commons, and the Nucleic Acids, Nucleotides, and Nucleosides Commons

Repository Citation
Chatterjee P, Lee J, Nip L, Koseki SR, Tysinger E, Sontheimer EJ, Jacobson JM, Jakimo N. (2020). A Cas9 with PAM recognition for adenine dinucleotides. Open Access Publications by UMass Chan Authors. https://doi.org/10.1038/s41467-020-16117-8. Retrieved from https://escholarship.umassmed.edu/oapubs/4255

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License. This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Open Access Publications by UMass Chan Authors by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
A Cas9 with PAM recognition for adenine dinucleotides

Pranam Chatterjee1,2✉, Jooyoung Lee3, Lisa Nip1,2, Sabrina R. T. Koseki1,2, Emma Tysinger1,2, Erik J. Sontheimer3, Joseph M. Jacobson1,2 & Noah Jakimo1,2

CRISPR-associated (Cas) DNA-endonucleases are remarkably effective tools for genome engineering, but have limited target ranges due to their protospacer adjacent motif (PAM) requirements. We demonstrate a critical expansion of the targetable sequence space for a type II-A CRISPR-associated enzyme through identification of the natural 5′-NAAN-3′ PAM preference of Streptococcus macacae Cas9 (SmacCas9). To achieve efficient editing activity, we graft the PAM-interacting domain of SmacCas9 to its well-established ortholog from Streptococcus pyogenes (SpyCas9), and further engineer an increased efficiency variant (iSpyMac) for robust genome editing activity. We establish that our hybrids can target all adenine dinucleotide PAM sequences and possess robust and accurate editing capabilities in human cells.
biotechnologies based on RNA-guided CRISPR systems have enabled precise and programmable genomic interfacing. However, CRISPR-associated (Cas) endonucleases are also collectively restrained from localizing to any position along double-stranded DNA (dsDNA) due to their requirement for targets to neighbor a protospacer adjacent motif (PAM). Current gaps in the PAM sequences that Cas enzymes are known to recognize prevent access to numerous genomic positions for powerful genome editing activities, such as base editing, prime editing, and homology-directed repair. Many adenine-thymine (AT)-rich regions, in particular, have been excluded from compelling CRISPR applications because previously reported endonucleases, such as Cas9 and Cas12a (formerly known as Cpf1), require targets to neighbor guanine–cytosine (GC)-content or more restrictive motifs, respectively.

In this work, we introduce an ortholog of the well-established Cas9 from Streptococcus pyogenes (SpyCas9), derived from Streptococcus macacae NCTC 11558, that can instead recognize a short 5′-NAA-3′ PAM. These sequences constitute 18.6% of the human genome, making adjacent adenines the most abundant dinucleotide (Supplementary Fig. 1A–B). The importance of this alternative PAM recognition for a Cas9 enzyme is reinforced by recent work exposing that Type-V DNA-targeting CRISPR nucleases (including Cas12 and Cas14 orthologs), while targeting dsDNA at AT-rich PAM sites with intrinsic high fidelity, will indiscriminately digest single-stranded DNA (ssDNA) once bound to their targets. Similar collateral activity may introduce unwanted risks around partially unpaired chromosomal structures, such as transcription bubbles, R-loops, and replication forks.

Here we present engineered nucleases derived from SmacCas9 and characterize their altered specificity and utility by means of transcriptional repression in bacterial culture, in vitro digestion reactions, and gene editing activity in human cells. Our results demonstrate complete 5′-NAA-3′ PAM recognition of our engineered variants in all tested contexts.

**Results**

**Discovery of SmacCas9.** To modify the ancestral 5′-NGG-3′ PAM specificity of SpyCas9, early and recent reports have employed directed evolution (e.g., "VQR", "EQR", "VRER", and "NRNH" variants) or rational design informed by crystal structure (e.g., "QQR", "NG", and "NR" variants). These reports focused on the PAM-contacting arginine residues R1333 and R1335 that abolish function when exclusively mutated. While those studies identified compensatory mutations resulting in altered PAM specificity, the Cas9 variants that they produced maintained a guanine preference at least one position of the PAM sequence for reported in vivo editing. Concurrent reports have used evolutionary information to further relax the canonical 5′-NGRRT-3′ PAM specificity of Staphylococcus aureus Cas9 (SaCas9) or to discover alternative 5′-NNNCC-3′ PAM specificity to the canonical 5′-NNNHTT-3′ PAM of Neisseria meningitidis Cas224. The nuclease from both of these new reports, however, still prefer GC-content in at least one position of the PAM sequence. We aimed to lift such GC-content prerequisites via a custom bioinformatics-driven workflow that mines existing PAM diversity in the Streptococcus genus. Using this strategy, we homed in on SmacCas9 as having the potential to bear altered non-GC PAM specificity upon aligning 115 orthologs of SpyCas9 from UniProt (limited to those with greater than a 70% pairwise BLOSUM62 score). From the alignment we found SmacCas9 was one of two close homologs, along with a Streptococcus mutans B112SM-A Cas9 (SmutCas9), possessing glutamines at both of the positions aligned to the otherwise highly conserved PAM-contacting arginines (Fig. 1a–b; Supplementary Fig. 2A).

Arginine residues are known to strongly prefer guanines in the amino-acid-base interaction landscape, as evidenced by the 5′-NGG-3′ specificity of SpyCas9. Glutamine residues, on the other hand, preferentially bind to adenines, through interaction with the major groove edge. We thus hypothesized that SmacCas9 had naturally coevolved the necessary compensatory mutations to gain new adenine-rich PAM recognition. A small sample size of 13 spacers from its corresponding genome’s CRISPR array prevented us from confidently inferring the SmacCas9 PAM in silico. Nevertheless, the possibility for SmacCas9 requiring less GC-content in its PAM was supported by sequence similarities to the "QQR" variant that has 5′-NAA-3′ specificity, in addition to the AT-rich putative consensus PAM for phage-originating spacers in CRISPR arrays associated with highly homologous SmutCas9, which were identified with the aid of our previously-described SPAMALOT pipeline and consistent with previous predictions (Fig. 1c; Supplementary Fig. 2B; Supplementary Fig. 3) [22,23].

**Engineering and PAM characterization of SpyMac.** We proceeded to empirically determine the PAM preferences of several Streptococcus orthologs that change one or both of the critical PAM-contacts. Based on demonstrated examples of the PAM-interaction domain (PID) and guide RNA (gRNA) having cross-compatibility between Cas9 orthologs that are closely related and active, we constructed new variants by rationally exchanging the PI region of catalytically “dead” SpyCas9 (dSpyCas9) with those of the selected orthologs (Supplementary Fig. 2A–B). Assembled variants, including dSpyMacCas9 (herein referred to as dSpyMac), were separately cotransformed into E. coli cells, along with guide RNA derived from S. pyogenes and an 8-mer PAM library of uniform base representation in the PAM-SCANR genetic circuit, established by Leenay et al. The circuit upregulates a green fluorescent protein (GFP) reporter in proportion to PAM-binding strength. Therefore, we collected the GFP-positive cell populations by flow cytometry (Supplementary Fig. 4) and Sanger sequenced them around the site of the PAM to determine position-wise base preferences in a corresponding variant’s PAM recognition. dSpyMac, more so than dSpyMutCas9, generated a trace profile that was most consistent with guanine-independent PAM recognition, along with a dominant specificity for adenine dinucleotides (Fig. 2a; Supplementary Fig. 2C).

Next, we purified nuclease-active enzymes to continue probing the DNA target recognition potential and uniqueness of SpyMac. (Supplementary Fig. 5A)27,32. We individually incubated the ribonucleoprotein complex enzymes (composed of Cas9 + crRNA + tracrRNA) with double-stranded target substrates of all 5′-3′-neighboring base combinations at an adenine dinucleotide PAM (5′-NAA-3′; Fig. 2b). A brief 16-min digestion indicated both wild-type SmacCas9 and the hybrid SpyMac cleaved adjacent to 5′-NAA-3′ motifs more broadly and evenly than the previously reported QQR variant. SpyMac distinguished itself further with rapid DNA-cutting rates that resemble the fast digest kinetics of SpyCas9 (Fig. 2c–d).33. We ran reactions that used varying crRNA spacer lengths and tracrRNA sequence, as the latter differs slightly between the S. macacae and S. pyogenes genomes (Supplementary Fig. 5B–E). Neither of these two parameters compensated for the slower cleavage rate of SmacCas9, but we did notice marginal improvement in the activity of the wild-type form with its native tracrRNA, which comports with the interface of the guide-Cas9 interaction being mostly outside of the PI domain.

To verify that an adenine dsDNA dinucleotide is sufficient for Cas9 PAM recognition and target cleavage, we assembled target sequences that switch the next four downstream bases to the same nucleotide (e.g., 5′-TAAGXXXX-3′, for X all fixed to A, C, G, or
**Fig. 1** Identification of features from natural PAM divergence through bioinformatics. a Sequence alignment of SpyCas9, its QQR variant, and SmacCas9. The step in the underlining red line marks the joining of SpyCas9 and SmacCas9 to construct a SpyMac hybrid. The sequence logo (Weblogo online tool) immediately below the alignment depicts the conservation at 11 positions around the PAM-contacting arginines of SpyCas9. b The domain organization of SpyCas9 juxtaposed over a color-coded structure of RNA-guided, target-bound SpyCas9 (PDB ID 5F9R). The two DNA strands are black with the exception of a magenta segment corresponding to the PAM. A blue–green–red color map is used for labeling the Cas9 PI domain and guide spacer sequence to highlight structures that confer sequence specificity and the prevalence of intra-domain contacts within the PI43. c A sequence logo generated online (WebLogo) that was input with putative PAM sequences found in Streptococcus phage and associated with close SmacCas9 homologs.

**Fig. 2** Validation of SmacCas9 recognition for adenine dinucleotide PAM sequences. a Chromatograms representing the PAM-SCANR based enrichment of variant-recognizing PAM sequences from a 5'-NNNNNNNNN-3' library. b SYBR-stained agarose gels showing in vitro digestion of 10 nM 5'-NAAN-3' substrates upon 16 minutes of incubation with 100 nM of purified ribonucleoprotein enzyme assemblies. Arrows distinguish banding of the cleaved products from uncleaved substrate (top band). Matrix plots summarize cleaved fraction calculations, which were carried out in a custom script for processing gel images. Samples were performed in independent biological duplicates (n = 2). c Time course measurements of target DNA substrate cleavage for SmacCas9 and SpyMac. d DNA substrate cleavage plotted as a function of 0.25:1, 1:1, and 4:1 molar ratios of ribonucleoprotein to target for wild-type SpyCas9 and hybrid SpyMac. Source data are provided as a Source Data file.
T; Supplementary Fig. 5F). We confirmed SpyMac remains active across this target set, albeit with some variation in cutting rate. Additionally, we observed moderate yield of cleaved products on examples of 5′-NBBAA-3′, 5′-NABAB-3′, 5′-NBABA-3′ PAM sequences (where B is C, G, or T; Supplementary Fig. 5G), revealing an even broader tolerance for increments to the dinucleotide position or adenine adjacency. We anticipate future measurements of guide-loading, target-dissociation and R-loop expansion/contraction will provide more insights on the serendipitous catalytic benefit over SmacCas9 from grafting its PI domain onto a truncated SpyCas9.

**Methods**

**Selection of Streptococcus Cas9 orthologs of interest.** All Cas9 orthologs from the Streptococcus genus were downloaded from the online UniProt database [https://www.uniprot.org/](https://www.uniprot.org/). These were the down-selected by pairwise alignment to SpyCas9 using a BLOSUM62 cost matrix in Genewiz software, discarding orthologs with less than 70% agreement with the Spy Cas9 sequence. The remaining 115 orthologs were used to generate a sequence logo (Weblogo http://weblogo.berkeley.edu/logo.cgi), and were manually selected for divergence at positions aligned to residues critical for the PAM interaction of SpyCas9. The SPAMALOT pipeline was implemented as we previously reported. Briefly, a set of scripts based around the Bowtie alignment tool (http://bowtie-bio.sourceforge.net) map the spacer sequences from CRISPR cassettes to putative targets in phage genomes. The SPAMALOT software can be downloaded at https://github.com/mitmedialab/SPAMALOT.

**PAM-SCANr bacterial fluorescence assay.** Sequences encoding the PAM-interaction domains of selected Cas9 orthologs were synthesized as gBlock fragments by Integrated DNA Technologies (IDT) and inserted into a New England Biolabs (NEB) Gibson Assembly reaction into the C-terminus of a low-copy plasmid containing iSpyCas9 (Beisel Lab, NCsu). The hybrid-protein constructs were transformed into electrocompetent E. coli cells with additional PAM-SCANr protein components as described. Overnight cultures were analyzed and sorted on a Becton Dickinson (BD) FACSAria machine. Sorted GFP-positive cells were grown to sufficient density, and plasmids from the pre-sorted and sorted populations were then isolated. The region flanking the nucleotide library was PCR-amplified and submitted for Sanger sequencing (Genewiz). The chromatograms from received trace files were inspected for post-sorted sequence enrichments relative to the pre-sorted library.

**Purification of and DNA cleavage with selected nucleases.** The gBlock (IDT) encoding the PAM-interaction domain of S. mucacae was inserted into a plasmid encoding the PAM-interaction domain of S. pyogenes. The resulting construct was transformed into E. coli construct were transformed into electrocompetent E. coli cells with additional PAM-SCANr protein components as described. Overnight cultures were analyzed and sorted on a Becton Dickinson (BD) FACSAria machine. Sorted GFP-positive cells were grown to sufficient density, and plasmids from the pre-sorted and sorted populations were then isolated. The region flanking the nucleotide library was PCR-amplified and submitted for Sanger sequencing (Genewiz). The chromatograms from received trace files were inspected for post-sorted sequence enrichments relative to the pre-sorted library.

**Discussion**

In summary, we have identified a homolog of SpyCas9 in *Streptococcus macacae* with native 5′-NAAN-3′ PAM specificity. By leveraging the high similarity in Cas9 sequences between different *Streptococcus* species and the substantial background in the development and characterization of SpyCas9, we have engineered variants of SmacCas9 that maintain its minimal adenine dinucleotide PAM specificity and achieve efficient and accurate activity for mediating edits on chromosones in human cells. This finding sets the path for engineering enzymes like iSpyMac with other desirable properties, control points, effectors, and activities.

**Acknowledgements**

This work was supported in part by NIH Grant 1R01HG006688 (D.M.L.) and a grant from the National Science Foundation (DMS 2041954, C.N.K.). The authors thank the NC State Flow Core for instrumentation. We thank the NC State Core Facility for DNA sequencing. We thank Dr. Raul Quiroga for an important discussion. We thank Dr. Thomas Beisel for providing plasmids and essential reagents. We thank Dr. Shannon Poteete for helpful discussions.

**Author contributions**

C.N.K., F.P., D.M.L. and D.S. designed the research. C.N.K. and F.P. generated and characterized the SpyMac and SpyCas9 enzymes. C.N.K. and E.H.R. performed all bacterial and yeast experiments. C.N.K., F.P., D.M.L. and D.S. analyzed and contributed to manuscript preparation.

**Competing interests**

The authors declare no competing interests.
amplified from assemblies of the PAM-SCANR plasmid with a fixed PAM sequence. In vitro digestion reactions with 10 nM target and typically a 10-fold excess of enzyme components were prepared on ice and then incubated in a thermal cycler at 37 °C. Reactions were halted after at least 1 min of incubation by subsequent heat denaturation at 65 °C for 5 min and run on a 2% TAE-agarose gel stained with DNA-intercalating SYBR dye (Invitrogen). Gel images were recorded from blue-light exposure and analyzed in a Python script adapted from https://github.com/jharman25/gelquant/. Cleavage fraction measurements were quantified, in ImageJ (imagej.nih.gov) by the relative intensity of substrate and product bands as follows:

\[
\% \text{ cleaved fraction} = \frac{\text{Integrated intensity of product bands}}{\text{Integrated intensity of all bands}}
\]

Cell culture and DNA modification analysis. HEK293T cells were maintained in DMEM supplemented with 100 units/ml penicillin, 100 mg/ml streptomycin, and 10% fetal bovine serum (FBS). sgRNA plasmids (100 ng) and nuclease plasmids (100 ng) were transfected into cells as duplicates (2 x 10⁴/well in a 96-well plate) with Lipofectamine 3000 (Invitrogen) in Opti-MEM (Gibco). After 5 days post-transfection, genomic DNA was extracted using QuickExtract Solution (Epicentre), and genomic loci were amplified by PCR utilizing the Phusion Hot Start Flex DNA Polymerase (NEB). Amplicons were enzymatically purified and submitted for Sanger sequencing or NGS sequencing. Sanger sequencing ABI files were analyzed using the TIDE algorithm (tide.deskgen.com) 41, in comparison to an unedited control to calculate indel frequencies. NGS FASTQ files were analyzed using a batch version of the software CRISPResso2 (https://github.com/pinellolab/CRISPResso2) 42. All samples were performed in independent duplicates (n = 2). Standard deviation was used to calculate error bars.

Statistical analysis. Data are shown as the mean of duplicate values, which are indicated by dots for each figure. Data were plotted using Matplotlib and the GraphPad Prism software (graphpad.com/scientific-software/prism/).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequence data that support the findings of this study are available via the NIH Sequence Read Archive via BioProject PRJNA623926. Data underlying Figs. 2-3 and Supplementary Fig. 6 are provided as Source Data.

Received: 30 May 2019; Accepted: 9 April 2020;
Published online: 18 May 2020

References

1. Komor, A. C., Badran, A. H. & Liu, D. R. Crispr-based technologies for the manipulation of eukaryotic genomes. Cell 168, 20–36 (2017).
2. Mojica, F. J. M., Diez-Villanueva, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic crispr defence system. *Microbiology* **155**, 733–740 (2009).

3. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the crispr rna-guided endonuclease cas9. *Nature* **507**, 62–67 (2014).

4. Leenay, R. T. & Beisel, C. L. Deciphering, communicating, and engineering the crispr pam. *J. Mol. Biol.* **429**, 177–191 (2017).

5. Komor, A. C., Kim, Y. R., Packer, M. S., Zunis, J. A. & Liu, D. R. Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature* **533**, 420–424 (2016).

6. Gaudelli, N. M. et al. Programmable base editing of a-t to g-c in genomic dna without dna cleavage. *Nature* **551**, 464–471 (2017).

7. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

8. Holtzman, L. & Gersbach, C. A. Editing the epigenome: reshaping the genomic landscape. *Annu. Rev. Genomics Hum. Genet.* **19**, 43–71 (2018).

9. Hu, J. H. et al. Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature* **556**, 57–63 (2018).

10. Brinkman, E. K. & van Steensel, B. in *Methods in Molecular Biology*, 29–44 (Springer New York, 2019).

11. Jang, F. et al. Structures of a crispr-cas9 r-loop complex primed for dna cleavage. *Science* **351**, 867–871 (2016).

**Acknowledgements**

This work was supported by the consortia of sponsors of the MIT Media Lab and the MIT Center for Bites and Atoms. NGS work was supported by a grant (GM115911) to E.J. S from the U.S. National Institutes of Health (NIH). We thank E. Boyden for access to cell culture, in addition to N. Gershenson and S. Zhang for shared lab equipment. We further thank A. Hennes for technical assistance.

**Author contributions**

N.J. identified SmacCas9 and related orthologs as proteins of interest, and L.N. assembled ortholog constructs for PAM characterization. P.C. and N.J. conceived identification strategies for PAM novelty, designed and implemented workflows for PAM discovery, and conducted data analysis for PAM validation. P.C. and N.J. formulated genome editing experiments for mammalian characterization. P.C. carried out genome editing assays, and J.L. conducted NGS experiments and data collection. L.N. optimized protein purification protocols and isolated nucleases for enzymology. S.R.T.K. and E.T. assisted in plasmid construction, transfection, and sample preparation. P.C., N.J., and L.N. wrote the manuscript with input from all authors. N.J. supervised the study, with assistance from E.J.S. and M.J.M.

**Competing interests**

P.C., N.J., L.N., and M.J.M. are inventors of US Patent WO2019217336A2: “Applications of Streptococcus-Derived Cas9 Nucleases on Minimal Adenine-Rich PAM Targets”.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16117-8.

**Correspondence** and requests for materials should be addressed to P.C.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are in the author’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020