PLOS BIOLOGY

RESEARCH ARTICLE

Identification of SaCas9 orthologs containing a conserved serine residue that determines simple NNGG PAM recognition

Shuai Wang[1☯], Chen Tao[1☯], Huilin Mao[1], Linghui Hou[1], Yao Wang[1], Tao Qi[1], Yuan Yang[1], Sang-Ging Ong[2,3], Shijun Hu[4]*, Renjie Chai[5,6,7]*, Yongming Wang[1,8]*

1 State Key Laboratory of Genetic Engineering, School of Life Sciences, Zhongshan Hospital, Fudan University, Shanghai, China, **2** Department of Pharmacology, University of Illinois College of Medicine, Chicago, Illinois, United States of America, **3** Division of Cardiology, Department of Medicine, University of Illinois College of Medicine, Chicago, Illinois, United States of America, **4** Department of Cardiovascular Surgery of the First Affiliated Hospital & Institute for Cardiovascular Science, Collaborative Innovation Center of Hematology, State Key Laboratory of Radiation Medicine and Protection, Suzhou Medical College, Soochow University, Suzhou, China, **5** State Key Laboratory of Bioelectronics, Department of Otolaryngology Head and Neck Surgery, Zhongda Hospital, School of Life Sciences and Technology, Advanced Institute for Life and Health, Jiangsu Province High-Tech Key Laboratory for Bio-Medical Research, Southeast University, Nanjing, China, **6** Co-Innovation Center of Neuroregeneration, Nantong University, Nantong, China, **7** Department of Otolaryngology Head and Neck Surgery, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China, **8** Shanghai Engineering Research Center of Industrial Microorganisms, Shanghai, China

☯ These authors contributed equally to this work.
* shijunhu@suda.edu.cn (SH); renjiec@seu.edu.cn (RC); ymw@fudan.edu.cn (YW)

OPEN ACCESS

Citation: Wang S, Tao C, Mao H, Hou L, Wang Y, Qi T, et al. (2022) Identification of SaCas9 orthologs containing a conserved serine residue that determines simple NNGG PAM recognition. PLoS Biol 20(11): e3001897. https://doi.org/10.1371/journal.pbio.3001897

Academic Editor: Jacob E. Corn, ETH Zurich, SWITZERLAND

Received: June 11, 2022

Accepted: October 31, 2022

Published: November 30, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants from the National Key Research and Development Program of China (2021YFA0910602, 2021YFC2701103 to YW), the National Natural Science Foundation of China (82070258, 81870199 to YW),Open Research Fund of State Key Laboratory of Genetic Engineering, Fudan University (No. SKLGE-2104 to YW) and Science

Abstract

Due to different nucleotide preferences at target sites, no single Cas9 is capable of editing all sequences. Thus, this highlights the need to establish a Cas9 repertoire covering all sequences for efficient genome editing. Cas9s with simple protospacer adjacent motif (PAM) requirements are particularly attractive to allow for a wide range of genome editing, but identification of such Cas9s from thousands of Cas9s in the public database is a challenge. We previously identified PAMs for 16 SaCas9 orthologs. Here, we compared the PAM-interacting (PI) domains in these orthologs and found that the serine residue corresponding to SaCas9 N986 was associated with the simple NNGG PAM requirement. Based on this discovery, we identified five additional SaCas9 orthologs that recognize the NNGG PAM. We further identified three amino acids that determined the NNGG PAM requirement of SaCas9. Finally, we engineered Sha2Cas9 and SpeCas9 to generate high-fidelity versions of Cas9s. Importantly, these natural and engineered Cas9s displayed high activities and distinct nucleotide preferences. Our study offers a new perspective to identify SaCas9 orthologs with NNGG PAM requirements, expanding the Cas9 repertoire.

Introduction

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-RNA-guided Cas endonuclease system is based on the bacterial adaptive immune system and has been utilized

and Technology Research Program of Shanghai (19DZ2282100 to YW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: authors have applied a patent related to the work.

Abbreviations: AAV, adeno-associated virus; ANOVA, analysis of variance; CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; crRNA, CRISPR RNA; indels, insertions or deletions; PAM, protospacer adjacent motif; PI, PAM-interacting; PMSF, phenylmethanesulfonyl fluoride; PVDF, polyvinylidene fluoride; sgRNA, single-guide RNA; tracrRNA, *trans*-activating crRNA.

as a fast and efficient method for precise genome editing [1–6]. This system is made up of two main components: a Cas9 nuclease and a chimeric single-guide RNA (sgRNA) derived from CRISPR RNA (crRNA) and the *trans*-activating crRNA (tracrRNA) [2]. Cas9 and sgRNA combine to form a complex that recognizes the target DNA that is complementary to the 5′ end of the sgRNA [2]. In addition to sgRNA-target DNA complementarity, DNA recognition requires a specific DNA sequence known as protospacer adjacent motif (PAM), flanking the target sequence [2]. The PAM allows the Cas nuclease to discriminate between the target DNA and the DNA sequence encoding the sgRNA but also restricts its ability to target any sequence in the genome.

Editing efficiency is a major hurdle of the CRISPR system. Every Cas nuclease has its own nucleotide preference [7]. For example, SpCas9 prefers guanine-rich sequences [8], while AsCas12a prefers adenine-rich sequences [9]. SpCas9 is generally considered the most efficient Cas nuclease, whose efficiency varies from 0% to approximately 100% depending on the target sequences [8]. Although previous studies have focused on limitations of the PAM [10–12], the sole presence of a PAM within a locus does not guarantee that it can be efficiently edited. For high efficiency of genome editing to be achieved, it is essential to establish a Cas9 repertoire that can accommodate all sequences.

Cas9 nucleases with flexible PAM requirements are crucial for large-scale genome editing. We previously developed Cas9 nucleases with highly flexible NNGG PAMs recognition [13,14]. To rapidly identify additional natural Cas9 nucleases recognizing NNGG PAMs, we compared the PAM-interacting (PI) domains of SaCas9 orthologs and found that the serine residue corresponding to SaCas9 N986 was associated with the NNGG PAM. We identified five additional SaCas9 orthologs recognizing the NNGG PAM. We further engineered two of them to improve the specificity. Our study expands the Cas9 repertoire and provides a foundation to search for Cas9s with NNGG PAMs in the future.

Results

A serine residue was associated with the NNGG PAM requirement among SaCas9 orthologs

We previously identified PAMs for 16 SaCas9 orthologs, where SauriCas9 and SlugCas9 recognized NNGG PAMs [13–15]. Nishimasu and colleagues have demonstrated that amino acids of N985, N986, R991, E993, and R1015 in the PI domain of SaCas9 are crucial for PAM recognition [16]. N985, E993, and R1015 are very conserved among these 16 orthologs (Fig 1A). In contrast, residues corresponding to N986 and R991 showed substantial diversity. Interestingly, SauriCas9 and SlugCas9 contain a serine residue corresponding to SaCas9 N986. We hypothesized that this serine residue is associated with the NNGG PAM.

We employed SaCas9 as a template to search for related orthologs from NCBI's Gene database, and our search identified five additional Cas9s that contained this serine residue with amino acid identity ranging from 58.4% to 64.5% (Fig 1A and Table 1). Genetic loci of these orthologs contain a conserved organization where Cas9 is followed by Cas1 and Cas2 (S1A Fig). The organization of CRISPR repeats and tracrRNAs does not appear to be conserved. SaCas9 encodes a tracrRNA upstream of Cas9 and CRISPR repeats downstream of Cas2. Sha2-Cas9 and SpeCas9 encode CRISPR repeats and tracrRNAs downstream of Cas2. SwaCas9 and Swa2Cas9 encode CRISPR repeats and tracrRNAs upstream of Cas9 and additional CRISPR repeats downstream of Cas2. SmiCas9 encodes CRISPR repeats-tracrRNA-CRISPR repeats upstream of Cas9.

Nevertheless, the 5′ end sequences of CRISPR repeats and tracrRNAs exhibited high conservation among these orthologs (S1B and S1C Fig). We fused the 3′ end of a direct repeat with

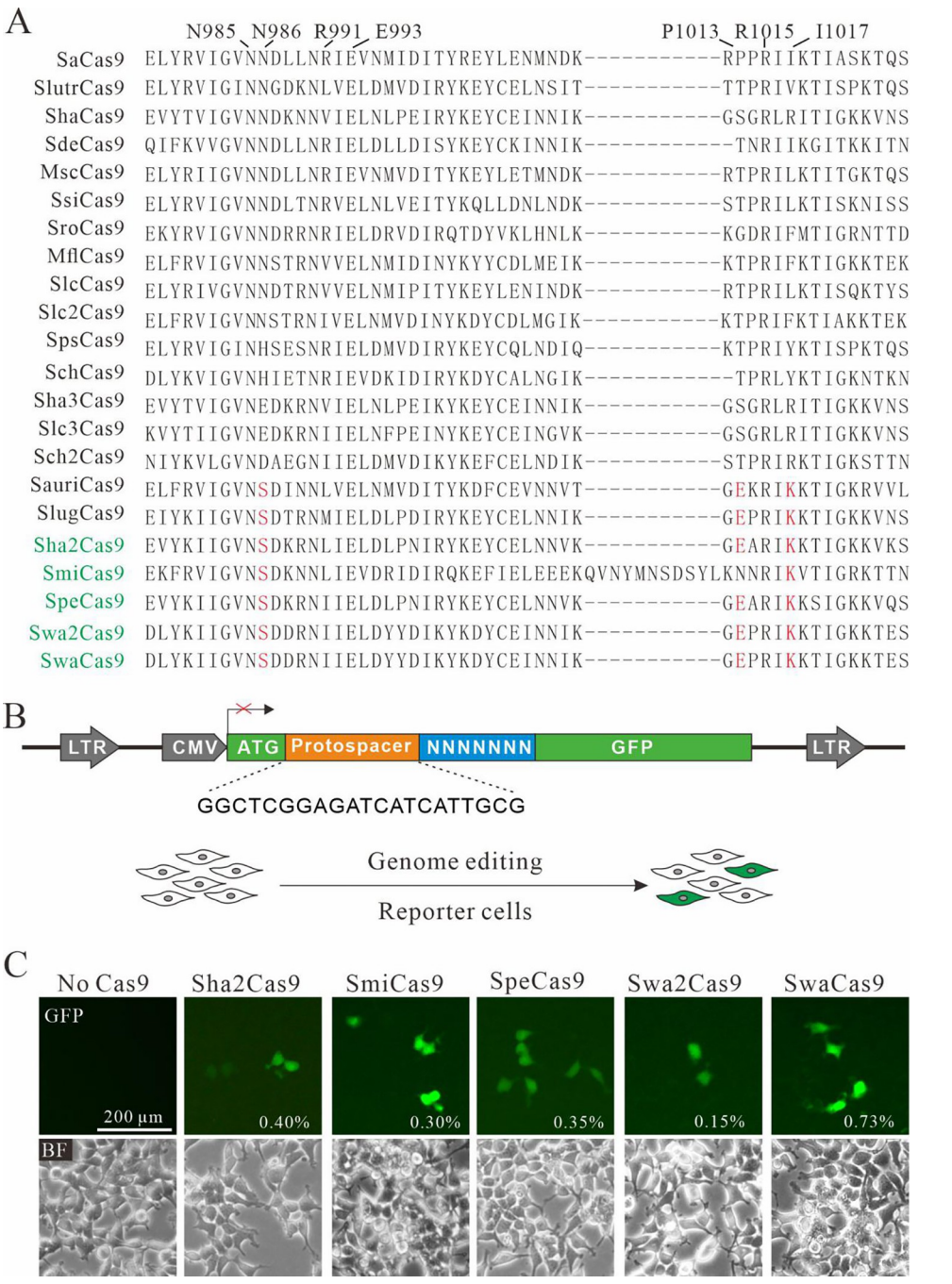

Fig 1. Analysis of five SaCas9 ortholog activities. (**A**) Amino acid sequences of the SaCas9 ortholog PI domain are aligned. The residues that are important for PAM recognition are indicated at the top; the conserved residues among newly identified SaCas9 orthologs are shown in red; the names of newly identified Cas9s are shown in green. (**B**) Design of the GFP activation reporter construct. A target sequence (protospacer) containing a 7-bp random sequence is inserted between ATG and the GFP-coding sequence. The library DNA is stably integrated into HEK293T cells by lentivirus. (**C**) Transfection of SaCas9 orthologs induced GFP expression. Percentage of GFP-positive cells was shown. The cells without transfection of Cas9 were used as a negative control.

https://doi.org/10.1371/journal.pbio.3001897.g001

Table 1. Five SaCas9 orthologs selected from the NCBI database.

NCBI ID	Host strain	Name	Length (aa)	Identity to SaCas9
Sha2Cas9	*Staphylococcus haemolyticus*	WP_154836552	1,058	63.2%
SmiCas9	*Staphylococcus microti*	WP_044361501	1,063	58.4%
SpeCas9	Staphylococcus petrasii	WP_115359133	1,058	63.5%
SwaCas9	*Staphylococcus warneri*	WP_107532850	1,054	64.5%
Swa2Cas9	*Staphylococcus warneri*	WP_114599540	1,054	64.3%

https://doi.org/10.1371/journal.pbio.3001897.t001

the 5′ end of the corresponding tracrRNA, including the full-length tail, via a 4-nt linker to form a sgRNA for each Cas9 (S2A Fig). Interestingly, these sgRNAs formed a similar secondary structures with three stem loops (S2B Fig), suggesting that these SaCas9 orthologs could share the same sgRNA scaffold for genome editing.

PAM screening

Next, we used a previously established GFP activation assay for PAM screening [13,15]. In this assay, the GFP expression is disrupted by a target sequence (protospacer) flanked by a 7-bp random sequence, which is inserted into the GFP coding sequence immediately downstream of the ATG start codon, inducing to a frameshift mutation. This reporter library is then stably integrated into HEK293T cells. If a Cas9 nuclease successfully edits the target sequence, small insertions or deletions (indels) will be generated at the target sequence, and a functional GFP cassette will be restored in a portion of cells (Fig 1B). Each Cas9 was human codon optimized, synthesized, and cloned into a mammalian expression construct that was developed by Ran and colleagues [17]. The canonical SaCas9 sgRNA scaffold was employed for sgRNA expression [17]. Three days after transfection of Cas9 with sgRNA expression plasmids, all five tested Cas9s induced GFP expression (Fig 1C). GFP-positive cells were sorted out and the target DNA was PCR amplified for deep sequencing. Sequencing results showed that indels occurred at target sites (Fig 2A). WebLogos and PAM wheels were generated based on deep sequencing data, which revealed that these Cas9s recognized NNGG PAMs (Fig 2B and 2C). These data validated our hypothesis that the serine residue is associated with the NNGG PAM.

To test whether the serine residue corresponding to SaCas9 N986 determined NNGG PAM recognition, we replaced N986 with a serine (S3A Fig). The GFP activation assay revealed that the substitution increased guanine preference at PAM position 3, but the favored PAM remained NNGRRT (S3B Fig). We reanalyzed the PI domain of these orthologs and identified two additional residues, a glutamic acid (E) corresponding to SaCas9 P1013 and a lysine (K) corresponding to SaCas9 I1017, conserved among orthologs that recognized an NNGG PAM, except SmiCas9 where there is a 13-amino acid insertion corresponding to SaCas9 P1013 (Fig 1A). We added either P1013E, I1017K, or both mutations to SaCas9-N986 (S3A Fig). Interestingly, all resulted SaCas9 variants recognized an NNGG PAM (S3B Fig). These data demonstrated that these three amino acids are important for determining the NNGG PAM.

Genome editing for endogenous loci

Next, we tested the capacity of these Cas9s for genome editing at selected endogenous sites in HEK293T cells. Five days after transfection of Cas9 and sgRNA expression plasmid DNA, we extracted genomic DNA and amplified target sites by PCR. As an initial screen, we used the T7EI assay to rapidly analyze the efficiency for each Cas9. SmiCas9, Sha2Cas9, and SpeCas9 displayed higher editing efficiency, while SwaCas9 and Swa2Cas9 displayed lower editing efficiency (S4A and S4B Fig). In the subsequent experiments, we only focused on SmiCas9, Sha2-Cas9, and SpeCas9.

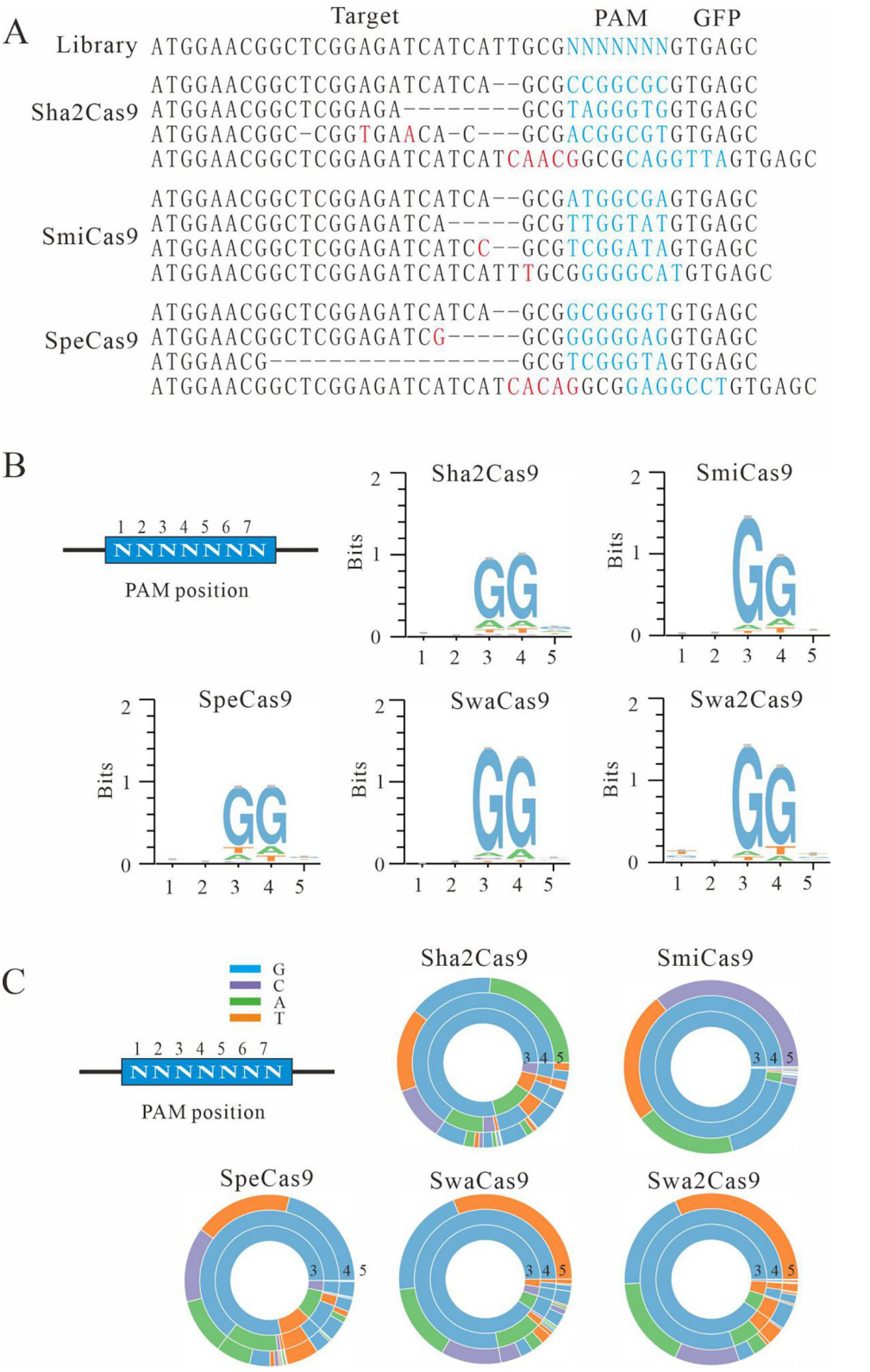

Fig 2. Analysis of the PAM sequence of Cas9. (**A**) Deep sequencing reveals that SmiCas9, Sha2Cas9, and SpeCas9 generated indels on the targets. (**B**) WebLogos were generated based on the deep sequencing data. (**C**) PAM wheels were generated based on the deep sequencing data.

https://doi.org/10.1371/journal.pbio.3001897.g002

We compared the activity of these three Cas9s to that of SaCas9 at 13 endogenous sites with NNGGRT PAMs. All tested Cas9s were expressed from the same construct and achieved similar expression levels, as revealed by western blot (Fig 3A and 3B). All four Cas9 nucleases generated indels with different efficiencies depending on the target sites in HEK293T cells (Fig 3C). Interestingly, these Cas9s displayed different activities at some sites. For example, Sha2-Cas9 displayed higher activity at site E0, while SmiCas9 and SpeCas9 displayed higher activity at site G10. SaCas9 displayed lower efficiency than newly identified Cas9s at sites G3 and G9. These data demonstrated that these Cas9s prefer distinct target sequences. Overall, SaCas9, Sha2Cas9, and SpeCas9 displayed comparable activities, while SmiCas9 displayed lower activity (Fig 3D).

Specificity of SmiCas9, Sha2Cas9, and SpeCas9

Next, we compared the specificity of SmiCas9, Sha2Cas9, SpeCas9, and SaCas9 using the GFP activation assay. A panel of sgRNAs with dinucleotide mutations along the protospacer was generated to detect the specificity of each Cas9. Off-target cleavage is considered to have occurred when the mismatched sgRNAs induce GFP expression. Overall, SaCas9 and SmiCas9 had negligible off-target effects, while Sha2Cas9 and SpeCas9 displayed moderate off-target effects (S5 Fig). Specifically, SaCas9 was highly sensitive to mismatches at PAM-proximal and PAM-distal positions but relatively less sensitive at middle positions; SmiCas9 displayed minimal off-target effects with mismatches at all positions; and Sha2Cas9 and SpeCas9 were sensitive to mismatches at PMA-proximal positions 18 through 20 but less sensitive at other positions.

Recently, Tan and colleagues unraveled the crystal structure of the SaCas9/sgRNA–target DNA complex and identified four amino acid residues (R245, N413, N419, and R654) forming polar contacts within a 3.0-Å distance from the target DNA strand [18]. When one or more of these residues were replaced by alanine, SaCas9 specificity was significantly improved [18]. To investigate whether the specificity of Sha2Cas9 can be improved, we used pairwise alignment to identify the corresponding residues (R247, N415, S421, and R656; S6 Fig) and generated single amino acid mutants by alanine substitution. The GFP activation assay revealed that the R247A and N415A mutations could significantly improve specificity without compromising the on-target activity (S7A and S7B Fig). The R656A mutation also improved the specificity although this was accompanied by markedly decreased on-target activity. We introduced the R247A and N415A double mutations into Sha2Cas9 to generate a high-fidelity version of Cas9 named Sha2Cas9-HF. The GFP activation assay revealed that double mutations further improved its specificity (Fig 4A).

We simultaneously identified the corresponding residues for SpeCas9 (R247, N415, S421, and R656; S6 Fig) and generated single amino acid mutants by alanine substitution (S8A Fig). The GFP activation assay revealed that the R247A, N415A, and S421A mutations could significantly improve specificity without compromising the on-target activity (S8B Fig). We introduced the R247A, N415A, and S421A triple mutations into SpeCas9 to generate a high-fidelity version of Cas9 named SpeCas9-HF. The GFP activation assay revealed that triple mutations further improved specificity (Fig 4A).

Genome-wide unbiased off-target effects of Sha2Cas9, Sha2Cas9-HF, SpeCas9, and SpeCas9-HF were next evaluated by GUIDE-seq [19]. We evaluated two sites targeting the EXM1 gene and one site targeting the RUNX1 gene. Five days after transfection of the Cas9 plasmid, the sgRNA plasmid, and the GUIDE-seq oligos, we prepared libraries for deep sequencing. Sequencing and analysis showed that on-target cleavage occurred for all Cas9 nucleases at 3 targets, as reflected by the high GUIDE-seq read counts (Fig 4B). High-fidelity versions of

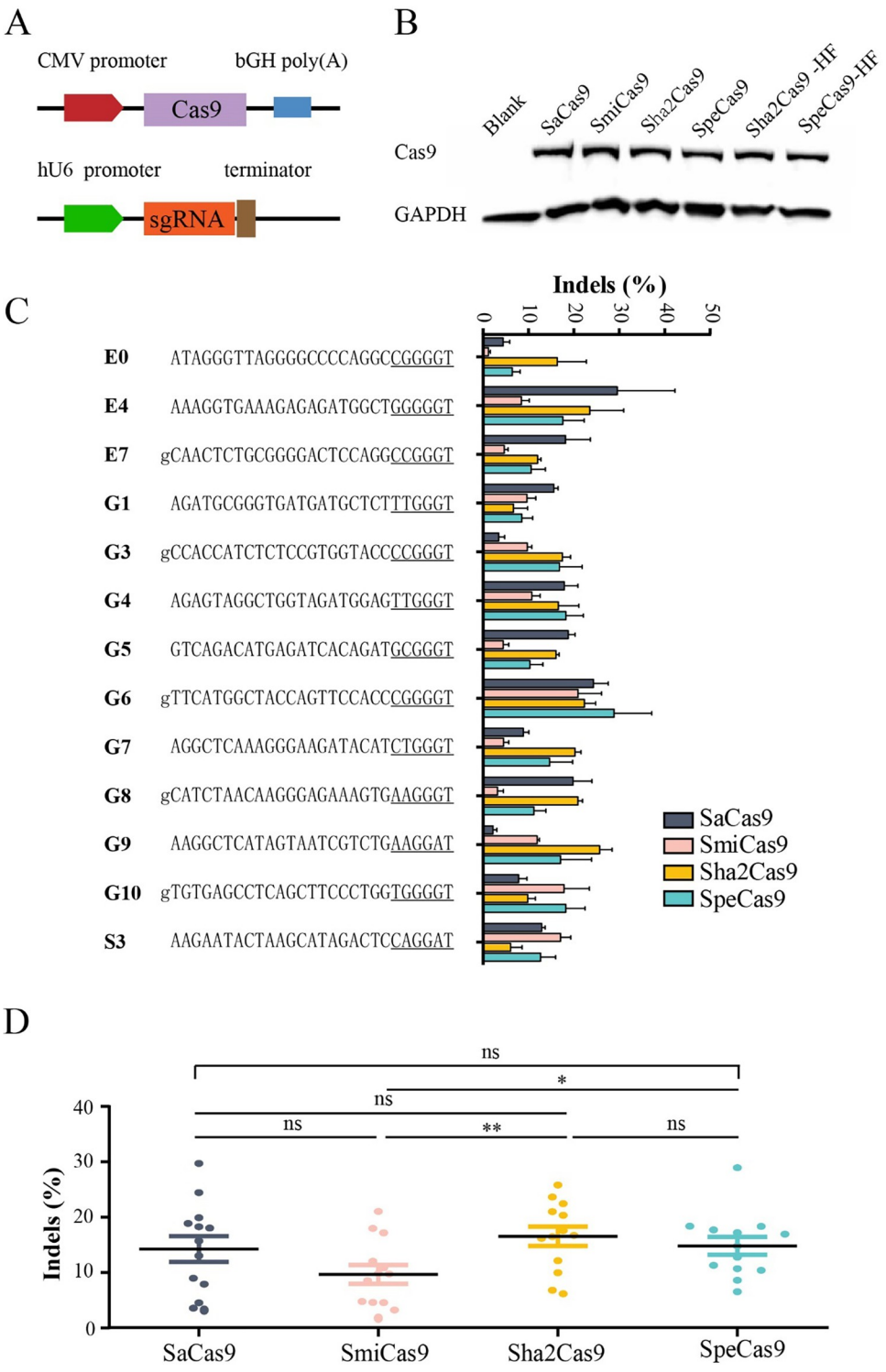

Fig 3. Genome editing for endogenous sites. (**A**) Schematic of the Cas9 expression constructs. (**B**) Protein expression level of Cas9s was measured by western blot. Cells without Cas9 transfection was used as a negative control. (**C**) Comparison of SaCas9, SmiCas9, Sha2Cas9, and SpeCas9 efficiency for genome editing at 13 endogenous loci. Additional "g" is added for U6 promoter transcription ($n = 3$). Underlying data for all summary statistics can be found in S1 Data. (**D**) Quantification of editing efficiency for SaCas9, SmiCas9, Sha2Cas9, and SpeCas9. Underlying data for all summary statistics can be found in S1 Data.

https://doi.org/10.1371/journal.pbio.3001897.g003

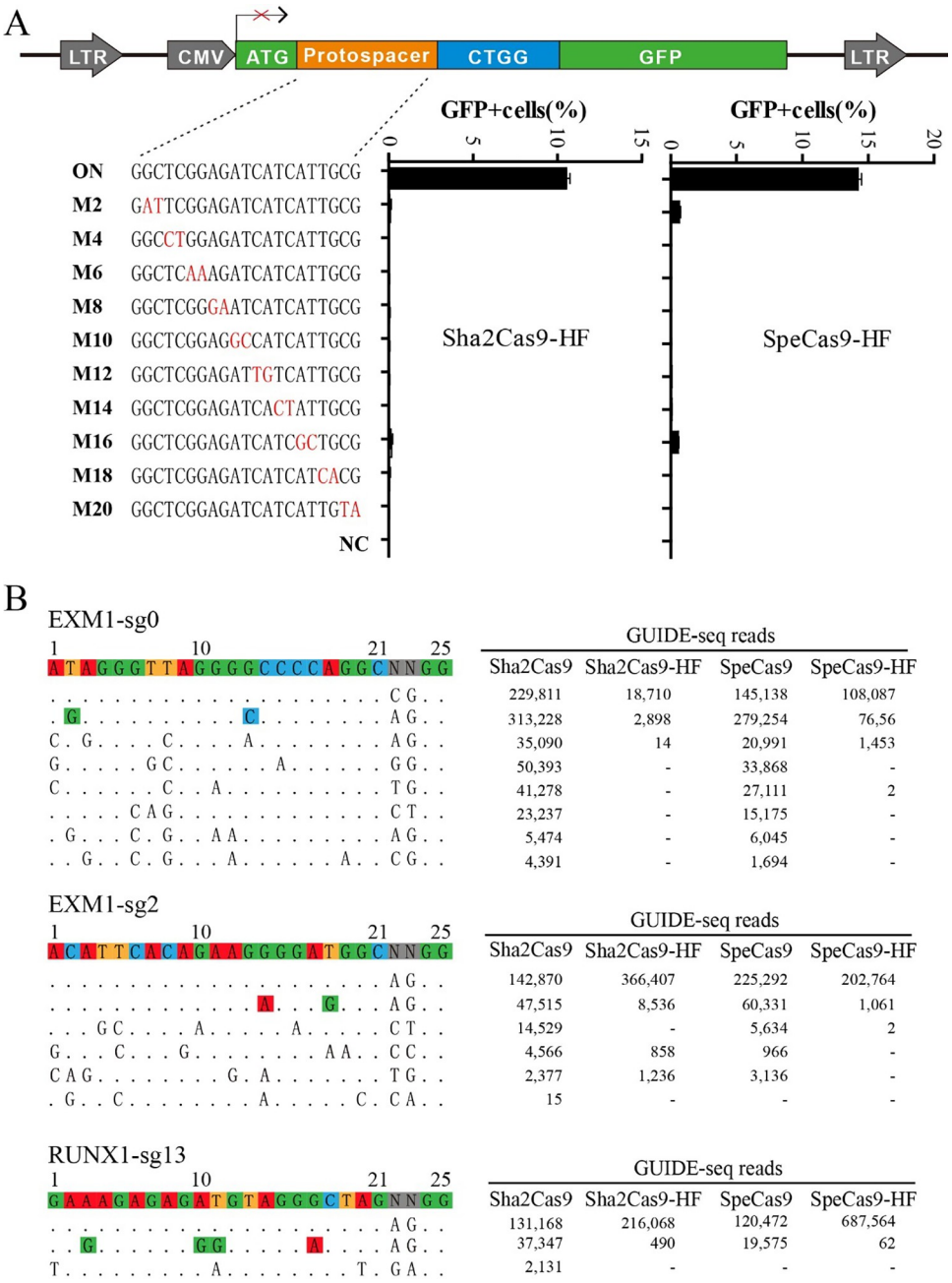

Fig 4. Analysis of Sha2Cas9-HF and SpeCas9-HF specificity. (**A**) Schematic of the GFP activation assay for specificity analysis is shown on the top. A panel of sgRNAs with dinucleotide mutations is shown below. sgRNA activities were measured based on GFP expression. Mismatches are shown in red ($n = 3$). Underlying data for all summary statistics can be found in S1 Data. (**B**) Off-targets for EMX1 locus are analyzed by GUIDE-seq. Read numbers for on- and off-targets are shown on the right. Mismatches compared with the on-target site are shown and highlighted in color.

https://doi.org/10.1371/journal.pbio.3001897.g004

Cas9s displayed significantly fewer off-target effects than wild-type Cas9s, reflected by the numbers of off-target sites and off-target read counts. For example, SpeCas9 and SpeCas9-HF generated similar read counts (225,292 versus 202,764) at the EXM1-sg2 site. SpeCas9 induced four off-target sites, while SpeCas9-HF induced two off-target sites. For one off-target, SpeCas9

generated 60,331 read counts, while SpeCas9-HF generated 1,061 read counts. For another off-target, SpeCas9 generated 5,634 read counts, while SpeCas9-HF generated 2 read counts. These data demonstrated that the occurrence of off-target events is significantly lower when using Sha2Cas9-HF and SpeCas9-HF.

Evaluation of Sha2Cas9-HF and SpeCas9-HF on-target activities

Next, we compared the activities of high-fidelity Cas9s to those of wild-type Cas9s (Sha2Cas9 versus Sha2Cas9-HF; SpeCas9 versus SpeCas9-HF) with a panel of 13 endogenous sites. Western blot analysis revealed that the protein expression levels of high-fidelity Cas9s and wild-type Cas9s were comparable (Fig 3B). All four Cas9s generated indels at targets with varying efficiencies (Fig 5A). Overall, high-fidelity Cas9s and wild-type Cas9s displayed comparable efficiencies (Fig 5B). However, different efficiencies were observed for a number of targets. For example, SpeCas9-HF displayed higher efficiency than SpeCas9 at the G5 site, whereas SpeCas9-HF displayed lower efficiency than SpeCas9 at the G8 site. These data demonstrated that the preference of high-fidelity Cas9s for nucleotides differs from that of wild-type Cas9s for genome editing.

Discussion

Different nucleotide preferences have been observed among natural Cas9 nucleases. For example, SpCas9 favors G-rich sequences but disfavors T-rich sequences [8]; AsCas12a favors A-rich sequences but disfavors G-rich sequences [9]. One possible strategy to achieve high efficiency of genome editing is to harness multiple natural Cas nucleases for genome editing, and a collection of these nucleases could cover all possible sequences. A number of Cas nucleases, such as SaCas9 [17], NmeCas9 [20], CjCas9 [21], AaCas12b [22], and Cas12f1 [23,24], have been harnessed for genome editing. We previously developed BlatCas9 [25], SauriCas9 [13], SlugCas9 [14], and SchCas9 [15] for genome editing. In this study, we further expanded the Cas repertoire by developing SmiCas9, Sha2Cas9, and SpeCas9. Importantly, they contain a compact genome, facilitating delivery by a single adeno-associated virus (AAV) for in vivo genome editing. These newly developed Cas9s will enhance our ability to achieve high efficiency genome editing.

Different nucleotide preferences have also been observed between natural Cas9s and their engineered variants. We and others previously screened thousands of sgRNA activities for SpCas9 and its engineered variants and observed different nucleotide preferences [8,26]. For example, SpCas9 slightly prefers A and G at sgRNA position 10, while SpCas9-HF1 strongly prefers C at this position [8]. In this study, we generated two high-fidelity versions of Cas9s. Although they only contain 2 or 3 amino acid modifications, distinct nucleotide preferences were observed for a number of targets. Therefore, engineered Cas9s not only change specificity or targeting scope [27–30] but also change nucleotide preferences.

Cas9s with flexible PAMs are crucial for precision positioning. In addition to SpCas9, several other natural Cas nucleases with dinucleotide PAMs have been identified, including FnCas9 [31], Nme2Cas9 [32], SauriCas9 [13], SlugCas9 [14], SchCas9 [15], and AaCas12b [22]. In this study, we identified the serine residue corresponding to SaCas9 N986 associated with the simple NNGG PAM requirement. This PAM occurs, on average, once in every approximately 8 randomly chosen genomic loci. We further identified three amino acids that determined the NNGG PAM requirement of SaCas9. With the continuous expansion of the Cas9 database, our strategy will offer a clue to identify more SaCas9 orthologs with NNGG PAMs.

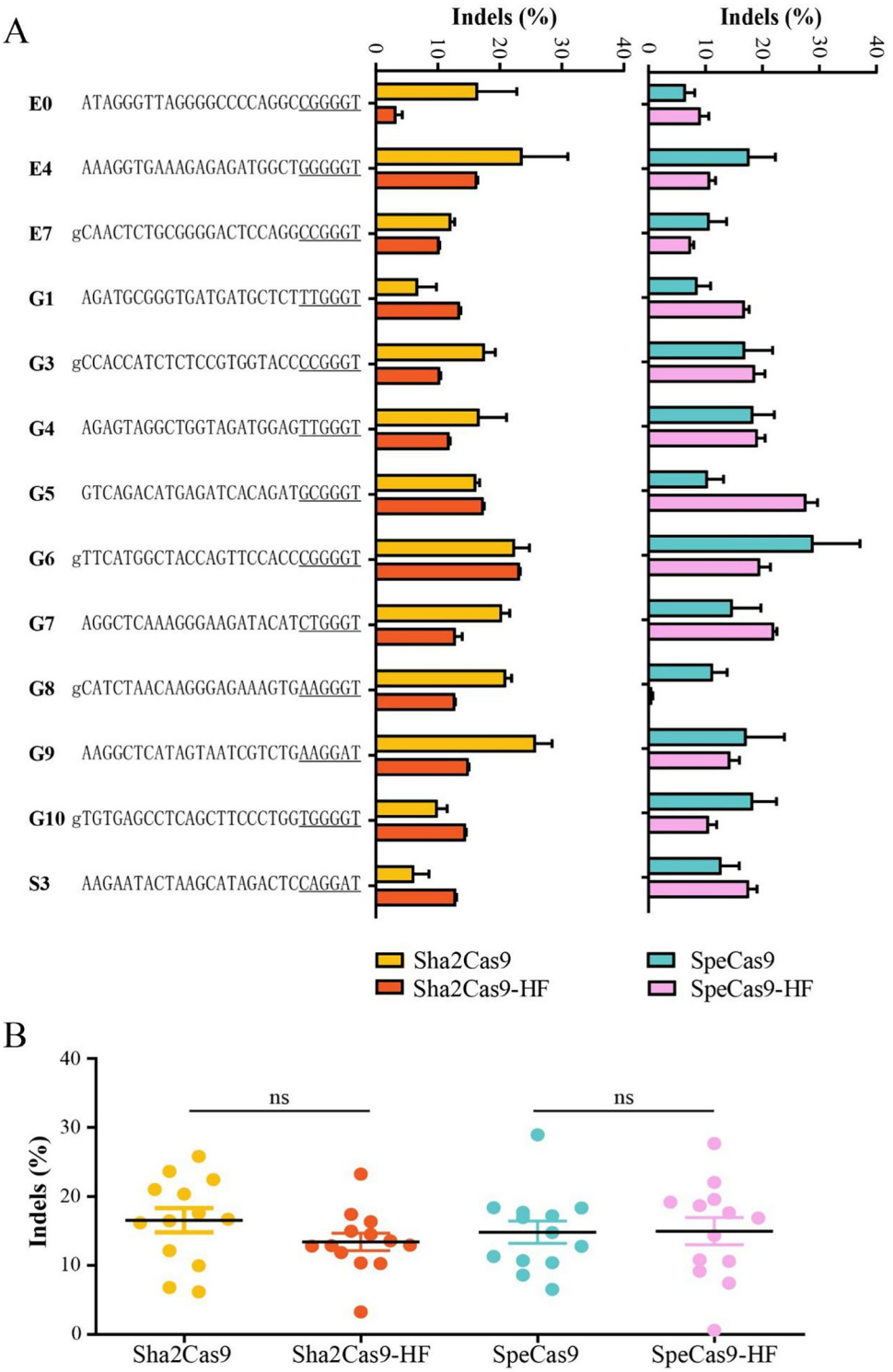

Fig 5. Evaluation of Sha2Cas9-HF and SpeCas9-HF on-target activities. (**A**) Comparison of activities of high-fidelity Cas9s to the wild-type Cas9s ($n = 3$). The target sequences are shown on the left. PAM is underlined. If the first nucleotide is C or T, additional "g" is added for U6 promoter transcription. Underlying data for all summary statistics can be found in S1 Data. (**B**) Quantification of editing efficiency for SaCas9, SmiCas9, Sha2Cas9, and SpeCas9. Underlying data for all summary statistics can be found in S1 Data.

https://doi.org/10.1371/journal.pbio.3001897.g005

Materials and methods

Cell culture and transfection

HEK293T cells were cultured in DMEM (Gibco) supplemented with 10% FBS (Gibco) and 1× penicillin–streptomycin (Gibco) at 37˚C with 5% CO_2. HEK293T cells were transfected with Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. For Cas9 PAM sequence screening, 1.2×10^7 HEK293T cells were transfected with a total of 10 μg of Cas9 plasmid and 5 μg of sgRNA plasmid in 10-cm dishes. For genome editing comparisons of Cas9, 10^5 cells were transfected with a total of 300 ng of Cas9 plasmid and 200 ng of sgRNA plasmid in 48-well plates.

Plasmid construction

Cas9 expression plasmid construction: The plasmid pX601 (Addgene#61591) was amplified by the primers px601-F/px601-R to obtain the pX601 backbone. The human codon–optimized Cas9 gene (S1 Table) was synthesized by HuaGene (Shanghai, China) and cloned into the pX601 backbone by the NEBuilder assembly tool (NEB) according to the manufacturer's instructions. Sequences of each Cas9 were confirmed by Sanger sequencing (GENEWIZ, Suzhou, China).

sgRNA expression plasmid construction: sgRNA expression plasmids were constructed by ligating sgRNA into the Bsa1-digested hU6-Sa_tracr plasmid. The primer sequences and target sequences are listed in S2 and S3 Tables, respectively.

PAM sequence analysis

Twenty base-pair sequences (AAGCCTTGTTTGCCACCATG/GTGAGCAAGG GCGAGGA GCT) flanking the target sequence (GAACGGCTCGGAGATCATC ATTGCGNNNNNNN) were used to fix the target sequences. GCG and GTGAGCAAGGGCG AGGAGCT were used to fix a 7-bp random sequence. Target sequences with in-frame mutations were used for PAM analysis. The 7-bp random sequence was extracted and visualized by WebLogo [33] and a PAM wheel chart to identify PAMs [29].

Genome editing for endogenous sites

HEK293T cells were seeded into 48-well plates and transfected with a total of 300 ng of Cas9 plasmid and 200 ng of sgRNA plasmid by Lipofectamine 2000 (1 μL). Cells were collected 5 days after transfection. Genomic DNA was isolated, and the target sites were PCR amplified and extracted by QuickExtract DNA Extraction Solution (Epicentre) for deep sequencing. For genomic HEK293T DNA, the PCR products were subjected to a T7E1 assay to check the editing efficiency. The primer sequences are listed in S2 Table.

Test of Cas9 specificity

To test the specificity of Cas9, we generated two GFP reporter cell lines with the CTGG PAM. The cells were seeded into 48-well plates and transfected with 300 ng of Cas9 plasmids and 200 ng of sgRNA plasmids by using Lipofectamine 2000. Five days after editing, the GFP-positive cells were analyzed on a Calibur instrument (BD). The data were analyzed using FlowJo.

GUIDE-seq

GUIDE-seq experiments were performed as described previously [19], with minor modifications. Briefly, 2×10^5 HEK293T cells were transfected with 500 ng of SchCas9/Sa-SchCas9, 500

ng of sgRNA plasmids, and 100 pmol of annealed GUIDE-seq oligonucleotides by electroporation and then seeded into 6 wells. The electroporation voltage, width, and the number of pulses were 1,150 V, 30 ms, and 1 pulse, respectively. Genomic DNA was extracted with the DNeasy Blood and Tissue kit (QIAGEN) 6 days after transfection according to the manufacturer's protocol. The genome library was prepared and subjected to deep sequencing [19].

Western blotting

One day before transfection, HEK293T cells were seeded into a 6-well plate. For each well, 2 μg of Cas9-expressing plasmid were transfected using 4 μL of Lipofectamine2000. Three days after transfection, cell samples were collected and total proteins were extracted using NP-40 buffer (Beyotime) supplemented with 1 mM phenylmethanesulfonyl fluoride (PMSF) (Beyotime). The protein was separated by SDS-PAGE gel and transferred onto polyvinylidene fluoride (PVDF) (Thermo) membrane. After transfer, the membrane was blocked with 5% (wt./vol.) BSA (Sigma) in TBS-T (0.1% Tween 20 in 1× TBS) buffer and then incubated in the primary antibody (anti-HA tag (1:1,000; ab236632, Abcam) and anti-GAPDH (1:2,000; 5174s, Cell Signaling) at 4˚C overnight. Wash membrane three times in TBS-T for 5 min each time. The second antibody (1:10,000; ab6721, Abcam) was incubated for 1 h at room temperature, and then washed three times and imaged.

Statistical analysis

All the data are shown as mean ± SD. Statistical analyses were performed using Microsoft Excel. Two-tailed, paired Student *t* tests were used to determine statistical significance when comparing two groups, whereas analyses of variance (ANOVAs) are used for comparisons between for three or more groups. A value of $P < 0.05$ was considered to be statistically significant ($^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$).

Supporting information

S1 Fig. Genetic locus of CRISPR/Cas9. (**A**) The structures of CRISPR loci for six SaCas9 orthologs. (**B**) Alignment of CRISPR repeat sequences for six SaCas9 orthologs. (**C**) Alignment of tracrRNA for six SaCas9 orthologs.
(TIF)

S2 Fig. Analysis of sgRNAs. (**A**) Alignment of sgRNA scaffolds for six SaCas9 orthologs. The GAAA linker are indicated by the black box. (**B**) Analysis of SaCas9 orthologs' secondary RNA structures. These structures were generated by an online tool named RNAfold WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi).
(TIF)

S3 Fig. Analysis of the SaCas9 variant PAMs. (**A**) Amino acid sequence of the SaCas9 variant PI domains. The residues that are important for PAM recognition are marked at the top; the mutations are highlighted in red. (**B**) SaCas9 variant PAMs were analyzed by the GFP activation assay. WebLogos generated by analyzing the deep sequencing data.
(TIF)

S4 Fig. Evaluation of the genome editing efficiency of 6 SaCas9 orthologs. (**A**) Examples of the gel pictures of T7EI assay for Sha2Cas9 and SpeCas9. Cleaved fragments are marked by red triangles. Indel frequencies are shown below. Underlying data for all summary statistics can be found in S1 Data. (**B**) Quantification of editing efficiency for 6 SaCas9 orthologs. Underlying

data for all summary statistics can be found in S1 Data.
(TIF)

S5 Fig. Analysis of four Cas9 ortholog specificity. Schematic of the GFP activation assay for specificity analysis is shown on the top. A panel of sgRNAs with dinucleotide mutations is shown below. sgRNA activities were measured based on GFP expression. Cells without Cas9 transfection were used as a negative control (NC). Mismatches are shown in red ($n = 3$). Underlying data for all summary statistics can be found in S1 Data.
(TIF)

S6 Fig. Protein sequence alignment of SaCas9, Sha2Cas9, and SpeCas9. The amino acid residues important for specificity are indicated by vertical lines above. The amino acid residue positions are shown on the right.
(TIF)

S7 Fig. Specificity of four Sha2Cas9 variants. (**A**) Schematic of Sha2Cas9 structure. The amino acid residues important for specificity are shown below. (**B**) Test of four Sha2Cas9 variant specificity. Schematic of the GFP activation assay for specificity analysis is shown on the top. A panel of sgRNAs with dinucleotide mutations is shown below. sgRNA activities were measured based on GFP expression. Cells without Cas9 transfection were used as a negative control (NC). Mismatches are shown in red ($n = 2$ or 3). Underlying data for all summary statistics can be found in S1 Data.
(TIF)

S8 Fig. Specificity of four SpeCas9 variants. (**A**) Schematic of SpeCas9 structure. The amino acid residues important for specificity are shown below. (**B**) Test of four SpeCas9 variant specificity. Schematic of the GFP activation assay for specificity analysis is shown on the top. A panel of sgRNAs with dinucleotide mutations is shown below. sgRNA activities were measured based on GFP expression. Cells without Cas9 transfection were used as a negative control (NC). Mismatches are shown in red ($n = 3$). Underlying data for all summary statistics can be found in S1 Data.
(TIF)

S1 Data. Underlying values for all reported summary statistics. Raw data from all reported summary statistics.
(XLSX)

S1 Table. The Cas9 ID and human codon–optimized Cas9 gene. The file contains the Cas9 ID, host strain, tracrRNA, and amino acid sequences of SaCas9 orthologs used in this study. The human codon–optimized Cas9 genes were synthesized.
(DOCX)

S2 Table. Primers used in this study. A list of oligonucleotide pairs and primers used for deep sequencing.
(DOCX)

S3 Table. Target sites used in this study. A list of the endogenous target sites of human and their downstream PAM. PAM, protospacer adjacent motif.
(DOCX)

S1 Raw Images. Raw images of Figs 3 and S4.
(JPG)

Author Contributions

Conceptualization: Yongming Wang.

Formal analysis: Tao Qi, Yuan Yang.

Funding acquisition: Yongming Wang.

Investigation: Shuai Wang, Chen Tao, Huilin Mao, Linghui Hou, Yao Wang.

Project administration: Yongming Wang.

Supervision: Shijun Hu, Renjie Chai, Yongming Wang.

Visualization: Shuai Wang, Chen Tao.

Writing – original draft: Yongming Wang.

Writing – review & editing: Sang-Ging Ong.

References

1. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013; 339(6121):819–823. https://doi.org/10.1126/science.1231143 PMID: 23287718; PubMed Central PMCID: PMC3795411.
2. 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(6096):816–821. https://doi.org/10.1126/science.1225829 PMID: 22745249.
3. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339(6121):823–826. https://doi.org/10.1126/science.1232033 PMID: 23287722; PubMed Central PMCID: PMC3712628.
4. Xie Y, Wang D, Lan F, Wei G, Ni T, Chai R, et al. An episomal vector-based CRISPR/Cas9 system for highly efficient gene knockout in human pluripotent stem cells. Sci Rep. 2017; 7(1):2320. https://doi.org/10.1038/s41598-017-02456-y PMID: 28539611; PubMed Central PMCID: PMC5443789.
5. Qi T, Wu F, Xie Y, Gao S, Li M, Pu J, et al. Base Editing Mediated Generation of Point Mutations Into Human Pluripotent Stem Cells for Modeling Disease. Front Cell Dev Biol. 2020; 8:590581. https://doi.org/10.3389/fcell.2020.590581 PMID: 33102492; PubMed Central PMCID: PMC7546412.
6. Wang B, Wang Z, Wang D, Zhang B, Ong SG, Li M, et al. krCRISPR: an easy and efficient strategy for generating conditional knockout of essential genes in cells. J Biol Eng. 2019; 13:35. https://doi.org/10.1186/s13036-019-0150-y PMID: 31049076; PubMed Central PMCID: PMC6480908.
7. Wang Y, Liu KI, Sutrisnoh NB, Srinivasan H, Zhang J, Li J, et al. Systematic evaluation of CRISPR-Cas systems reveals design principles for genome editing in human cells. Genome Biol. 2018; 19(1):62. https://doi.org/10.1186/s13059-018-1445-x PMID: 29843790; PubMed Central PMCID: PMC5972437.
8. Wang DQ, Zhang CD, Wang B, Li B, Wang Q, Liu D, et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat Commun. 2019; 10. Artn 428410.1038/S41467-019-12281-8. WOS:000486566700001. https://doi.org/10.1038/s41467-019-12281-8 PMID: 31537810
9. Kim HK, Song M, Lee J, Menon AV, Jung S, Kang YM, et al. In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat Methods. 2017; 14(2):153–159. https://doi.org/10.1038/nmeth.4104 WOS:000393539700018. PMID: 27992409
10. Kleinstiver BP, Prew MS, Tsai SQ, Nguyen NT, Topkar VV, Zheng Z, et al. Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat Biotechnol. 2015; 33(12):1293–1298. https://doi.org/10.1038/nbt.3404 PMID: 26524662; PubMed Central PMCID: PMC4689141.
11. Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z, et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature. 2016; 529(7587):490–495. https://doi.org/10.1038/nature16526 PMID: 26735016; PubMed Central PMCID: PMC4851738.
12. Nishimasu H, Shi X, Ishiguro S, Gao L, Hirano S, Okazaki S, et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science. 2018; 361(6408):1259–1262. https://doi.org/10.1126/science.aas9129 PMID: 30166441; PubMed Central PMCID: PMC6368452.
13. Hu Z, Wang S, Zhang C, Gao N, Li M, Wang D, et al. A compact Cas9 ortholog from Staphylococcus auricularis (SauriCas9) expands the DNA targeting scope. PLoS Biol. 2020; 18(3):e3000686. https://doi.org/10.1371/journal.pbio.3000686 PMID: 32226015; PubMed Central PMCID: PMC7145270.

14. Hu Z, Zhang C, Wang S, Gao S, Wei J, Li M, et al. Discovery and engineering of small SlugCas9 with broad targeting range and high specificity and activity. Nucleic Acids Res. 2021; 49(7):4008–4019. https://doi.org/10.1093/nar/gkab148 PMID: 33721016; PubMed Central PMCID: PMC8053104.

15. Wang S, Mao HL, Hou LH, Hu ZY, Wang Y, Qi T, et al. Compact SchCas9 Recognizes the Simple NNGR PAM. Adv Sci. 2021. Artn 2104789 https://doi.org/10.1002/advs.202104789 WOS:000727233600001. PMID: 34874112

16. Nishimasu H, Cong L, Yan WX, Ran FA, Zetsche B, Li Y, et al. Crystal Structure of Staphylococcus aureus Cas9. Cell. 2015; 162(5):1113–1126. https://doi.org/10.1016/j.cell.2015.08.007 PMID: 26317473; PubMed Central PMCID: PMC4670267.

17. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015; 520(7546):186–191. https://doi.org/10.1038/nature14299 PMID: 25830891; PubMed Central PMCID: PMC4393360.

18. Tan Y, Chu AHY, Bao S, Hoang DA, Kebede FT, Xiong W, et al. Rationally engineered Staphylococcus aureus Cas9 nucleases with high genome-wide specificity. Proc Natl Acad Sci U S A. 2019; 116 (42):20969–20976. https://doi.org/10.1073/pnas.1906843116 PMID: 31570596; PubMed Central PMCID: PMC6800346.

19. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol. 2015; 33(2):187–197. https://doi.org/10.1038/nbt.3117 PMID: 25513782; PubMed Central PMCID: PMC4320685.

20. Hou Z, Zhang Y, Propson NE, Howden SE, Chu LF, Sontheimer EJ, et al. Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. Proc Natl Acad Sci U S A. 2013; 110(39):15644–15649. https://doi.org/10.1073/pnas.1313587110 PMID: 23940360; PubMed Central PMCID: PMC3785731.

21. Kim E, Koo T, Park SW, Kim D, Kim K, Cho HY, et al. In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. Nat Commun. 2017; 8:14500. https://doi.org/10.1038/ncomms14500 PMID: 28220790; PubMed Central PMCID: PMC5473640.

22. Teng F, Cui T, Feng G, Guo L, Xu K, Gao Q, et al. Repurposing CRISPR-Cas12b for mammalian genome engineering. Cell Discov. 2018; 4:63. https://doi.org/10.1038/s41421-018-0069-3 PMID: 30510770; PubMed Central PMCID: PMC6255809.

23. Kim DY, Lee JM, Moon SB, Chin HJ, Park S, Lim Y, et al. Efficient CRISPR editing with a hypercompact Cas12f1 and engineered guide RNAs delivered by adeno-associated virus. Nat Biotechnol. 2022; 40 (1):94–102. Epub 2021/09/04. https://doi.org/10.1038/s41587-021-01009-z PMID: 34475560; PubMed Central PMCID: PMC8763643.

24. Wu Z, Zhang Y, Yu H, Pan D, Wang Y, Wang Y, et al. Programmed genome editing by a miniature CRISPR-Cas12f nuclease. Nat Chem Biol. 2021; 17(11):1132–8. Epub 2021/09/04. https://doi.org/10.1038/s41589-021-00868-6 PMID: 34475565.

25. Gao N, Zhang C, Hu Z, Li M, Wei J, Wang Y, et al. Characterization of Brevibacillus laterosporus Cas9 (BlatCas9) for Mammalian Genome Editing. Front Cell Dev Biol. 2020; 8:583164. Epub 2020/11/17. https://doi.org/10.3389/fcell.2020.583164 PMID: 33195228; PubMed Central PMCID: PMC7604293.

26. Kim N, Kim HK, Lee S, Seo JH, Choi JW, Park J, et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. Nat Biotechnol. 2020; 38(11):1328–36. Epub 2020/06/10. https://doi.org/10.1038/s41587-020-0537-9 PMID: 32514125.

27. Chen JS, Dagdas YS, Kleinstiver BP, Welch MM, Sousa AA, Harrington LB, et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. Nature. 2017; 550(7676):407–410. https://doi.org/10.1038/nature24268 PMID: 28931002; PubMed Central PMCID: PMC5918688.

28. Gao L, Cox DBT, Yan WX, Manteiga JC, Schneider MW, Yamano T, et al. Engineered Cpf1 variants with altered PAM specificities. Nat Biotechnol. 2017; 35(8):789–792. https://doi.org/10.1038/nbt.3900 PMID: 28581492; PubMed Central PMCID: PMC5548640.

29. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell. 2016; 62(1):137–147. https://doi.org/10.1016/j.molcel.2016.02.031 PMID: 27041224; PubMed Central PMCID: PMC4826307.

30. Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX, Zhang F. Rationally engineered Cas9 nucleases with improved specificity. Science. 2016; 351(6268):84–88. https://doi.org/10.1126/science.aad5227 PMID: 26628643; PubMed Central PMCID: PMC4714946.

31. Hirano H, Gootenberg JS, Horii T, Abudayyeh OO, Kimura M, Hsu PD, et al. Structure and Engineering of Francisella novicida Cas9. Cell. 2016; 164(5):950–61. Epub 2016/02/16. https://doi.org/10.1016/j.cell.2016.01.039 PMID: 26875867; PubMed Central PMCID: PMC4899972.

32. Edraki A, Mir A, Ibraheim R, Gainetdinov I, Yoon Y, Song CQ, et al. A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. Mol Cell. 2019; 73(4):714–26 e4. https://doi.org/10.1016/j.molcel.2018.12.003 PMID: 30581144; PubMed Central PMCID: PMC6386616.

33. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004; 14(6):1188–1190. https://doi.org/10.1101/gr.849004 WOS:000221852400021 PMID: 15173120