A Novel CRISPR-Cas9 Platform with Divergent Targeting Capabilities

by

Pranam Chatterjee

S.B., Massachusetts Institute of Technology (2016)

Submitted to the Program in Media Arts and Sciences
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author .....
Program in Media Arts and Sciences
May 4, 2018

Signature redacted

Certified by .....
Joseph M. Jacobson
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Tod Machover
Academic Head, Program in Media Arts and Sciences

A Novel CRISPR-Cas9 Platform with Divergent Targeting Capabilities

by

Pranam Chatterjee

Submitted to the Program in Media Arts and Sciences
on May 4, 2018, in partial fulfillment of the
requirements for the degree of
MASTER OF SCIENCE

Abstract

RNA-guided DNA endonucleases of the CRISPR-Cas system are widely used for genome engineering and thus have numerous applications in a wide variety of fields. The range of sequences that CRISPR endonucleases can recognize, however, is constrained by the need for a specific protospacer adjacent motif (PAM) flanking the target site. In this thesis, we demonstrate the natural PAM plasticity of a highly-similar, yet previously uncharacterized, Cas9 from *Streptococcus canis* (ScCas9) through rational manipulation of distinguishing motif insertions. To this end, we report a divergent affinity to 5'-NNGT-3' PAM sequences, as well as preferences for expanded 5'-NNG-3' motifs, and demonstrate the editing capabilities of the ortholog in both bacterial and human cells. We subsequently build an automated bioinformatics pipeline, the **S**earch for **PAM**s by **AL**ignment Of **T**argets (SPAMALOT), which further explores the microbial PAM diversity of otherwise-overlooked *Streptococcus* Cas9 orthologs. Our results establish that ScCas9 can be utilized both as an alternative genome editing tool and as a functional platform to discover novel *Streptococcus* PAM specificities. Finally, we develop original machine learning-based tools to both predict the efficacy of single guide RNA (sgRNA) sequences targeting specific loci, as well as to classify and characterize the recently-discovered anti-CRISPR proteins.

Thesis Supervisor: Joseph M. Jacobson
Title: Associate Professor of Media Arts and Sciences

A Novel CRISPR-Cas9 Platform with

Divergent Targeting Capabilities

by

Pranam Chatterjee

Submitted to the Program in Media Arts and Sciences
on May 4, 2018, in partial fulfillment of the
requirements for the degree of
MASTER OF SCIENCE
at the
Massachusetts Institute of Technology

The following people served as readers for the thesis:

Signature redacted

Academic Adviser.....

Joseph M. Jacobson, PhD

Associate Professor of Media Arts and Sciences, MIT

Signature redacted

Reader .....

Kevin Esvelt, PhD

Assistant Professor of Media Arts and Sciences, MIT

Signature redacted

Reader .....

Shuguang Zhang, PhD

Principal Investigator, Center for Bits and Atoms, MIT

Acknowledgments

I would like to thank the following people who have enabled me to complete this thesis:

The Partner, Teammate, and Collaborator

First and foremost, I would especially like to give my gratitude and thanks to Noah Jakimo for his unwavering support, collaboration, inspiration, and overall mentorship. This work, and all of our projects, would not have been possible without your ingenuity, compassion, and hard work — I will be forever grateful for your role in my current and future research career.

The Lab

I give my utmost thanks to the Molecular Machines group, Joe J., Lisa, Thras, Kfir, and Maksym, for your incredibly meaningful support, both in terms of research and personal guidance. I'd also like to thank Neil, Shuguang, Rui, Fillipos, Grace, Ben, Sam, Amanda, Will, Jake, Erik, Nadya, Amira, Qiuyi, Soma, and Allan for making the CBA the best lab at MIT! Special thanks goes to Joe M., James, Ryan, Teri, Linda, and Keira (Tom, John, Jamie, Jessie, and Blaire, too!) for enabling my research and academic progress, and providing the administrative support every step of the way. And to the overall Media Lab family, thank you for accepting a hardcore bioengineer into your inventive and creative ranks.

The Core

Finally, I send all of my love to my supportive and loving family (Mommy, Baba, Mr. Johnson, and Dida), my amazing twin sister, Priyanka, and my sweet and caring Willow, for making all of my accomplishments possible. I love you all very much!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats (CRISPR) system for genome engineering has promised numerous future breakthroughs in medicine, agriculture, bioenergy, food security, nanotechnology, and a host of other applications [1][2]. Deriving from the prokaryotic adaptive immune system, CRISPR, with its associated RNA-guided Cas endonucleases, improves upon more tedious gene editing techniques, such as Zinc-Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs) which require specifically engineered proteins [3], by targeting distinct regions of DNA using a single guide RNA (sgRNA) molecule. This RNA molecule comprises of two distinct modules: one component, the tracrRNA, that allows the sgRNA to bind the Cas enzyme and another, the crRNA, that directs the targeting of a specific 20 nucleotide loci in the genome of interest [1][2][3]. Once an appropriate sgRNA is synthesized and co-expressed with the Cas endonuclease in an *in vivo* or *in vitro* setting, the Cas enzyme can generate either a double-stranded DNA break, inducing either non-homologous end joining (NHEJ), which can cause an insertion-deletion (indel) mutation to permanently silence a gene, or homology directed repair (HDR) in the presence of a donor DNA template, which can be incorporated at that site as a means of gene insertion [2]. Furthermore, catalytically inactive forms of the Cas endonuclease have been repurposed to regulate endogenous gene expression or to label specific chromosomal loci in living cells or organisms [2][4].

1.1 Motivation

While CRISPR endonucleases, such as Cas9 [2] and Cas12a [5], have proven to be versatile tools for genome editing and regulation, the range of targetable sequences is limited, however, by the need for a specific protospacer adjacent motif (PAM), which is determined by DNA-protein interactions, to immediately follow the DNA sequence specified by the sgRNA (Figure 1-1) [1][2][6][7][8][9]. For example, the most widely used variant, *Streptococcus pyogenes* Cas9 (SpCas9), requires an "NGG" motif downstream of its RNA-programmed DNA target [1][2][7][8][9]. In applications that require targeting a precise position along DNA, such as homology repair induction [10] or specific base conversion [11][12], the current sequence-limitation imposed by the small set of known PAM motifs has constrained the impact of synthetic genome engineering efforts. Thus, there is a pressing need to expand the toolkit of CRISPR endonucleases with diverse PAM sequences, so to enable the targeting of currently inaccessible genomic loci.

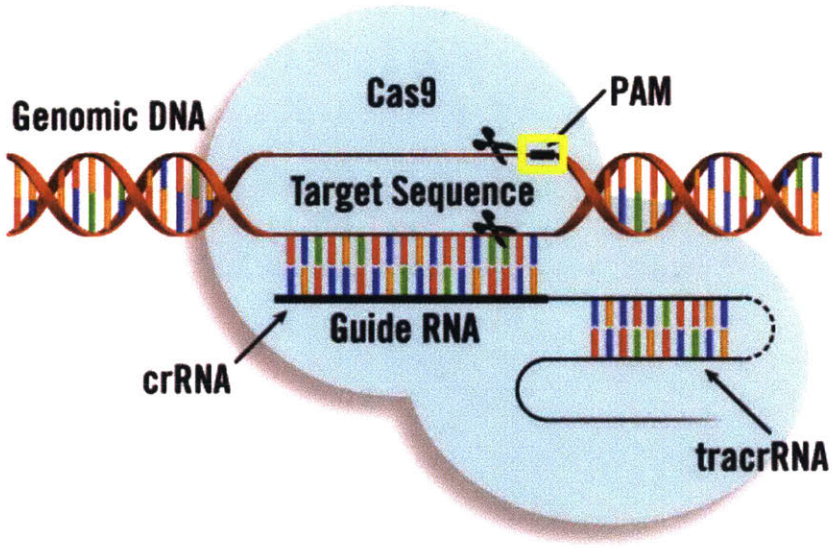

Figure 1-1: PAM requirement for CRISPR enzymes.

1.2 Related Work

To relax the PAM constraint, additional Cas9 and Cas12a variants with distinct PAM requirements have been discovered in nature or engineered to diversify the range of targetable DNA sequences. Bioinformatics tools have been developed to align CRISPR cassettes of numerous bacterial species with presumed protospacers in phage or other genomes [13][14]. This mapping helps to infer PAM sequences of naturally occurring Cas9 orthologs that possess useful properties, such as decreased size [14][15][16] and thermostability [17]. Alternatively, functionally efficient CRISPR effectors, such as SpCas9 and *Acidaminococcus sp.* Cas12a (AsCas12a), have been utilized as scaffolds for engineering to produce variants with altered PAM specificities, using methodologies such as random mutagenesis and phage-assisted continuous evolution (PACE)[18][19][20]. In total, these studies have provided only a handful of CRISPR effectors with minimal PAM requirements that enable wide targeting capabilities (Table 1-1). Combining these approaches by utilizing an analogous, and potentially more versatile, Cas9 platform with a natively short PAM sequence is thus a critical goal for developing the next generation of CRISPR tools and expanding their utility.

Enzyme	Species	PAM (5' to 3')	# Specific Bases
SpCas9	*Streptococcus pyogenes*	NGG, NGA*, NGD*	2
FnCas9	*Francisella novicida*	NGG, YG*	2
FnCas12a	*Francisella novicida*	TTN	2
AsCas12a	*Acidoaminococcus sp.*	TTTV, TYCV*, TATV*	3
SaCas9	*Staphylococcus aureus*	NNNRRT	3
CjCas9	*Campylobacter jejuni*	NNNNRYAC	4
NmeCas9	*Neisseria meningitidis*	NNNNGMTT	4
StCas9	*Streptococcus thermophilus*	NNAGAAW	5

N=Any Base, **D**=A or G or T, **Y**=C or T, **V**=A, C, or G **R**=A or G **M**=A or C **W**=A or T ***=engineered

Table 1.1: Commonly-utilized CRISPR effectors and their PAM requirements.

1.3 Contributions

To help augment the list of CRISPR effector proteins with short PAM sequences, we characterize an orthologous Cas9 protein from *Streptococcus canis*, ScCas9 (UniProt I7QXF2), possessing 89.2% sequence similarity to SpCas9. We find that despite such homology, ScCas9 prefers a distinct 5'-NNGT-3' PAM, with additional activity on certain 5'-NNG-3' sequences. To explain this divergence, we identify two significant insertions within its open reading frame (ORF) that differentiate ScCas9 from SpCas9 and contribute to its PAM-recognition flexibility. We show that ScCas9 can efficiently edit genomic DNA in mammalian cells, and construct bioinformatics pipelines to explore the PAM specificities of other *Streptococcus* orthologs and to expand the functionality and utility of other CRISPR proteins. Together, we anticipate that the development of these novel CRISPR technologies, combining both experimental and computational methodologies, will be welcome additions to the evergrowing genome engineering toolkit.

Chapter 2

Characterization of ScCas9

2.1 Overview

This chapter describes the identification and characterization of the single-effector Cas9 endonuclease from *Streptococcus canis* across both *in silico* and bacterial contexts. Harnessing both sequence and structural information of ScCas9, we bioinformatically predict its putative PAM sequence and subsequently assess its PAM binding preference using a fluorescence-based assay within *E. coli* cells. The chapter concludes with a putative binding sequence for ScCas9, as well as an examination of the role of specific sequence elements within its open reading frame (ORF).

2.2 *In silico* Characterization of ScCas9

While numerous Cas9 homologs have been sequenced, only a handful of *Streptococcus* orthologs have been characterized or functionally validated. To explore this space, we curated all *Streptococcus* Cas9 protein sequences from UniProt [21], performed global pairwise alignments using the BLOSUM62 scoring matrix [22], and calculated percent sequence homology to SpCas9. From them, the Cas9 from *Streptococcus canis* (ScCas9) stood out, not only due to its remarkable sequence homology (89.2%) to SpCas9, but also because of a positive-charged insertion of 10 amino acids within the highly-conserved REC3 domain, in positions 367-376 (Figure 2-1).

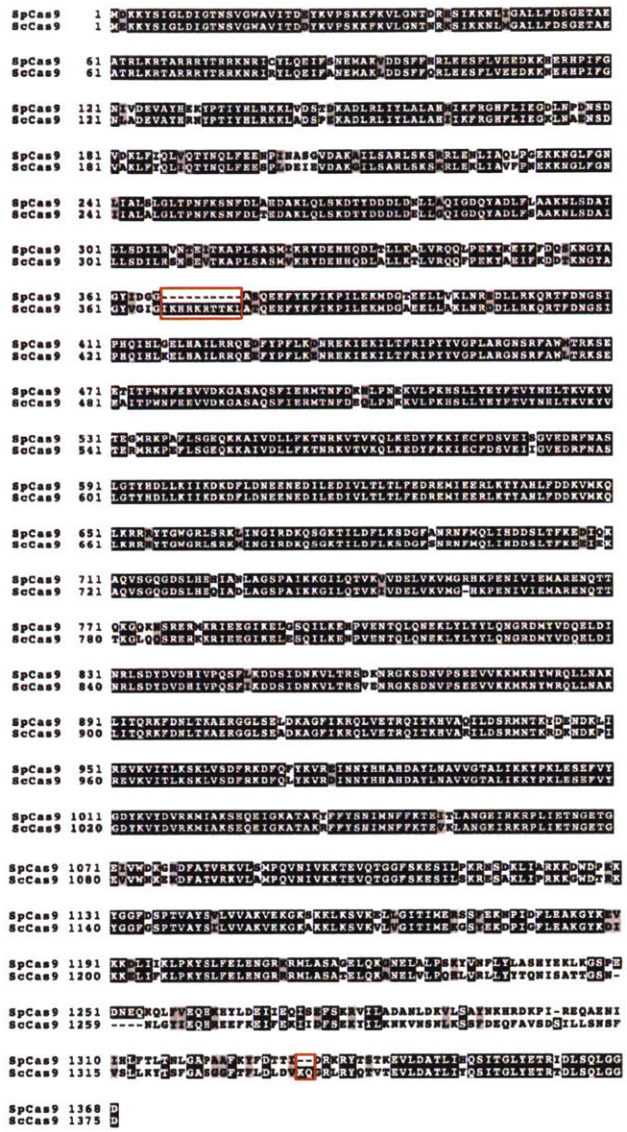

Figure 2-1: Global pairwise sequence alignment of SpCas9 and ScCas9.

Exploiting both of these properties, we modeled the insertion within the corresponding domain of PDB 4OO8 [23] and, when viewed in PyMol, noticed that it formed a "loop"-like structure, of which several of its positive-charged residues come in close proximity with the target DNA near the PAM (Figure 2-2).

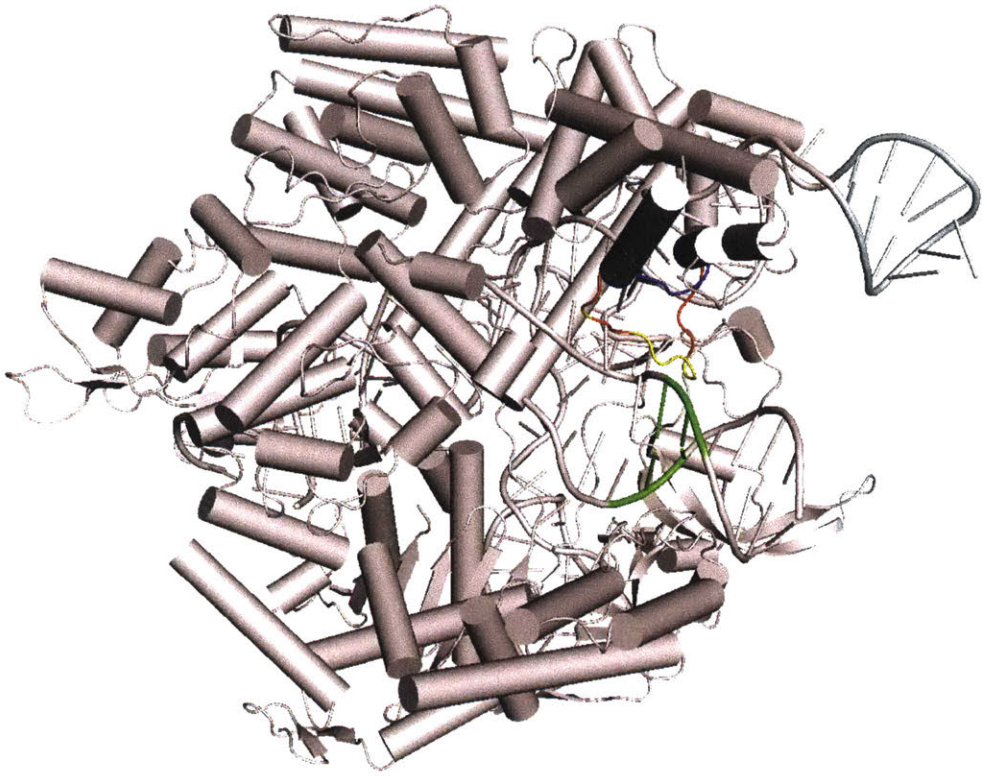

Figure 2-2: Insertion of novel REC motif into PDB 4OO8 [18]. The 367-376 insertion demonstrates a loop-like structure (red). Several of its positive-charged residues (yellow) come in close proximity to the target DNA near the PAM (green).

We further identified an additional insertion of two amino acids (KQ) immediately upstream of the two critical arginine residues necessary for PAM binding [24], in positions 1337-1338 (Figure 2-1). We thus hypothesized that these insertions may affect the PAM specificity of this enzyme. To support this prediction, we computationally characterized the PAM for ScCas9, by first mapping spacer sequences from the Cas9-associated type II CRISPR loci in the *Streptococcus canis* genome [25] to viral and plasmid genomes using BLAST [26], extracting the sequences 3' to the mapped protospacers, and subsequently generating a WebLogo [27] representation of the aligned PAM sequences. Our analysis suggested an 5'-NNGTT-3' PAM (Figure 2-3). Intrigued by these novel motifs and motivated by its predicted, divergent PAM, we selected ScCas9 as a candidate for further PAM characterization and engineering.

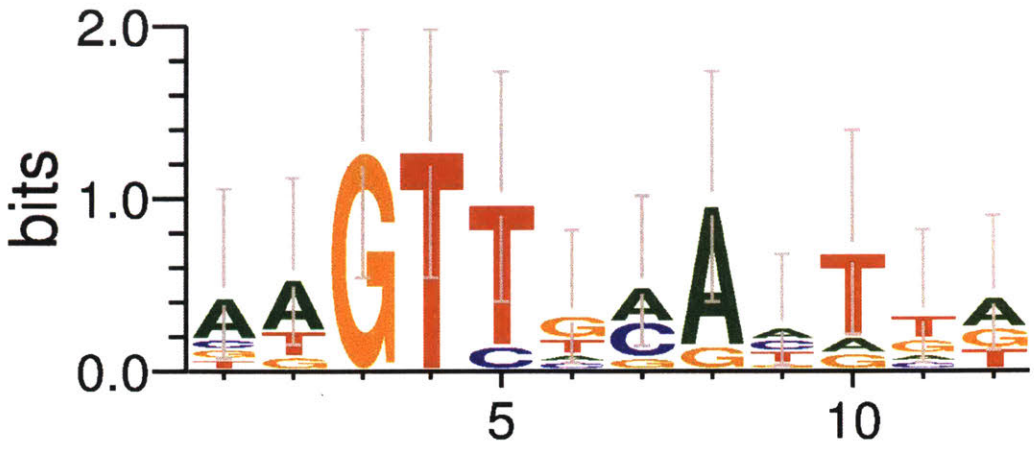

Figure 2-3: WebLogo [27] for sequences found at the 3' end of protospacer targets identified in plasmid and viral genomes using Type II spacer sequences within *Streptococcus canis* as BLAST [26] queries.

2.3 Determination of PAM Sequences Recognized by ScCas9

Due to the the relatively low number of protospacer targets, we first validated the PAM binding sequence of ScCas9 utilizing an existent positive selection bacterial screen based on green fluorescent protein (GFP) expression conditioned on PAM binding, termed PAM-SCANR [28]. A plasmid library containing the target sequence followed by a randomized 5'-CNNNNC-3' PAM sequence was bound by a nuclease-deficient ScCas9 (and dSpCas9 as a control) and an sgRNA both specific to the target sequence and general for SpCas9 and ScCas9, allowing for the repression of *lacI* and expression of GFP. Plasmid DNA from FACS-sorted GFP-positive cells and pre-sorted cells were extracted and amplified, and enriched PAM sequences were identified by Sanger sequencing. Our results provide initial evidence that ScCas9 can bind to a more minimal 5'-NNGT-3' PAM, distinct to that of SpCas9's 5'-NGG-3' (Figure 2-4).

We hypothesized that the previously described insertions may contribute to this flexibility, and thus engineered ScCas9 to remove either insertion or both, and subjected these variants to the same screen. Only removing the loop (ScCas9 Δ367-376 or ScCas9 ΔLoop) extended the PAM of ScCas9 to 5'-NAGT-3', while only removing the KQ insertion (ScCas9 Δ1337-1338 or ScCas9 ΔKQ), reverted the PAM specificity to a more 5'-NGG-3'-like PAM with minimal requirements for T at position 4 (Figure 2A). Finally, the most SpCas9-like variant, where both insertions are removed (ScCas9 Δ367-376 Δ1337-1338 or ScCas9 ΔLoop ΔKQ), showed a strong preference for G in position 3 while also exhibiting reduced affinity for T at position 4 (Figure 2-4).

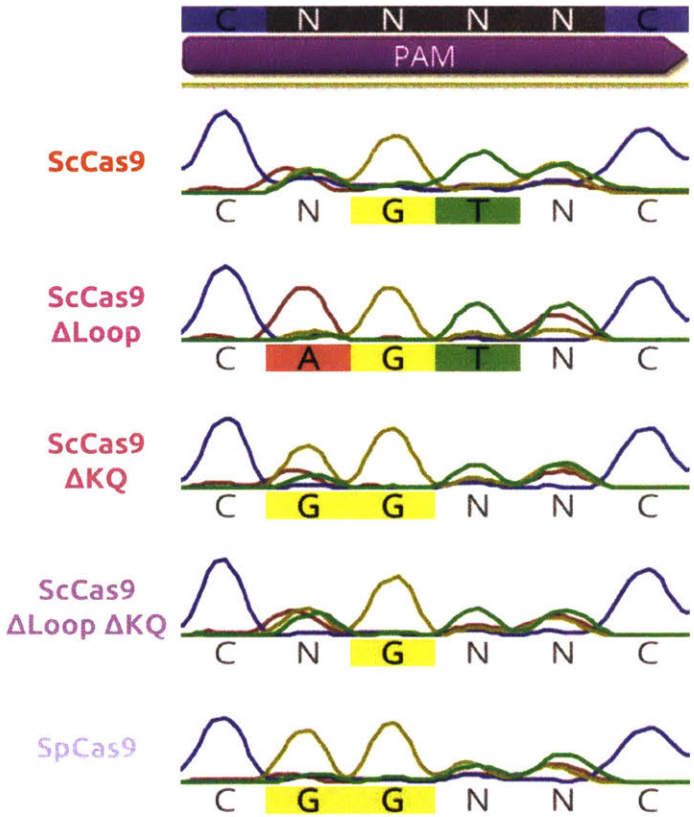

Figure 2-4: PAM Determination of ScCas9 Variants. GFP binding enrichment on a 5'-CNNNNC-3' PAM library.

To confirm the results of the library assay, we decided to elucidate the minimal PAM requirements of ScCas9 and ScCas9 ΔLoop ΔKQ by utilizing fixed PAM sequences. We replaced the PAM library with individual PAM sequences, which were varied at positions 2 and 5 to test each possible base. Our results demonstrate that while ScCas9 exhibits a clear 5'-NNGTN-3' preference, with activity for all bases at both positions, ScCas9 ΔLoop ΔKQ demonstrates significant binding at 5'-NGG-3' PAM sequences and at some, but not all, 5'-NNGTN-3' motifs, indicating an intermediate PAM specificity between that of SpCas9 and ScCas9 (Figure 2-5).

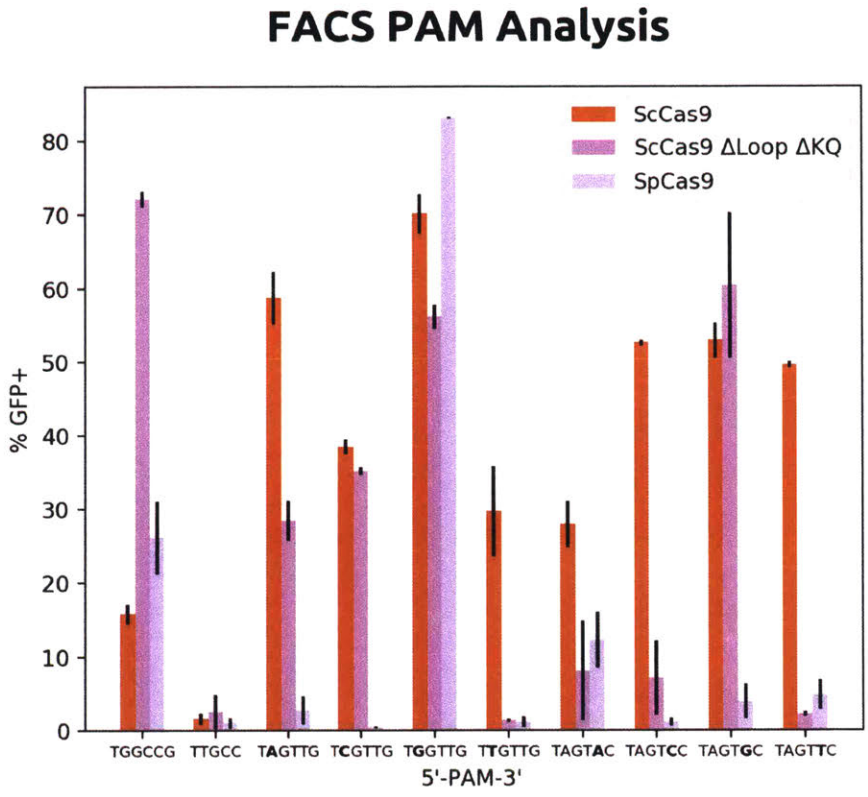

Figure 2-5: Examination of PAM preference for ScCas9. For individual PAMs, we varied a single position (2 and 5) to test each possible base. All samples were electroporated in duplicates, subjected to FACS analysis, and gated for GFP expression. Subsequently, GFP expression levels were averaged. Standard deviation was used to calculate error bars.

2.4 Expanded PAM Specificity of ScCas9

To fully assess the scope of PAM sequences recognized by ScCas9, we constructed a plasmid library containing the target sequence followed by a fully-randomized 5'-NNNNNNNN-3' (8N) PAM sequence. Employing the PAM-SCANR assay, we demonstrate that ScCas9 is able to bind to most 5'-NNG-3' PAM sequences (Figure 2-6), similar to the newly-engineered xCas9(3.7) obtained by PACE [19]. ScCas9 ΔLoop exhibits a stronger preference for A and a weaker preference for G at position 2, while ScCas9 ΔKQ demonstrates a strong preference for G at position 2 with a weak affinity to A (Figure 2-6). The most SpCas9-like variant, ScCas9 ΔLoop ΔKQ, corroborates the 5'-NGG-3' PAM specificity of SpCas9 (Figure 2-6). Thus, from a functional perspective, these insertions operate in tandem to reduce the specificity of ScCas9 to its more minimal 5'-NNG-3' PAM, in addition to its affinity toward 5'-NNGT-3'.

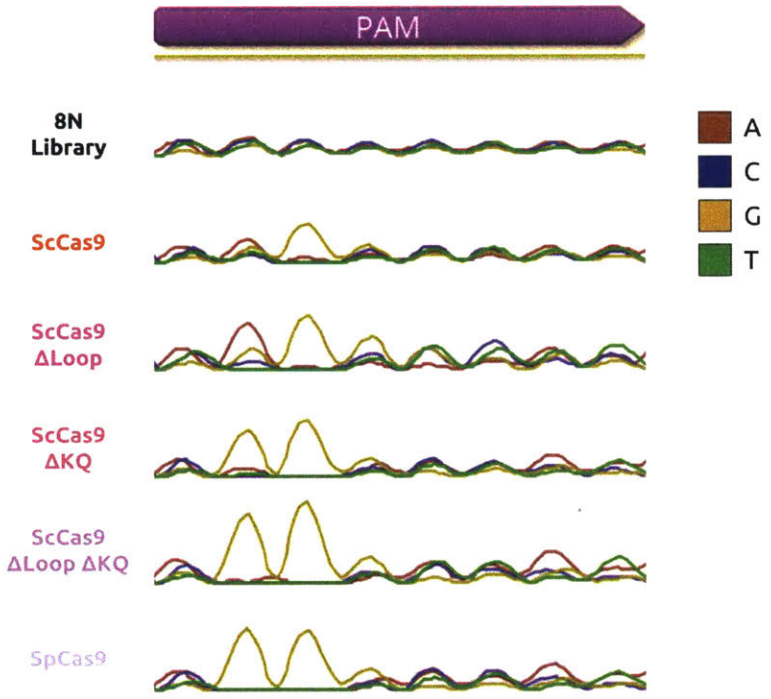

Figure 2-6: PAM Determination of ScCas9 Variants. GFP binding enrichment on a 5'-NNNNNNNN-3' PAM library.

2.5 Methodologies

2.5.1 Identification of SpCas9 Homologs and Generation of Plasmids

The UniProt database [21] was mined for all *Streptococcus* Cas9 protein sequences, which were used as inputs to either the BioPython *pairwise2* module or Geneious to conduct global pairwise alignments with SpCas9, using the BLOSUM62 scoring matrix [22], and subsequently calculate percent homology. The Cas9 from *Streptococcus canis* was codon optimized for *E. Coli*, ordered as gBlocks from Integrated DNA Technologies (IDT), and assembled using Golden Gate Assembly. Engineering of the coding sequence of ScCas9 was conducted using either the Q5 Site-Directed Mutagenesis Kit (NEB) or Gibson Assembly. Plasmid backbones for expression in alternate contexts were manipulated to individually insert the ORFs of SpCas9, ScCas9 variants, or the sgRNA targeting sequence using Gibson or Golden Gate Assembly.

2.5.2 PAM-SCANR Assay

Plasmids for the SpCas9 sgRNA and PAM-SCANR genetic circuit, as well as BW25113 ΔlacI cells, were generously provided by the Beisel Lab (North Carolina State University). Plasmid libraries containing the target sequence followed by either a fully-randomized 8-bp 5'-NNNNNNNN-3' library, a more constrained 4-bp 5'-CNNNNC-3' library, or fixed PAM sequences were constructed by conducting site-directed mutagenesis on the PAM-SCANR plasmid. Nuclease-deficient mutations were introduced to the Cas9 variants using Gibson Assembly. The provided BW25113 cells were made electrocompetent using standard glycerol wash and resuspension protocols. The PAM library and sgRNA plasmids, with resistance to kanamycin (Kan) and carbenicillin (Crb) respectively, were co-electroporated into the electrocompetent cells at 2.4 kV, outgrown, and recovered in Kan+Crb Luria Broth (LB) media overnight. The outgrowth was diluted 1:100, grown to ABS600 of 0.6 in Kan+Crb LB liquid media, and made electrocompetent. Indicated Cas9 plasmids, with resistance to chloramphenicol

(Chl), was electroporated in duplicates into the electrocompetent cells harboring both the dCas9 and sgRNA plasmids, outgrown, and collected in 5 mL Kan+Crb+Chl LB media. Overnight cultures were diluted to an ABS600 of 0.01 and cultured to an OD600 of 0.2. Cultures were analyzed and sorted on a FACSAria machine (Becton Dickinson). Events were gated based on forward scatter and side scatter and fluorescence was measured in the FITC channel (488 nm laser for excitation, 530/30 filter for detection), with at least 30,000 gated events for data analysis. Sorted GFP-positive cells were grown to sufficient density, and plasmids from the pre-sorted and sorted populations were then isolated, and the region flanking the nucleotide library was PCR amplified and submitted for Sanger sequencing (Genewiz). Bacteria harboring non-library PAM plasmids, performed in duplicates, were analyzed by FACS analysis following electroporation and overnight incubation, and represented as the percent of GFP-positive cells in the population, utilizing standard deviation to calculate error bars. Additional details on the PAM-SCANR assay can be found in Leenay, et al. [28].

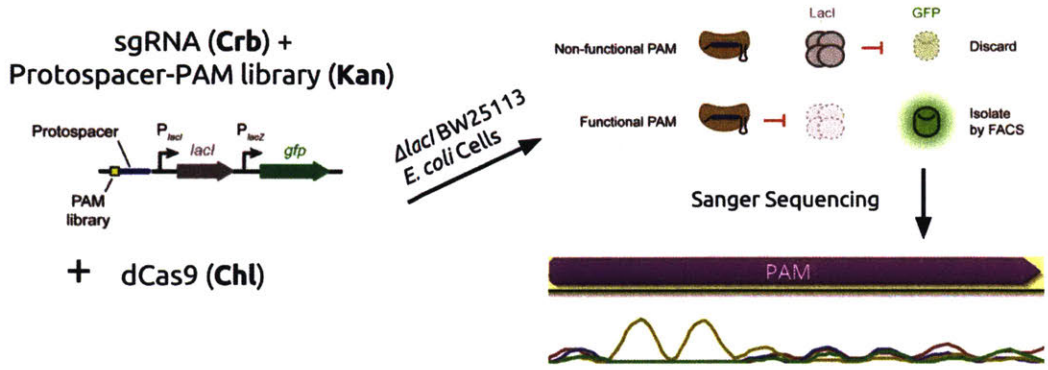

Figure 2-7: PAM-SCANR Schematic. ΔlacI *E. coli* cells are transformed with sgRNA and PAM library plasmids [28], followed by a dCas9 variant, each with indicated antibiotic resistance. GFP-positive cells are collected by FACS, PAM libraries are amplified by PCR, and are subjected to Sanger sequencing, producing chromatograms as readouts.

Chapter 3

Genome Editing by ScCas9 in Human Cells

3.1 Overview

This chapter describes the introduction of the ScCas9 machinery into mammalian cell lines to edit both endogenous genomic loci and synthetic plasmid targets. To this end, we co-transfect guide RNA plasmids and Cas9 plasmids into human embryonic kidney (HEK293T) cells and utilize indel analysis or FACS analysis to measure either cleavage or base editing outcomes, respectively. The focus of this chapter is to demonstrate the targeting of previously inaccessible loci utilizing ScCas9 in human cell contexts.

3.2 Indel Analysis of ScCas9 Variants in HEK293T Cells

We assessed the ability of ScCas9 and ScCas9 ΔLoop ΔKQ to edit mammalian genomes by co-transfecting HEK293T cells with plasmids constitutively expressing these variants along with sgRNAs directed to sites with varying PAM sequences within a native genomic locus (VEGFA). We first tested editing efficiency at a site containing an overlapping PAM (5'-GGGT-3'). After 48 hours post-transfection, mutation

rates detected by the T7 endonuclease I (T7E1) assay demonstrated comparable editing activities of SpCas9, ScCas9, and ScCas9 ΔLoop ΔKQ (Figure 3-1A and 3-2). Additionally, we constructed sgRNAs to endogenous VEGFA sites with various non-overlapping 5'-NNGT-3' PAM sequences (Figure 3-1A), iterating through all bases at position 2. Other than at the well-described weakly-preferred 5'-NAG-3' PAM sequence [29], SpCas9's ability to form indels was abrogated to background levels for other non-overlapping 5'-NNGT-3' sequences (Figure 3-1A and 3-2). Alternatively, ScCas9 maintained detectable activity (Figure 3-1A and 3-2), as expected from the results of the PAM-SCANR assay.

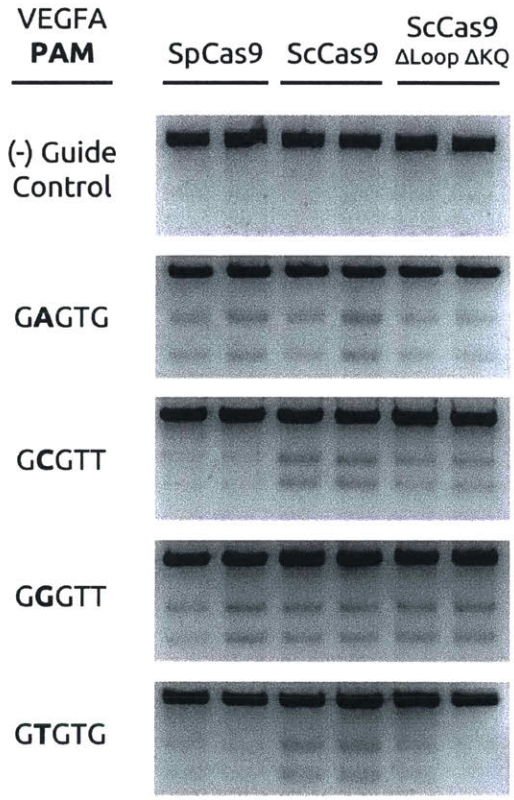

Figure 3-1: T7E1 analysis of indels produced at VEGFA loci with indicated 5'-NNGT-3' PAM sequences. The Cas9 used is indicated above each lane. All samples were performed in biological duplicates. As a background control, SpCas9, ScCas9, and ScCas9 ΔLoop ΔKQ were transfected without targeting guide RNA vectors.

Consistent with the bacterial data, ScCas9 ΔLoop ΔKQ was able to generate indels at most, but not all, non-overlapping 5'-NNGT-3' sites (Figure 3-1 and 3-2).

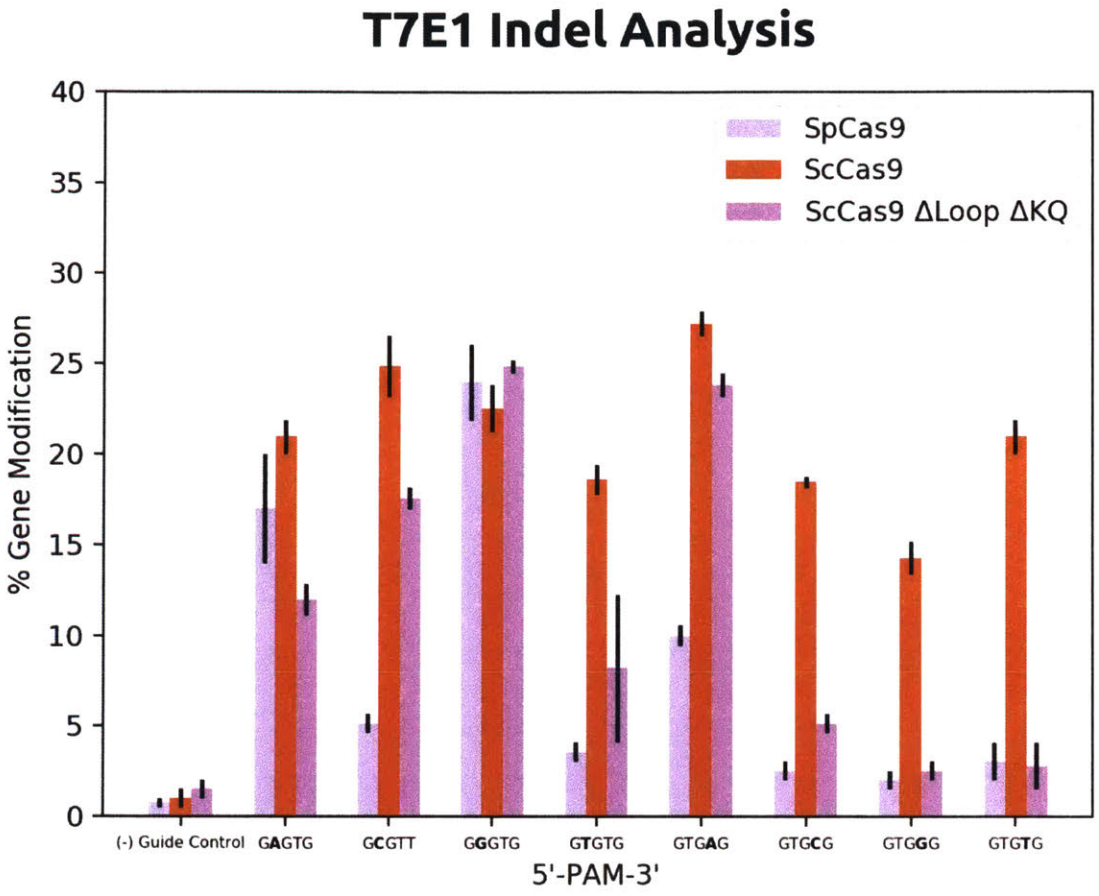

Figure 3-2: Quantitative analysis of T7E1 products. Unprocessed gel images were quantified by line scan analysis using Fiji [30], the total intensity of cleaved bands were calculated as a fraction of total product, and percent gene modification was calculated. All samples were performed in duplicates and quantified modification values were averaged. Standard deviation was used to calculate error bars.

These results verify that ScCas9 can serve as an effective alternative to SpCas9 for genome editing in mammalian cells, both at overlapping 5'-NGGT-3' and non-overlapping 5'-NNGT-3' PAM sequences, while also confirming ScCas9 ΔLoop ΔKQ's intermediate PAM specificity.

Finally, to fully examine ScCas9's possible 5'-NNG-3' cleavage specificity in human cells, we targeted endogenous VEGFA loci with PAM sequences that iterate through all four bases at position 4, with positions 1, 2, 3, and 5 being held constant. T7E1 analyses on these loci confirmed that ScCas9 can cleave all examined 5'-NNGN-3' targets (Figure 3-1, 3-2, and 3-3). Thus, to this point of validation, ScCas9 displays strong affinity to 5'-NNGT-3' PAM sequences, and can also recognize and cleave various 5'-NNGN-3' targets, exhibiting broader PAM specificity than SpCas9, and similar to that of xCas9(3.7) [19], while potentially offering a greater than 2-fold advantage on 5'-NNGC-3' targets (Figure 3-2). Future side-by-side analysis is necessary to establish the complete set of non-overlapping PAM specificities of ScCas9 with xCas9(3.7).

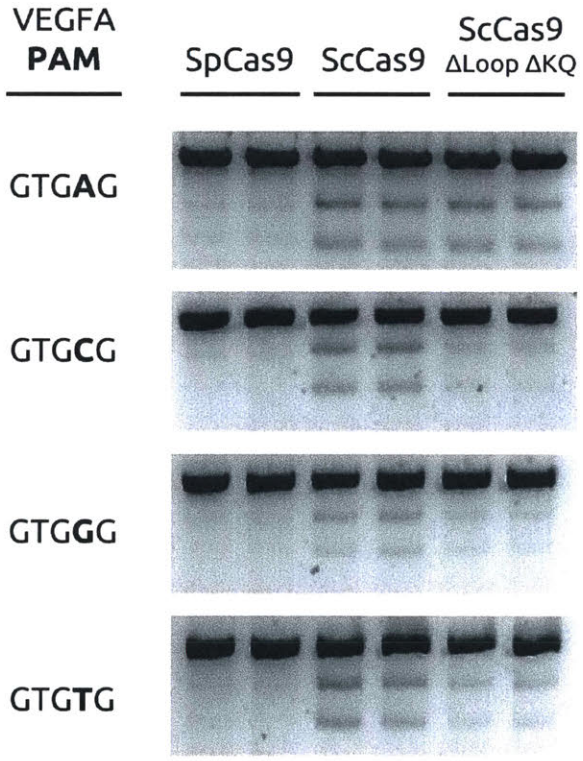

Figure 3-3: T7E1 analysis of indels produced at VEGFA loci with indicated 5'-NNGN-3' PAM sequences. The Cas9 used is indicated above each lane. All samples were performed in biological duplicates.

3.3 Base Editing by ScCas9 in HEK293T Cells

We further assessed the base editing capabilities of ScCas9 and ScCas9 ΔLoop ΔKQ using a synthetic Traffic Light Reporter (TLR) [31] plasmid, containing an early stop codon upstream of a GFP ORF and downstream of an mCherry ORF. Successful A→G base editing using the ABE7.10 architecture, as described in Gaudelli, et al. [12], will convert a TAG stop codon to a TGG tryptophan codon, thus restoring GFP expression. After gating cells based on mCherry expression, we observed ~35% base editing efficiency at an 5'-AAGT-3' PAM sequence for ScCas9, as compared to only ~15% for the standard SpCas9 architecture (Figure 3-4), thus suggesting ScCas9's improved base editing capabilities at certain, non-overlapping PAM sequences.

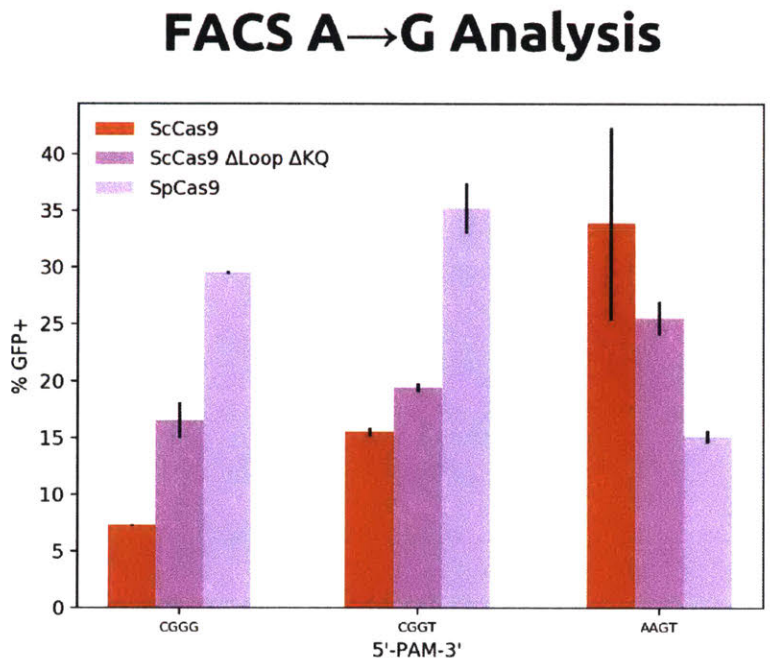

Figure 3-4: FACS Analysis of A→G base editing outcomes with ScCas9 variants. Cells expressing mCherry were gated and percent GFP calculation of the subset were calculated. All samples were performed in duplicates and quantified expression percentages were averaged. Standard deviation was used to calculate error bars.

3.4 Methodologies

3.4.1 Cell Culture and Indel Analysis

HEK293T cells were maintained in DMEM supplemented with 100 units/ml penicillin, 100 mg/ml streptomycin, and 10% fetal bovine serum (FBS). sgRNA plasmid (500 ng) and Cas9 plasmid (500 ng) were transfected into cells as duplicates (2 x 10^5/well in a 24-well plate) with Lipofectamine 2000 (Invitrogen) in Opti-MEM (Gibco). After 48 hours post-transfection, genomic DNA was extracted using QuickExtract Solution (Epicentre), and VEGFA loci were amplified by PCR. The T7E1 reaction was conducted according to the manufacturer's instructions and equal concentration of products were analyzed on a 2% agarose gel stained with SYBR Safe (Thermo Fisher Scientific). Unprocessed gel image files were analyzed in Fiji [30]. The cleaved bands of interest were isolated using the rectangle tool, and the areas under the corresponding peaks were measured and calculated as the fraction cleaved of the total product. Percent gene modification was calculated as follows:

$$\% \text{ gene modification} = 100 \text{ x } (1 - (1 - \text{ fraction cleaved})^{\frac{1}{2}})$$

All samples were performed in duplicates and percent gene modifications were averaged. Standard deviation was used to calculate error bars.

3.4.2 Cell Culture and Base Editing Analysis

HEK293T cells were maintained as previously described, and transfected with the corresponding sgRNA plasmids (333 ng), ABE7.10 plasmids (Addgene) (333 ng), and synthetically constructed TLR plasmids (333 ng) were transfected into cells as duplicates (2 x 10^5/well in a 24-well plate) with Lipofectamine 2000 (Invitrogen) in Opti-MEM (Gibco). After 72 hours post-transfection, cells were harvested and analyzed on a FACSCelesta machine (Becton Dickinson) for mCherry (561 nm laser excitation) and GFP (488 nm laser excitation) fluorescence. Cells expressing mCherry were gated and percent GFP calculation of the subset were calculated.

Chapter 4

Computational Methods for CRISPR Discovery and Utility

4.1 Overview

In the culminating chapter, several computational tools are showcased that either enable the discovery of novel CRISPR proteins or enhance current CRISPR technologies. Specifically, we present a PAM prediction pipeline for *Streptococcus* Cas9 orthologs, a target activity prediction software for SpCas9-based sgRNA selection, and a Support Vector Machine (SVM)-based classifier for the characterization of novel anti-CRISPR proteins.

4.2 Genus-wide Prediction of Divergent *Streptococcus* Cas9 PAMs

Demonstrations of efficient genome editing by Cas9 nucleases with distinct PAM specificity from several *Streptococcus* species, including *S. canis*, motivated us to develop a bioinformatics pipeline for discovering additional Cas9 proteins with novel PAM requirements in the *Streptococcus* genus. We call this method the **S**earch for **PAM**s by **AL**ignment **O**f **T**argets (**SPAMALOT**). Briefly, we mapped a 20 nt portion of

spacers flanked by known *Streptococcus* repeat sequences to candidate protospacers that align with no more than two mismatches in phages associated with the genus [32]. We grouped 12 nt protospacer 3'-adjacent sequences from each alignment by genome and CRISPR repeat, and then generated group WebLogos [27] to compute presumed PAM features.

Figure 4-1A shows that resulting WebLogos accurately reflect the known PAM specificities of Cas9 from *S. canis* (this work), *S. pyogenes*, *S. thermophilus*, and *S. mutans* [7, 33, 34]. We identified a notable diversity in the WebLogo plots derived from various *S. thermophilus* cassettes with common repeat sequences (Figure 4-1B), each of which could originate from any other such *S. thermophilus* WebLogo upon subtle specificity changes that traverse intermediate WebLogos among them. We observe a similar relationship between two *S. oralis* WebLogos that also share this repeat, as well as unique putative PAM specificities associated with CRISPR cassettes containing *S. mutans*-like repeats from the *S. oralis*, *S. equinis*, and *S. pseudopneumoniae* genomes (Figure 4-1C).

4.2.1 SPAMALOT Pipeline

All 11,440 *Streptococcus* bacterial and 53 *Streptococcus* associated phage genomes were downloaded from NCBI. CRISPR repeats catalogued for the genus were downloaded from CRISPRdb hosted by University of Paris-Sud [35]. For each genome, spacers upstream of a specific repeat sequence were collected with a toolchain consisting of the fast and memory-efficient Bowtie 2 alignment [36]. Each genome and repeat-type specific collection of spacers were then matched to all phage genomes using the original Bowtie short-sequence alignment tool [37] to identify candidate protospacers with at most one, two, or no mismatches. Unique candidates were input into the WebLogo 3 [27] command line tool for prediction of PAM features.

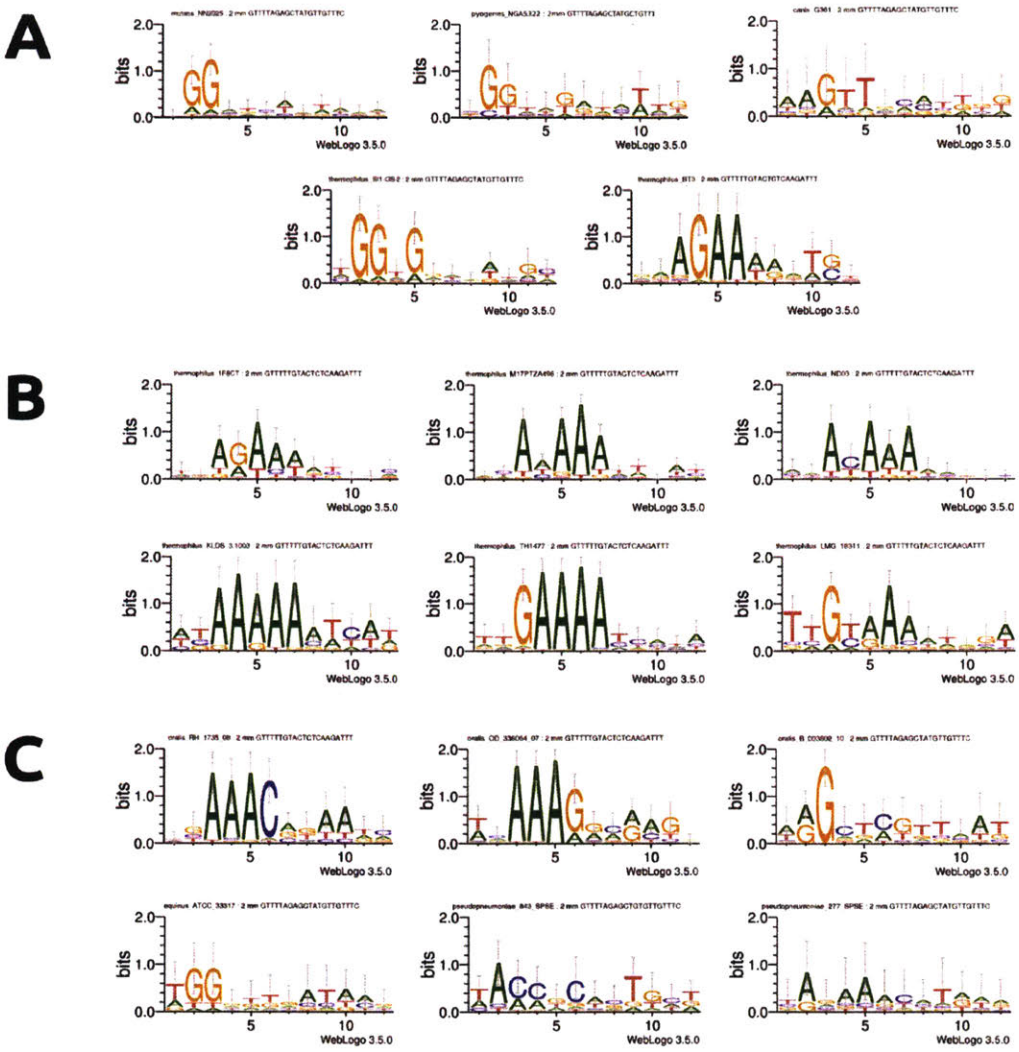

Figure 4-1: SPAMALOT PAM Predictions for *Streptococcus* Cas9 Orthologs. Spacer sequences found within the Type II CRISPR cassettes associated with Cas9 ORFs from specified *Streptococcus* genomes were aligned to *Streptococcus phage* genomes to generate spacer-protospacer mappings. WebLogos [27], labeled with the relevant species, genome, and CRISPR repeat, were generated for sequences found at the 3' end of candidate protospacer targets with no more than two mismatches (2mm). A) PAM predictions for experimentally validated Cas9 PAM sequences in previous studies. SPAMALOT correctly predicts the PAM of experimentally characterized Cas9 enzymes from the specified *S. mutans*, *S. pyogenes*, *S. canis*, and *S. thermophilus* genomes [7, 33, 34]. B) Novel PAM predictions of alternate *S. thermophilus* Cas9 orthologs with putative divergent specificities. C) Novel PAM predictions of uncharacterized *Streptococcus* orthologs with distinct specificities.

4.3 A Deep Learning Model for Predicting CRISPR sgRNA Performance

With the growing usage of CRISPR in a variety of fields [2], there is a pressing need for effective *in silico* prediction softwares that allow users to select sgRNA sequences that maximize on-target activity, thus leading to efficient gene disruption, insertion, regulation, and modification outcomes. Recently, "deep learning" has achieved record-breaking performance in a variety of information technology applications [38]. Here, we develop **CRISPRedict**, a suite of deep learning methods to predict and classify the effect of an sgRNA given its sequence. We utilize library-on-library sgRNA activity data [39] to develop an artificial neural network architecture based on sequence alone. We then use the generated model to predict on-target efficacies of held-out "test" sgRNA sequences by utilizing a comprehensive human sgRNA database acquired through high-throughput phenotypic screens [40].

4.3.1 Prediction of Mutation Rate from sgRNA Sequence

Using Keras, a minimalist, highly modular neural networks library written in Python [41], We trained three separate model architectures for the regression task of predicting mutation rate from sgRNA sequence: a model with only fully connected layers (FCNN), a model with convolutional layers (CNN), and one with recurrent LSTM layers (LSTM). Hyperas [42] was utilized to optimize the hyperparameters of each model architecture, and the best model for each was selected based on the mean-squared error (MSE) loss on the test data after training using the self-weighted MSE loss function. Our results demonstrate that a two-layered fully connected layer, with the number of units in the first layer equivalent to the number of sequences in the training set and 810 units in the second layer, followed by a final output layer, proved to have the lowest loss value (Table 4-1).

FCNN

Hyperparameter	Value
# of FC Layers	2
Layer 1 Units	917
Layer 2 Units	810

Training MSE: 0.9435
Testing MSE: 1.0112

CNN

Hyperparameter	Value
# of Conv Layers	1
Filter Size	4
l2-regularization	0.14
Dropout	0.46
Activation	Sigmoid

Training MSE: 0.9134
Testing MSE: 1.2604

LSTM

Hyperparameter	Value
# of LSTM Layers	1
Layer 1 Units	900
Activation	Sigmoid

Training MSE: 1.2613
Testing MSE: 1.2771

Table 4.1: Optimized Hyperparameter Settings and Results for Mutation Rate Prediction. Hyperparameter values for each optimized architecture (FCNN, CNN, LSTM) along with the MSE-loss values on both the training and validation datasets are shown.

With an optimized model chosen, we next applied held-out test data to visualize its performance.

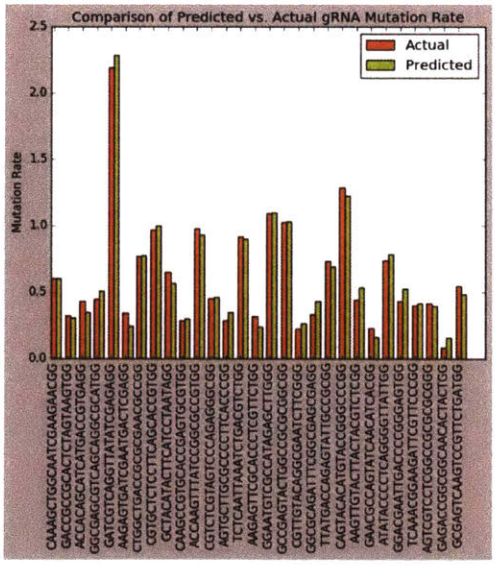

Figure 4-2: Model predictive efficacy. 30 example test sgRNA sequences are shown comparing predicted mutation rates with experimentally-observed mutation rates.

As this is a regression problem, we compared the predicted mutation rate given by the model against the actual experimentally-validated mutation rate. The visualization demonstrates, for the subsampling of the sgRNA sequences shown, the effective predictive ability of the chosen model (Figure 4-2).

4.3.2 Classification of sgRNA Sequences

For the task of classifying sgRNA sequences based on their efficacy, we initially thresholded the sgRNA sequences based on a biologically relevant mutation rate (1.5), given the library-on-library experimental setup [39]. This relatively low threshold derives from the fact that bacterial genes within plasmids were transfected into human HEK293T cells, and targeted by a corresponding library of sgRNA sequences, thus yielding the relatively low efficiencies. This preprocessing step generated 300 positive sequences and 934 negative sequences. Due to the imbalance of the dataset favoring underperforming sgRNA sequences, we trained and optimized two model architectures, and selected the optimal model based on area under the receiver-operating characteristic (AUROC) curve, which, among other interpretations, indicates the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example [38].

FCNN

Hyperparameter	Value
# of FC Layers	3
Layer 1 Units	917
Layer 2 Units	800
Layer 3 Units	300

Testing AUC: 0.792

CNN

Hyperparameter	Value
# of Conv Layers	1
Filter Size	3
l2-regularization	0.2
Dropout	0.5

Testing AUC: 0.714

Table 4.2: Optimized Hyperparameter Settings and Results of sgRNA Classification. Hyperparameter values for each optimized architecture (FCNN and CNN). The optimal test-set AUROC values are shown below the corresponding tables.

Our results demonstrate that a 3-layered FCNN outperforms the optimal CNN with 1 layer, and yields a 0.792 AUROC, indicating effective classification performance (Table 4-2 and Figure 4-3), though lower than the 0.92 AUROC of the state-of-the-art WU-CRISPR SVM model, which coalesces numerous heterogeneous features characteristic of highly active sgRNAs [43].

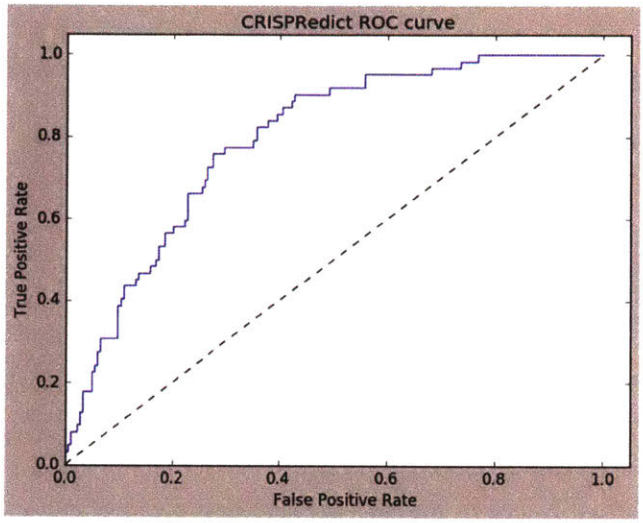

Figure 4-3: CRISPRedict ROC Curve. The ROC curve of the optimized 3-layer FCNN. The AUROC score of 0.792 was calculated by integration under the curve through the use of the scikit-learn library.

4.4 A Binary Classifier for Anti-CRISPR Prediction

While the function of Cas9 endonucleases is well documented [1][2][3][7][8], recently, a novel class of small CRISPR-interacting proteins, known as anti-CRISPRs, have been discovered [44]. Anti-CRISPRs offer a powerful defense system that helps phages to escape injury from the CRISPR-Cas system, by utilizing a variety of techniques to prevent Cas9-mediated DNA binding, unwinding, and subsequent cleavage [44][45]. Furthermore, anti-CRISPRs have been shown to reduce off-target events in both research and therapeutic applications [44][45].

4.4.1 SVM Model Training for Binary Classification

We generated a dataset containing 433 entries for anti-CRISPR proteins, acquired from anti-CRISPRdb [46], along with random "negative" protein sequences of various properties to the datasets, employing standard class-balancing techniques. We constructed a linear SVM classifier to generate a hyperplane with furthest distance from the nearest datapoints of each of two classes, anti-CRISPRs and non-anti-CRISPRs, respectively. To represent the data, we constructed frequency vectors for each sequence, which holds the count total of all unique 4-mers in the entire sequence space at each position in the vector. After splitting and shuffling the input vectors into training and test sets, we utilized a linear kernel to calculate the dot product of two frequency vectors and trained the models. To assess the model's performance, we calculated both the accuracy and AUROC metrics on held-out test set data for anti-CRISPR SVM model. Our model achieved a 99.1% classification accuracy and a 0.989 AUROC (Figure 4-4) — to our knowledge, the most accurate anti-CRISPR predictor available.

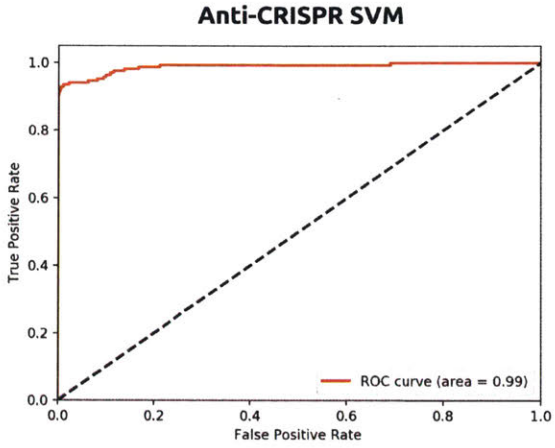

Figure 4-4: ROC Curve for Binary Classification of Anti-CRISPRs. The trained models for anti-CRISPRs was evaluated on held-out test data using the AUROC metric, calculating the true positive rate vs. false positive rate between predicted output values and expected output values.

4.4.2 Identification of Charge-Dependency on Binary Classification of Cas9 and Anti-CRISPR Proteins

With a highly accurate binary classifier for anti-CRISPRs, we next sought to identify sequence-based features that contribute to anti-CRISPR identity. Using intuition from experimental protein engineering practices, we first conducted a sliding-window alanine (A) scan of different motif lengths over the entire sequence for each anti-CRISPR in our database to attempt to break anti-CRISPR identification. This method yielded negligible results, thus indicating that each SVM-based model was learning overall features rather than overfitting on single motifs.

As a proof of concept, using similar methodologies as for our anti-CRISPR model, we trained a Cas9 binary classifier that exhibited an AUROC of 0.999 on held-out test data. Due to Cas9's affinity for binding and cleaving DNA, we hypothesized that converting positive-charged residues to alanines would result in decreased Cas9 identification. Our results confirmed this hypothesis, as mutating the three positive-charged residues, histidine (H), lysine (K), and arginine (R), completely abrogated Cas9 identification by our classifier. Specifically, lysine and arginine together prove to be the most critical residues for Cas9 identity. Mutating all negative-charged residues or single amino acids to alanine had negligible effects (Figure 4-5).

Due to their activity and identity as Cas9 inhibitors, anti-CRISPRs rely heavily upon their ability to competitively inhibit PAM binding, DNA unwinding, and appropriate HNH conformational changes for cleavage, which are all dependent on residue-based interactions with Cas9 [44]. Thus, we hypothesized that negative-charged residues are critical for anti-CRISPR identity. Our results further confirmed this hypothesis, as mutating all negative-charged residues, aspartic acid (D) and glutamic acid (E), to alanine reduced the number of identified anti-CRISPRs by nearly 44% (Figure 4-5).

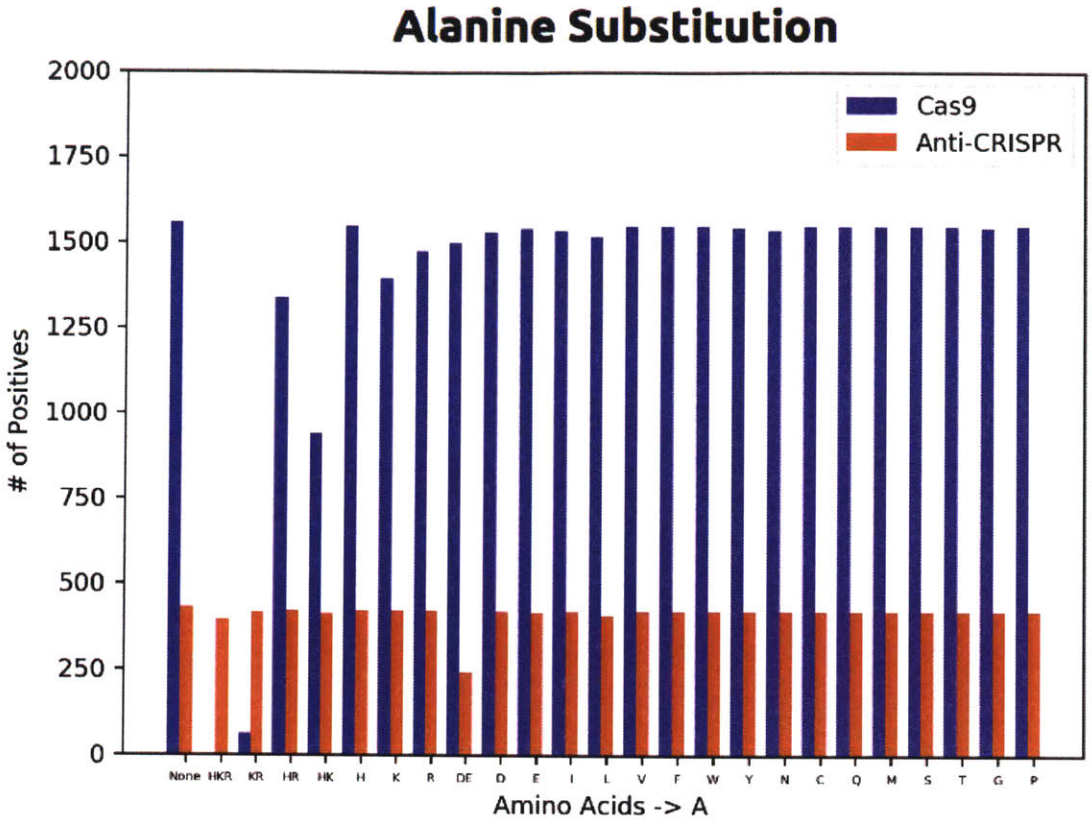

Figure 4-5: Alanine substitution of Cas9 and anti-CRISPRs. Indicated amino acids were mutated to alanine and number of positively-labeled sequences were counted after prediction with respective SVM models. For Cas9 sequences, either mutating all positively charged residues (H,K,R) or only two (K,R) was sufficient to abrogate Cas9 identification. Conversely, mutating all negatively charged residues (D,E) for anti-CRISPR proteins reduced the number of positive hits by 44%, with other changes having negligible effects.

Overall, of the known anti-CRISPRs, most are only able to inhibit the standard SpCas9 and reduce its off-target effects. Thus, there is a need to fill this gap by identifying or engineering anti-CRISPRs that can function on other Cas9 orthologs. Our developed tools will aid in this discovery process.

Appendix

5'-PAM-3'	VEGFA Target	Chr 6 Position (+)
GAGTGTGT	GTGTGTCTGTGTGGGTGAGT	43469703
GCGTGTGG	GTGTGGGTGAGTGAGTGTGT	43469711
GGGTTGAG	GTGAGTGAGTGTGTGCGTGTG	43769717
GTGTGCGT	GTCTGTGTGGGTGAGTGAGT	43469707
GTGAGTGA	GGACGTGTGTGTCTGTGTGG	43469697
GTGCGTGT	CTGTGTGGGTGAGTGAGTGT	43469691
GTGGGGTG	GCTCGGCCACCACAGGGAAG	43469709
GTGTGTGC	GTGTCTGTGTGGGTGAGTGA	43469705

Table A-1: VEGFA Target Sequences. sgRNAs were designed to target the indicated VEGFA loci, provided the indicated PAM sequences.

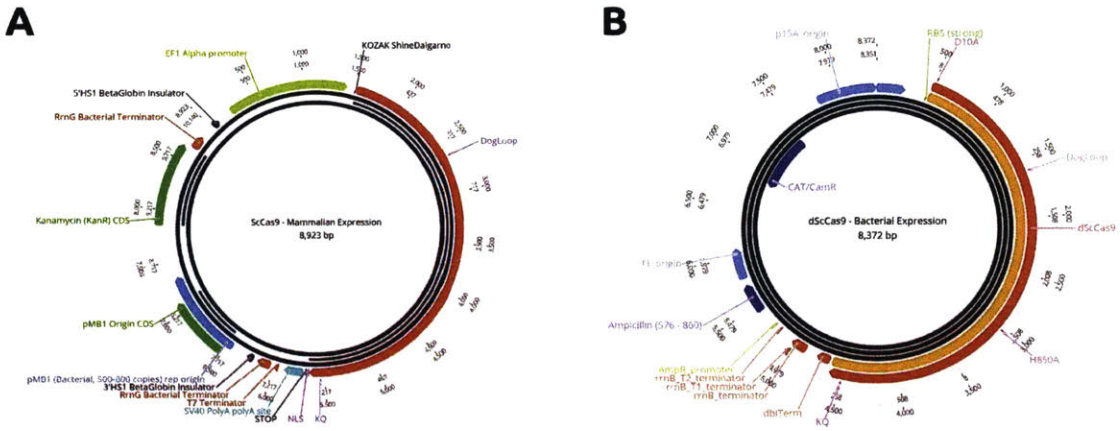

Figure A-1: ScCas9 Plasmid Maps. A) ScCas9 mammalian expression plasmid. ScCas9 is under the control of a constitutive EF1-α promoter. B) dScCas9 bacterial expression plasmid for PAM-SCANR assay. Nuclease-deficient ScCas9 (D10A and H841A) is under the control of a constitutive J23108 promoter within a pBAD33 vector.

Bibliography

[1] Doudna, J.A., et al. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* (2014).

[2] Sander, J., et al. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* (2014).

[3] Gaj, T., et al. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnology* (2014).

[4] Qi, L.S., et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* (2013).

[5] Zetsche, B., et al. Cpf1 is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* (2015).

[6] Shah, S.A., et al. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biology* (2013).

[7] Jinek, M., et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* (2012).

[8] Sternberg, S.H., et al. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* (2014).

[9] Mojica, F.J., et al. Short motif sequences determine the targets of the prokaryotic CRISPR defense system. *Microbiology* (2009).

[10] Richardson, C.D., et al. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature Biotechnology* (2016).

[11] Komor, A.C., et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* (2016).

[12] Gaudelli, N.M., et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* (2017).

[13] Biswas, A., et al. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biology* (2013).

[14] Ran, F.A., et al. *In vivo* genome editing using Staphylococcus aureus Cas9. *Nature* (2015).

[15] Kim, E., et al. *In vivo* genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nature Communications* (2017).

[16] Esvelt, K., et al. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods* (2013).

[17] Harrington, L. et al. A thermostable Cas9 with increased lifetime in human plasma. *bioRxiv* (2017).

[18] Kleinstiver, B.P., et al. Engineered CRISPR-Cas9 nucleases with altered specificities. *Nature* (2015).

[19] Hu, J.H., et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* (2018).

[20] Gao, L., et al. Engineered Cpf1 variants with altered specificities. *Nature Biotechnology* (2017).

[21] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* (2017).

[22] Henikoff, S., et al. Amino acid substitution matrices from protein blocks. *PNAS* (1992).

[23] Nishimasu, H., et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* (2014).

[24] Anders, C., et al. Structural plasticity of PAM recognition by engineered variants of the RNA-guided endonuclease Cas9. *Molecular Cell* (2016).

[25] Lefébure, T., et al. Gene Repertoire Evolution of *Streptococcus pyogenes* Inferred from Phylogenomic Analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. *PLOS ONE* (2012).

[26] Altschul, S.F, et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology* (1990).

[27] Crooks, G.E., et al. WebLogo: A Sequence Logo Generator. *Genome Research* (2004).

[28] Leenay, R., et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell* (2016).

[29] Hsu, P., et al. DNA targeting specificities of RNA-guided Cas9 nucleases. *Nature Biotechnology* (2013).

[30] Schindelin, J., et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods* (2012).

[31] Certo, M.T., et al. Tracking genome engineering outcome at individual DNA breakpoints. *Nature Methods* (2011).

[32] Shmakov, S., et al. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* (2017).

[33] Müller, M., et al. *Streptococcus thermophilus* CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome. *Molecular Therapy* (2016).

[34] Fonfara, I., et al. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Research* (2014).

[35] Grissa, I., et al. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* (2007).

[36] Langmead, B., et al. Fast gapped-read alignment with Bowtie 2. *Nature Methods* (2012).

[37] Langmead, B., et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* (2009).

[38] Krizhevsky, A., et al. *Advances in Neural Information Processing Systems*, 2012.

[39] Chari, R., et al. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods*, 2015.

[40] Rauscher, B., et al. GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. *Nucleic Acids Research*, 2016.

[41] Chollet, F. Keras: Deep Learning library for Theano and TensorFlow. GitHub; 2015. Available from: https://github.com/fchollet/keras.

[42] Pumperla, M. Hyperas. Available from: https://github.com/maxpumperla/hyperas.

[43] Wong, N., et al. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biology*, 2014.

[44] Pawluk, A., et al. Anti-CRISPR: discovery, mechanism and function. *Nature Reviews Microbiology* (2017).

[45] Shin, J., et al. Disabling Cas9 by an anti-CRISPR DNA mimic. *Science Advances* (2017).

[46] Dong, C., et al. Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Research* (2017).