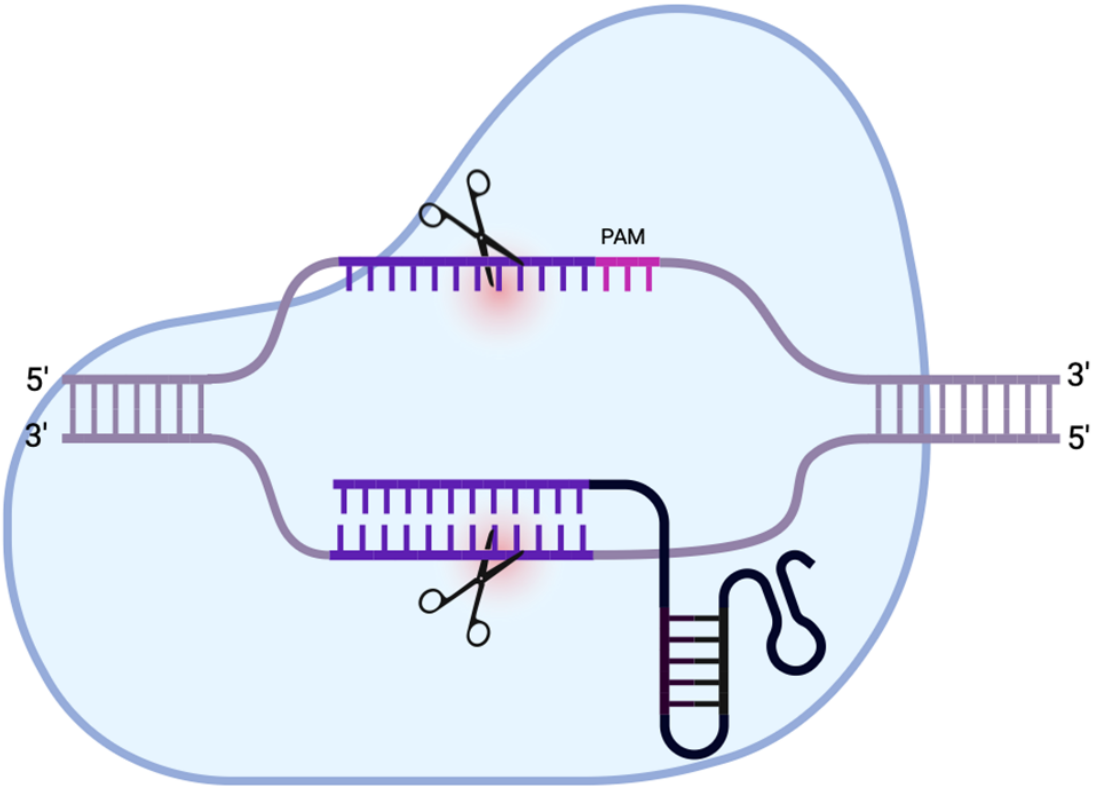
CHALMERS
UNIVERSITY OF TECHNOLOGY



Expanding the Target Range of a Type II-B Cas9 Enzyme Through Evolution-Guided Engineering

Master's thesis in Biotechnology

CAESAR NELSON

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Expanding the Target Range of a Type II-B Cas9 Enzyme Through Evolution-Guided Engineering

CAESAR NELSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Chalmers University of Technology
Gothenburg, Sweden 2026
Telefone: +46 (0)31-772 1000

Sponsored and performed in collaboration with
the Genome Editing Technology Team
Genome Engineering Department
ASTRAZENECA AB R&D
Mölndal, Sweden 2026
Telephone: +46 (0)31-776 1000

Expanding the Target Range of a Type II-B Cas9 Enzyme Through Evolution-Guided Engineering
CAESAR NELSON

Supervisor: Oi Kuan Choong, AstraZeneca AB
Examiner: Verena Siewers, Life Sciences

Master's Thesis 2026
Department of Life Sciences
Division of Systems Biology
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Figure depicts a Cas9-mediated cleavage reaction. Created with Biorender.com.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Expanding the Target Range of a Type II-B Cas9 Enzyme Through Evolution-Guided Engineering
CAESAR NELSON
Department of Life Sciences
Chalmers University of Technology

Abstract

The discovery and development of CRISPR-Cas9 has revolutionized genome editing and enabled the development of novel therapeutic approaches. However, many disease-causing mutations are still inaccessible due to the absence of relevant PAM sequences within the therapeutic editing window. This master's thesis employed directed evolution to generate ePsCas9 variants with expanded PAM recognition capabilities, enabling the targeting of disease-causing mutations previously inaccessible. Different library generation approaches and selection strategies were systematically explored. Selected variants obtained from the directed evolution were screened using a luciferase reporter system and four lead candidates (pCN1, pCN12, pCN15, and pCN27) were identified. A novel PAM-determination assay was established to profile motif recognition, revealing that pCN1, pCN12, and pCN15 exhibit relaxed recognition of NRG or NDG motifs. This PAM assay provides a valuable tool for characterizing Cas9 variants and was successfully applied to determine the NNNGAA recognition motif of another novel proprietary AstraZeneca Cas9 enzyme. Genome editing analysis demonstrated that pCN12 and pCN15 achieved equal or greater activity than the established SpRY variant at two of three NAG-PAM target sequences, while pCN27 showed enhanced NGC-PAM activity compared to wild-type ePsCas9.

Keywords: CRISPR-Cas9, directed evolution, ePsCas9, library generation, selection system.

Acknowledgements

I am deeply grateful to my supervisor, Oi Kuan Choong, for her invaluable guidance throughout this project. I also wish to extend my heartfelt thanks to Jakub Wiktor, who served as co-supervisor for all the bioinformatics work and provided steady support and insight. I am sincerely thankful to my line manager, Saša Sviković, for his enthusiasm for this project and for generously sharing his valuable knowledge.

I would like to acknowledge and appreciate the invaluable advice and encouragement from the members of the Genome Engineering Department led by Marcello Maresca. I am especially grateful to the members of the Genome Editing Technologies Team for their thoughtful feedback and continuous support.

I extend my sincere thanks to my examiner, Verena Siewers at Chalmers University of Technology, for her excellent support and mentorship in both the academic and administrative aspects of this project.

I thank AstraZeneca for providing students with the opportunity to contribute, learn, and gain experience within the pharmaceutical industry.

Finally, I want to thank my family and friends for their steadfast support and encouragement throughout this project.

Caesar Nelson, Gothenburg, April 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

Cas	CRISPR-associated protein
CDS	Coding sequence
crRNA	CRISPR RNA
ePsCas9	Engineered *Parasutella secunda* Cas9
HDR	Homology-directed Repair
MMEJ	Microhomology-mediated End Joining
NHEJ	Non-homologous End Joining
PCR	Polymerase Chain Reaction
PAM	Protospacer adjacent Motif
sgRNA	Single-guide RNA
SpCas9	*Streptococcus pyogenes* Cas9
tracerRNA	trans-activating RNA

Contents

List of Figures

List of Tables

1
Introduction

1.1 Theoretical Background

The concept of editing the human genome using specific endonucleases has evolved from a theoretical possibility to approved treatments within three decades [1]. Although early genome editing tools, such as zinc finger nucleases (ZFNs) and transcriptor activator-like effector nucleases (TALENs), provided a platform for introducing targeted genome modifications, they require great engineering effort to target specific sequences [2, 3]. The discovery of CRISPR-Cas9 and its subsequent development into a programmable genome editing tool fundamentally transformed the efficacy of producing treatments for genetic diseases [4, 5, 6]. In contrast, to ZFNs and TALENs, CRISPR-Cas9 consists of a ribonucleoprotein (RNP), where a programmable single-guide RNA (sgRNA) directs the Cas9 protein to induce precise double-stranded DNA breaks (DSBs) [5].

The clinical impact of CRISPR-Cas9 has been demonstrated through several approved gene therapies, such as the first *ex vivo* treatment for sickle cell disease that received regulatory approval in 2023 [7]. Beyond inducing DSBs breaks for gene disruption, CRISPR-Cas9 has been further developed into base editors that enable conversions of specific nucleotides without creating DSBs, and prime editors that can insert, delete, or replace DNA sequences with single nucleotide precision through the use of reverse transcription [8, 9]. Despite these remarkable achievements, CRISPR-Cas9 technologies are constrained by the requirement for a protospacer adjacent motif (PAM), a short DNA sequence located adjacent to the target sequence [4]. This PAM dependency significantly restricts targeting flexibility, as mutations in PAM-deficient genomic regions can become inaccessible. Engineering Cas9 with altered or relaxed PAM specificity therefore represents a viable strategy for expanding the targeting range and therapeutical applicability of CRISPR-Cas9-based genome editing.

1.1.1 CRISPR-Cas9

In bacteria and archaea, clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated proteins (Cas) was first discovered as adaptive defense systems that integrates short stretches of foreign DNA sequences (spacers) into the host's genome between repetitive CRISPR sequences [10, 11]. These arrays of short foreign and CRISPR sequenes are called CRISPR arrays and act as memory of past infections. Elements of the CRISPR array, together with Cas enable the

host to recognize and destroy DNA or RNA of returning genetic invaders, such as viruses [11]. CRISPR-Cas systems are divided into two main classes of CRISPR-Cas systems, class 1 and class 2, containing a total of six types of different CRISPR-Cas systems, type I-VI, and 33 subtypes [12]. CRISPR-Cas class 1 systems (type I, III, and IV), consists of several Cas proteins forming a multi-Cas protein complex, while CRISPR-Cas class 2 (type II, V, and VI) consists of a single Cas protein. Only CRISPR-Cas type II and V target DNA, while type VI targets RNA [12].

Type II CRISPR-Cas systems have Cas proteins called Cas9, and follow a distinct maturation pathway. Here, the long RNA sequences containing the spacer-repeat sequences (CRISPR arrays) are transcribed into pre-crRNA, which is then proccessed into crRNA in a pathway involving trans-activating RNA (tracrRNA), Cas9, and RNAse III [13]. The tracrRNA contains a 25-nt sequence complementary to the repeat in the pre-crRNA and annealing of these complementary sequences recruits the Cas9 enzyme to form a Cas9-pre-crRNA-tracrRNA complex. The formation of this ribonucleoprotein-complex recruits RNAse III that cleaves the tracrRNA-pre-crRNA at two distinct positions, leaving a mature crRNA still annealed to the tracrRNA and the Cas9 protein. The mature crRNA consists of 20 nt derived from the spacer sequence, and 22 nt derived from the repeat sequence that hybridizes with the tracrRNA forming a hairpin structure that interacts and binds with the Cas9 enzyme. Once assembled, the Cas9-ribonucleoprotein (Cas9-RNP) initiates target search by randomly colliding with the DNA until it encounters a PAM sequence, and if the spacer sequence matches the target sequence (protospacer), induces a DSB [14].

However, more than a decade ago, Jinek et al. (2012), found that a truncated tracrRNA retaining the endogenous nucleotides 23 to 48, compared to the native 72-nt tracrRNA, was capable of inducing DSBs at equal efficiency [4]. This discovery led the same researchers to design a chimeric RNA, a single-guide RNA or a sgRNA, by fusing the tracrRNA and the mature crRNA [4]. The development of an engineered sgRNA, awarded the noble prize in chemistry in 2020, eliminated the need for both the tracrRNA processing and crRNA maturation pathway and enabled target reprogramming by modifying the spacer sequence [15].

1.1.2 Cas9 Protein Structure and PAM Recognition

The structure of Cas9 type II subtypes (IIA-C) proteins are similar, with type IIA *Streptococcus pyogenes* Cas9 (SpCas9) being the most well understood among them. In general, the Cas9 in its apo state, meaning that it's not bound to a sgRNA or a tracrRNA-crRNA duplex, consists of two different lobes, the alpha-helical recognition (REC) lobe and the nuclease (NUC) lobe containing the divided RuvC domains, the HNH domain, and the C-terminal domain (CTD).

Two linking segments, one forming a bridge helix and the other a disordered linker segment, connect the two lobes. The REC lobe is organized into three alpha-helical domains (Hel-I, Hel-II, Hel-III). The CTD domain, which can be further divided into a WED and PI domain, is composed of PAM-interacting sites responsible for

PAM-recognition. In the Cas9 enzyme's apo state, the PAM-interacting regions are mostly disordered, which keeps the enzyme in an inactive conformation until the RNA binding occurs. The HNH domain, referred to as the His-Me finger domain, is a conserved catalytic region characteristic of enzymes that catalyze phosphodiester bonds in nucleic acids [16].

The RuvC is split into three smaller subdomains (RuvC I-III) and uses a two-metal mechanism and requires Mg^{2+} to catalyze the cleavage reaction of the phosphodiester bond [17]. During the assembly of Cas9 and sgRNA (or tracrRNA-crRNA duplex) the Cas9 undergoes a structural rearrangement into its active form and starts searching for its target sequence [18]. The target recognition for CRISPR-Cas9 is a two-step process requiring both the sgRNA spacer to bind to the protospacer and the recognition of an adjacent PAM sequence, located on the opposite strand, to the protospacer sequence.

Once the activated Cas9 enzyme encounters a PAM sequence, another conformational change initiates the seed nucleation. During the seed nucleation 10-12 nucleotides at the 3' end of the spacer sequence are probed against the protospacer sequence. If the seed sequence matches, the remaining sgRNA spacer sequence will bind, and if instead mismatches are identified, the Cas9 dissociates from the DNA [19]. A fully bound DNA-sgRNA forms an R-loop state that repositions the HNH and RuvC domains, which cleave both the target-strand and the non-target strand. For SpCas9, the PAM sequence recognized is NGG and the induced DSB occurs approximately 3 nt upstream of the PAM sequence on both strands, causing a blunt ended cut. In Figure 1.1, the Cas9 ribonucleoprotein formation, PAM recognition, seed nucleation, and DNA cleavage is depicted.

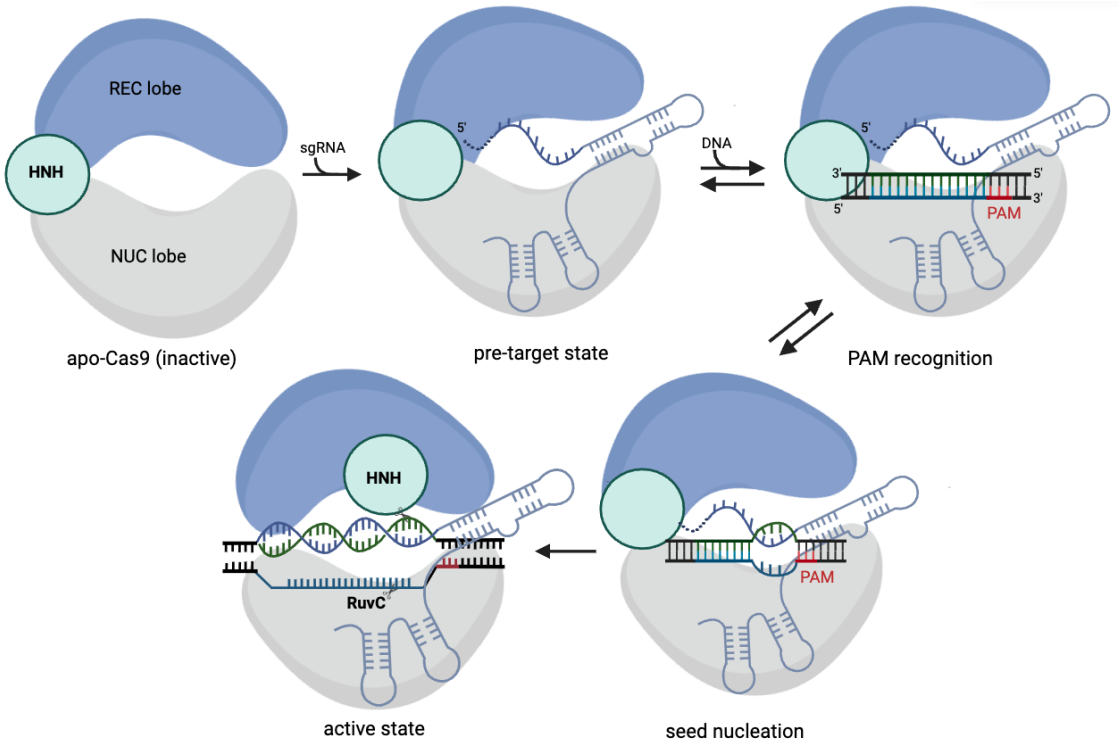


Figure 1.1: Cas9-mediated DNA targeting and cleavage mechanisms. The apo-Cas9 binds to the sgRNA and forms a Cas9 ribonucleoprotein (pre-target state). Once DNA containing a PAM sequence is encountered, a short segment of the sgRNA spacer is seeded (seed nucleation) and if the seed sequence is a match, then the remaining spacer is bound. When the spacer is fully bound to the DNA, two Cas9 protein domains are repositioned and the dsDNA is cleaved (active state). Created with biorender.com.

1.1.3 DNA Repair Mechanisms

When a double-strand break occurs in DNA, such as those induced by Cas9 enzymes, cells repair the DNA damage using two primary repair mechanisms found in both prokaryotes and eukaryotes: non-homologous end joining (NHEJ) and homologous recombination (HR). NHEJ works by directly joining the broken DNA ends together without the need for a homologous template, making it a relatively fast but error-prone process that can introduce small insertions or deletions [20]. In contrast, HR repairs the break using a homologous DNA sequence (typically the sister chromatid) as a template to accurately restore the original DNA sequence [21], making it a more precise but also more complex compared to NHEJ. In certain cases a third repair mechanism can occur: microhomology-mediated end-joining (MMEJ) [22]. MMEJ is an alternative repair pathway to NHEJ, and repairs a double-strand break using microhomologies of 1-16 nucleotides flanking the cut site, resulting in deletions.

1.1.4 ePsCas9

Since the discovery and development of SpCas9 into a genome editing tool, finding and developing Cas9 enzymes with higher sequence specificity and lower off-target effects has been a priority. For this reason many Cas9 orthologs have been found and characterized, among them *Parasutterella secunda* Cas9 (PsCas9) [23]. PsCas9 recognizes an NGG PAM sequence, has an optimal sgRNA spacer sequence length of 22-nt and is more sensitive to mismatches between the sgRNA spacer and the protospacer sequence, making it a safer therapeutic tool [23]. A similar increased specificity has been reported for *Francisella novicida* Cas9 (FnCas9), which belongs to the same type II-B subcategory as PsCas9 [24]. However, it was found that the editing efficiency of PsCas9 was highly dependent on the *in vivo* concentration of Cas9 RNP, likely due to its low affinity for DNA in mammalian cells. Therefore an engineered variant of PsCas9, ePsCas9, was developed, achieving similarly high sequence specificity but with much higher editing activity [25].

1.1.5 Directed Evolution

The engineering effort of Degtev et al. (2024) to increase the editing efficiency of PsCas9 was carried out using a protein engineering strategy called rational design, in which strategic mutations were introduced to increase DNA binding affinity of the Cas9 enzyme [25]. Rational design has also been applied to obtain Cas9 PAM-engineered variants, such as enhanced FnCas9 (enFnCas9) [26]. However, rational design is limited by the availability of structural and mechanistic insight. As an alternative, directed evolution has been employed to generate Cas9 PAM-engineered variants, such as xCas9 [27]. Directed evolution consists of iterative cycles of library generation and selection for determined phenotypical traits [28, 29, 30, 31, 32]. The purpose of the library generation is to create a diverse set containing the gene of interest (GOI) or selected parts of the GOI with randomized changes, such as mutations, deletions, or insertions [28]. Numerous methods and strategies for both *in vitro* and *in vivo* mutagenesis have been developed [33, 34, 35, 36, 37, 38, 39]. Examples of mutagenesis methods that have previously been used to evolve

Cas9 variants include *in vitro* methods such as error-prone PCR (epPCR), which uses a low-fidelity polymerase that misincorporates nucleotides during PCR amplification, and *in vivo* methods such as phage-assisted continuous evolution (PACE), in which genetically modified *E. coli* with impaired replication mechanistic introduce mutations into a mutagenesis plasmid [40, 41]. The purpose of the selection is to identify variants of the protein that exhibit traits of interest, such as improved substrate binding, solubility, thermostability, PAM-recognition, and more. There exist a wide range of methods for selection, but it is common that selection is coupled to increased growth, or survival, of cells that harbor variants of interest [40, 41].

1.2 Aim

The aim of this master's thesis is to evolve an AstraZeneca proprietary type II-B Cas9 enzyme, ePsCas9, Degtev et al. (2024), to obtain efficient ePsCas9 variants capable of recognizing alternative PAM sequences and to characterize the PAM-recognition of these variants. In the future, this will allow for therapeutic targeting of currently inaccessible loci.

1.3 Objectives

To achieve the stated aim, four objectives were established:

- The first objective is to evolve novel ePsCas9 PAM-relaxed variants through directed evolution and explore different strategies to improve the evolution process.
- The second objective is to evaluate the efficiency of ePsCas9 variants in recognizing alternative PAM sequences. An in-house established luciferase reporter system integrated at the *HBEGF* locus in HEK293T cells serves as an excellent reporter assay to evaluate the editing efficiency of new variants in a mammalian cell context.
- The third objective is to establish a Cas9 PAM determination assay, and profile the PAM sequences recognized by novel ePsCas9 PAM-relaxed variants.
- The fourth objective is to compare the editing efficiency of the novel ePsCas9 variants with established Cas9 PAM-engineered variants at various loci with different PAM sequences.

2
Methods

In this section, the methodology will be described. Primer and oligonucleotide sequences can be found in Appendix B. Note that due to confidentiality primer sequences binding to the coding sequence of ePsCas9 will not be disclosed, instead explanations of the primers will be provided.

2.1 Directed Evolution

2.1.1 ePsCas9 Plasmid Library Generation

The WED and PI domian, approximately 1.3 kb in length, for ePsCas9 was targeted for mutagenesis using GeneMorph II Random Mutagenesis Kit [42]. During mutagenesis via error-prone PCR (epPCR) there are two main ways of controlling the error rate: varying the number of PCR cycles (denaturation, annealing, elongation) or varying the amount of input DNA in the PCR reaction. In this project, the mutation rate was controlled by adjusting the amount of linear template in the epPCR reaction. In Figure 2.1, the Cas9 plasmid used in library generation and the CDS for ePsCas9 are shown.

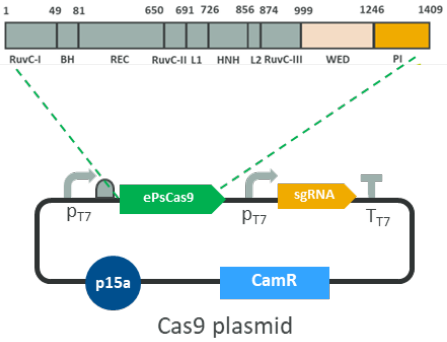


Figure 2.1: The bacterial Cas9 plasmid used in the library generation. Selected regions of the ePsCas9 CDS was mutagenised using epPCR and assembled with an unmodified backbone vector into Cas9 plasmid libraries.

The unmodified WED and PI domain of ePsCas9 (Fig. 2.1) was amplified from the Cas9 plasmid using 2X Phusion Flash PCR Master Mix (Thermo Fisher Scientific, cat. #F548S) and primers CN9 and CN10 (Appendix B.1) according to the manufacturer's protocol. Following amplification, reaction mixures were treated with DpnI (1 µL, New England Biolabs, cat. #R0176S) at 37^{o}C for 15 min, then

80^{o}C for 20 min to digest template plasmids, then purified by gel extraction using QIAquick Gel Extraction Kit (Qiagen, cat. #28704).

The unmodified linear WED and PI domain was subsequently used as input template in an error-prone PCR reaction using the Mutazyme II DNA Polymerase Kit (Agilent Technologies, cat. #200550) and primers CN9G and CN10G (Appendix B.1) according to the manufacturer's protocol. After amplification, reaction mixtures were purified by gel extraction using QIAquick Gel Extraction Kit.

The mutation frequency was empirically determined to be 4–9 mutations per 1.3 kb of linear template, consistent with the GeneMorph II Random Mutagenesis Kit specifications [42]. However, to introduce some variation in mutation rate the input template was slightly varied between library generations. For L1, L2, L3, and L4 the input template in the PCR reaction was 300 ng, 200 ng, 250 ng, and 250 ng, respectively. Note that the primers (CN9G and CN10G) used in the epPCR reactions contain the same binding sequence as the primers (CN9 and CN10) in the linear template amplification but have additional overhang sequences harboring BsaI recognition sites, allowing for downstream Golden Gate assembly.

A linear backbone vector containing the remaining ePsCas9 CDS was amplified from the Cas9 plasmid (Fig. 2.1) using High-fidelity 2X Phusion Flash PCR Master Mix and primers (CN11 and CN13; Appendix B.1) that harbor BsaI recognition sites and complementary overhangs to the epPCR-products. Following amplification, the reaction mix was DpnI treated and purified by gel extraction.

The Golden Gate plasmid assembly was performed according to a standardized assembly protocol (see Tab. 2.1 for thermocycler protocol) proposed by Daffern et al., 2023 [43]. For each generated library, a 200 µL reaction containing 200 fmol of linear vector, 200 fmol error-prone PCR product, 20 µL 10x T4 ligase buffer (New England Biolabs, cat. #B0202S), 5 µL BsaI-HF-V2 (New England Biolabs, cat. #R3733S), 5 µL T4 DNA ligase (New England Biolabs, cat. #M0202S), and nuclease-free H_2O up to 200 µL.

Table 2.1: Thermocycler conditions for the standardised Golden Gate assembly protocol.

Temperature	Time (s)	Cycles
37^{o}C	1 min	60
16^{o}C	1 min	
37^{o}C	5 min	1
**80^{o}C	20 min	1
4^{o}C	∞	

The heat inactivation of enzymes step, marked as "**", has been altered to the recommended temperature for BsaI-HF-V2 from NEB.

All plasmid assembly reactions were purified and eluted in 6 µL nuclease-free H_2O using Monarch DNA cleanup kit (New England Biolabs, cat. #T1030). The purified assembly reactions were measured using NanoDrop (Thermo Fisher Scientific, cat. #ND-ONE-W) to ensure sufficient purity, after which the remaining plasmid was transformed into high-competency electrocompetent DH10B (New England Biolabds, cat. #C3020K) according to the manufacturers recommendations. For the first generation libraries, 100 mL TB medium with 25 µg/mL chloramphenicol was inoculated with the transformation mix, grown overnight at 37^{o}C with shaking, and the plasmid libraries were extracted using the QIAGEN Plasmid Plus Maxi Kit (Qiagen, cat. #12963) according to the manufacturer's protocol. Transformation efficiency was determined by plating serial dilutions (50 µL aliqouts diluted 10^{-3} and 10^{-4}) of representative transformations.

For second generation libraries, each transformation mix was plated on a 500 cm^2 LB agar (Thermo Fisher Scientific, cat. #22700025) plate with 25 µg/mL chloramphenicol, incubated at 37^{o}C overnight, collected in 25 mL TB, inoculated for 3 h, and the plasmid library were extracted using the QIAGEN Plasmid Plus Maxi Kit. Agar-based selection was used to minimize potential selection bias, as slower antibiotic diffusion in solid medium may better preserve library diversity compared to direct liquid culture inoculation.

The transformation efficiency, defined as colony forming units (cfu), is calculated as follows:

$$\text{TE (CFU}/\mu\text{g)} = \frac{\text{colonies}}{\text{pDNA } (\mu\text{g})} \times \text{dilution factor} \tag{2.1}$$

The number of colonies on an agar plate is divided by the amount of plasmid DNA (μg) and the dilution.

2.1.1.1 Second-Generation Library Construction

The second-generation Cas9 plasmid libraries were constructed according to four approaches (A-D). Second-generation Cas9 libraries using approach A-C were generated using error-prone PCR and Golden Gate assembly as described in section 2.1.1, with different ePsCas9 variants as input templates:

Approach A: template consisted of a single Cas9 variant selected on TTG PAM containing one missense mutation.

Approach B: template consisted of two Cas9 variants selected on TGC PAM (carrying 2 or 3 missense mutations in the WED and PI domain, respectively) mixed in equal proportions.

Approach C: template consisted of pooled plasmids from approximately 600 colonies across all PAM-selections from the first selection of the first-generation libraries.

Approach D employed a hybrid mutagenesis strategy (Fig. 2.2). Eight ePsCas9 variants (pCN1-pCN8) selected on TTG PAM were pooled equally. The WED and

PI domain, and plasmid backbone were amplified from this pool using high-fidelity polymerase, while the upstream 3/4 of the ePsCas9 CDS (Fig. 2.1) was amplified from wild-type ePsCas9 plasmid and mutagenized by error-prone PCR using the GeneMorph II Random Mutagenesis Kit [42]. Both fragments were assembled into Cas9 plasmid libraries using the Golden Gate protocol described previously.

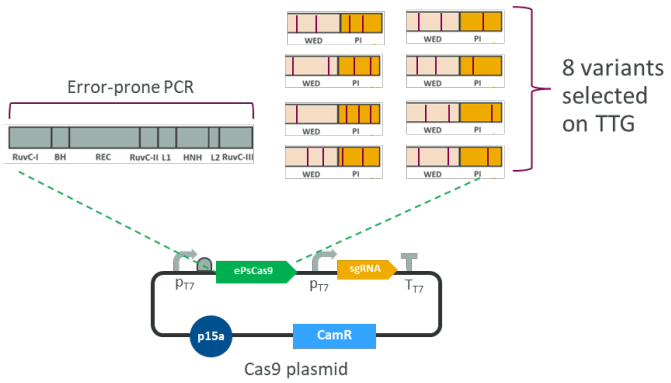


Figure 2.2: Principle of library generation using approach D. The non-PAM interacting upstream CDS was mutagenised and assembled with the WED and PI domain of pCN1-pCN8 into a Cas9 plasmid library.

The unmodified coding sequence upstream of the WED and PI domain of ePsCas9 (Fig. 2.1) was amplified from the Cas9 plasmid using 2X Phusion Flash PCR Master Mix and primers CN38 and CN39. Following amplification, reaction mixtures were treated with DpnI and purified by gel extraction QIAquick Gel Extraction Kit.

The unmodified linear coding sequence was used as the input template (300 ng) in the following epPCR reaction with using the Mutazyme II DNA Polymerase Kit and CN12G and CN11G. Following amplification, reaction mixtures were purified by gel extraction using the QIAquick Gel Extraction Kit. In addition, the linear backbone vector was amplified from the 8-TTG-Cas9 plasmid pool, that contained the WED and PI domain from the 8 selected variants, using 2X Phusion Flash PCR Master Mix and primers CN9G and CN14G (Appendix B.1). Following amplification the linear backbone vector was DpnI treated and purified by gel extraction.

2.1.2 Selection System and Retransformation Procedure

2.1.2.1 Overview of the Survival Selection System

A survival selection system was established in-house according to B. Kleinstiver, et al. (2015) , with modifications [40]. It consists of two plasmids, the selection plasmid and a Cas9 library plasmid. The first plasmid, the selection plasmid, encodes a DNA gyrase inhibitor called CcdB, and the *CcdB* gene expression is driven by an arabinose-inducible (p_{BAD}) promoter. The selection plasmid contains a defined 22-nt protospacer sequence extracted from the *EMX1* gene (EMX1b) followed by a distinct PAM sequence consisting of three nucleotides. Additionally, the selection

plasmid contains a ColE1 origin of replication, which ensures a high copy number, and an ampicillin resistance selection marker.

The second plasmid belongs to the Cas9 plasmid library, which contains a mutagenized ePsCas9 gene driven by a T7 promoter, followed by a gene encoding an sgRNA targeting the EMX1b protospacer sequence on the selection plasmid, which is flanked by its own T7 promoter and T7 terminator (see Fig. 2.3). The T7 promoter allows for strong and orthogonal expression of both Cas9 enzymes and sgRNA, but requires expression of T7 polymerase in the host cell as it is not natively expressed in *E. coli*. Furthermore, the plasmid harbors a p15a origin of replication as well as a chloramphenicol selection marker. p15a leads to a low to medium copy number and, more importantly, is compatible with ColE1 origin of replication, allowing both plasmids to replicate independently in the same cell.

Six selection plasmids, each carrying a distinct PAM sequence (TGG, TTT, TCG, TTG, TGT, or TGC), were individually transformed into *E. coli* BW25141 (DE3) to generate six in-house electrocompetent cell stocks, hereafter, referred to as PAM (sequence)-cell. The transformation efficiency of each stock was verified prior to selection.

During selection, six different electrocompetent cell stocks, each harboring a different selection plasmids, were transformed with Cas9 plasmid libraries and the transformation mixes were streaked onto agar plates containing arabinose and chloramphenicol. The presence of chloramphenicol selects for cells carrying Cas9 plasmid variants and the presence of arabinose induces expression of CcdB, killing the cells carrying the selection plasmid. Consequently, only the cells carrying Cas9 variants able to recognize the PAM sequence on the selection plasmid survive (Fig. 2.3).

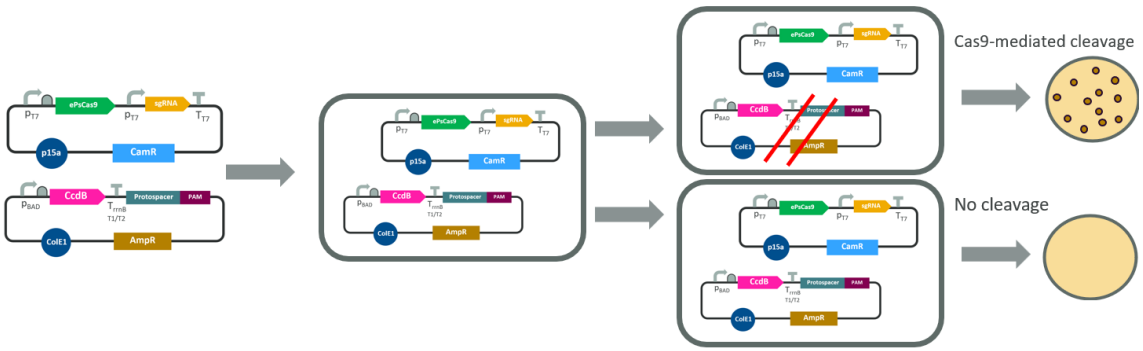


Figure 2.3: Principle of the bacterial selection system. Cas9 plasmid libraries were transformed into competent cell stocks harboring a selection plasmid and streaked onto agar plates containing chloramphenicol and L-arabinose. If the PAM sequence on the selection plasmid was recognized then toxic gene was not expressed and the variants survived.

2.1.2.2 Selection and Retransformation

The electrocompetent cell stocks harboring different selection plasmids were transformed with first-generation Cas9 plasmid libraries, according to the protocol de-

scribed in Appendix C.1. Surviving colonies from the same selection cells were pooled in groups of 6, inoculated in 5 mL TB containing 25 µg/mL chloramphenicol, and plasmids were extracted using the QIAprep Spin Miniprep Kit (QIAGEN, cat. #27104). Each of the pooled plasmid preparation was then retransformed (Appendix C.1) into its original selection cells. From these retransformations, 5 individual colonies were picked, inoculated in 5 mL TB containing 25 µg/mL chloramphenicol, and plasmids were extracted using the QIAprep Spin Miniprep Kit.

The second-generation Cas9 plasmid libraries were selected and retransformed using the same strategy, but varied by which selection cells they were transformed into. Libraries generated using approach A and D were transformed into TTG-selection cells, approach B were transformed into TGC-selection cells, and libraries generated using approach C were transformed with all six selection cells. Each second-generation Cas9 plasmid library was transformed according to the protocol found in Appendix C.1. Surviving colonies from the same plasmid library were pooled together, inoculated in 25 mL containing 25 µg/mL chloramphenicol, and plasmids were extracted using the QIAGEN Plasmid Plus Maxi Kit. Each of the Cas9 plasmid libraries were subjected to a total of three iterative rounds of transformation, pooling surviving colonies, inoculation, and plasmid extraction. The plasmid libraries were transformed a fourth time and 10 colonies from each plasmid library were individually picked, inoculated in 5 mL TB containing 25 µg/mL chloramphenicol, and extracted using the QIAprep Spin Miniprep Kit.

Protocols for preparing TB medium and agar plates containing chloramphenicol and L-arabinose can be found in Appendix C.3-C.4.

2.1.3 Variant Characterization

Each variant was sequenced using Sanger sequencing, which often is referred to as the gold standard of sequencing due to its low error rate of 0.001 % [44]. Variants in the WED and PI domains were bidirectionally sequenced using primers CN9 and CN10 (Tab. B.1) to ensure adequate coverage of the targeted regions. Selected variants from the second-generation Cas9 library (approach D) were sequenced after screening in the luciferase reporter assay (see Results and Discussion 3.3). Mutations were identified by mapping Sanger sequencing files to the reference Cas9 plasmid (Fig. 2.1) in Geneious 2024.0.7 (https://www.geneious.com).

2.2 Molecular Cloning

2.2.1 Cloning of Mammalian Vectors

The coding sequence in the bacterial Cas9 plasmid (Fig. 2.1) is codon optimized for mammalian cells allowing for direct cloning into a mammalian Cas9 vector (Fig. 2.4) with an identical coding sequence. The ePsCas9 variant coding sequence and the mammalian backbone were amplified using 2X PlatinumSuperFi Master Mix (Thermo Fisher , cat. #12358010). Primers CN30 and CN31 were used for amplifying the ePsCas9 coding sequence, while primers CN40 and CN33 were used for

amplifying the mammalian backbone vector (Tab. B.1). Each reaction mix was DpnI-treated, purified by gel extraction, and the PCR products were assembled using InFusion (TAKARA, cat. #638943). Primers used during amplification were designed so that each linear fragment would contain a 15 bp overlap on each side, allowing for assembly using InFusion.

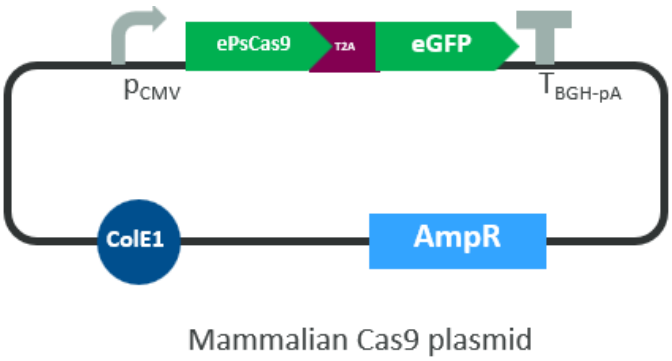


Figure 2.4: Mammalian Cas9 plasmid.

2.2.2 Site-directed Mutagenesis

Prior to using the Cas9 plasmid (Fig. 2.1) as template for downstream PCR, it was screened for common type IIS endonuclease recognition sites and BsaI was found to have the fewest recognition sites (one recognition site). At the BsaI-recognition site in the Cas9 plasmid (Fig. 2.1) forward and reverse primers were designed with an overlap of 15-nt, allowing for assembly using InFusion. The overlap additionally contained one mutated nucleotide, which removed the BsaI-recogntion site. 2X Phusion Flash PCR Master Mix and primers CN35 and CN36 were used during amplification. The subsequent reaction mix was DpnI-treated and purified by gel extraction.

Site-directed mutagenesis was also used to introduce the S1223A mutation or combined I1222A/S1223A mutations in the following variants: pCN1, pCN9, pCN12, pCN15, pCN24 and pCN27. Each of the variants resided in mammalian Cas9 plasmids (Fig. 2.4) and primer pairs were designed with overlapping sequences containing the desired mutations, as described earlier. The linear fragment with introduced mutations were amplified using 2X platinum SuperFi Master Mix, DpnI-treated and purified by gel extraction. Primers CN45 and CN47 were used to introduce the S1223A mutation, while primers CN46 and CN48 were used to introduce the I1222A and S1223A mutations. See table 3.2 for details on which variants contain the respective mutations.

2.3 Luciferase Reporter Assay

2.3.1 Overview of the Luciferase Reporter Assay

The Luciferase reporter system (Fig. 2.5) is a gene cassette integrated at the *HBEGF* locus in HEK293T cells [25, 45]. The gene cassette consists of *puroR*, *mScarlet*, and nanoluciferase (*Nluc*), each separated by a ribosomal skipping element (2A peptide) and coexpressed under a strong mammalian promoter (EF1a). Upstream of the CDS for *NLuc* there also exists a STOP codon followed by a specific PAM sequence. The STOP codon and the PAM sequence are furthermore flanked by a 9 nt MMEJ-repeat sequence. An sgRNA was designed to target a region spanning the upstream MMEJ repeat sequence and the STOP codon. Upon recognition of the adjacent PAM sequence, the Cas9 variant introduces a double-strand break positioned between the two MMEJ repeat sequences. Because the MMEJ-repeat sequences now flanks the dsDNA break, the MMEJ-repair mechanism is strongly favored as the main repair mechanism of action. The recruitment of the MMEJ-repair mechanism results in an excision of the STOP codon, which allows for the translation of *NLuc*.

The reporter cassette is expressed under a strong constitutive promoter, the EF1a promoter, and transcription is terminated under a dual terminator, including both the rabbit β-globulin (rbGlob) terminator and the *Bovine growth hormone polyadenylation* (BGH-pA) terminator. Additionally, the polyadenylation (pA) element serves to enhance the mRNA stability.

Once NLuc is translated, the enzyme is excreted by the cells into the cell medium from which a small sample can be taken and after adding a substrate containing furimazine a luminescence readout is measured using a fluorescent plate reader. The luminescence output is generated because the *NLuc* enzyme converts furimazine in the presence of oxygen into furimamide, which releases carbon dioxide and light. However, alternative repair pathways, such as NHEJ-repair, could result in repair patterns that do not remove the STOP codon. Selected samples were therefore submitted for amplicon sequencing for editing efficiency evaluation.

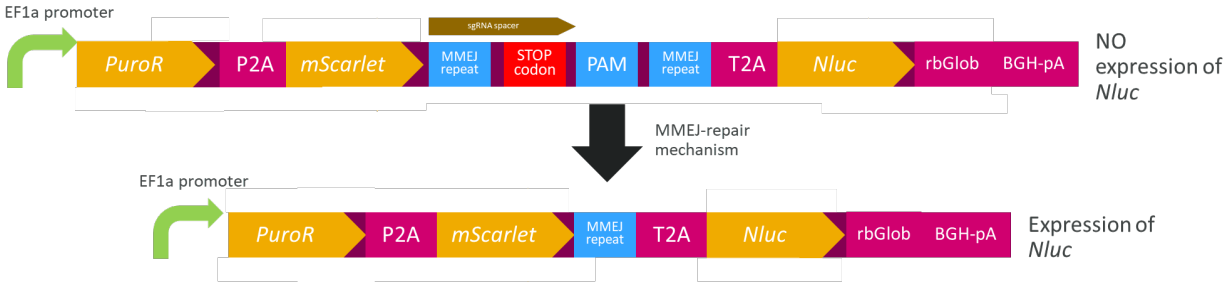


Figure 2.5: Schematic of the luciferase reporter system. The sgRNA targets a sequence in the reporter assay gene cassette, and if the PAM sequence is recognized by the Cas9 variant, a double-stranded break is induced between the MMEJ repeats. DNA repair via the MMEJ mechanism leads to the reactivation of NLuc, which is secreted into the cell medium.

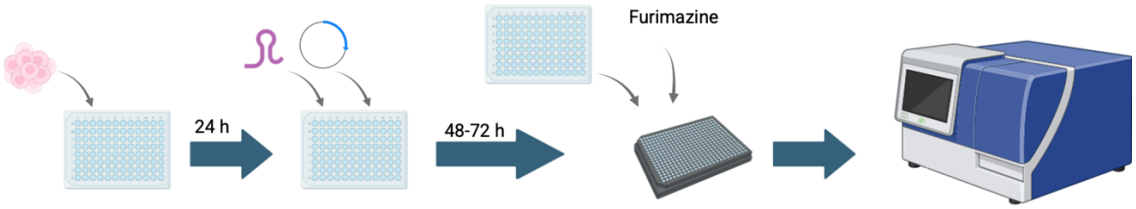


Figure 2.6: Illustration of the workflow for the luciferase reporter assay. The HEK293T reporter cell lines were seeded in a 96-well plate. After 24 h, the cells were co-transfected with synthetic sgRNA and a Cas9 plasmid, followed by transfer of the medium to a 384-well plate to which the NLuc substrate was added. The luminescence readout was measured using a plate reader. Created with BioRender.com

2.3.2 Luciferase Reporter Assay

HEK293T cells were maintained in DMEM (1X) + GlutaMAX (gibco, cat. #31966-021) supplemented with 10% FBS (Thermo Scientific, cat. #A5256701) and 1% penicillin-streptomycin at 37 ^{o}C in a humidified atmosphere with 5 % CO_2. Cells were passaged every 2-3 days upon reaching 80-90 % confluency.

Several HEK293T reporter cell lines have been generated in-house, each containing a distinct reporter cassette integrated into the *HBEGF* locus. Each cassette contained a unique PAM sequence to enable assessment of Cas9 variant activity against different PAM sequences.

HEK293T reporter cells were seeded in 96-well plates at a density of 1.5 * 10^4 in 100 µL complete DMEM and incubated for 24 hours to allow the cell to attach properly. Cells were then co-transfected with mammalian Cas9 plasmids (100 ng/well) and synthetic sgRNA (0.3pmol/well) using Lipofectamine 3000 (Thermo Fisher, cat. #L3000008) according to the manufacturer's protocol.

After 48-72 hours post-transfection, 7.5 µL of cell medium was transferred to a 384-well plate. Luminescence was measured immediately after dispensing 7.5 µL PBS containing NanoLuc substrate (Promega, cat. #N1110) at a 1:1000 (substrate:PBS) dilution using a microplate reader (Molecular DEVICES, SpectraMax iD3). The general workflow of the reporter assay is shown in Figure 2.6.

For selected samples, genomic DNA was extracted using QuickExtract DNA extraction solution (Lucigen, cat. #QE09050) according to the manufacturer's instructions. The targeted *HBEGF* locus was amplified by PCR using barcoded primers (Table B.2). Amplicons were purified using AMPure magnetic beads (Beckman Coulter, cat. #A63881) and sequenced with paired-end reads on an Illumina NextSeq500. Sequencing data were analyzed using an in-house software called RIMA4 to determine the Cas9 editing efficiency [46].

2.4 PAM assay

2.4.1 Overview of the PAM assay

The PAM assay is a novel assay inspired by the previously established *high-throughput PAM determination assay* (HT-PAMDA) by Walton, et al. [47]. HT-PAMDA quantifies PAM sequences not recognized by Cas enzymes and infers PAM preference based on depletion. The novel PAM assay developed in this thesis instead utilizes a dual-approach that quantifies both the recognized and non-recognized sequences. This dual-approach allows for determining the PAM preference through both inference and direct measurement.

The assay begins with a cleavage reaction in which a Cas9 enzyme and sgRNA are incubated with a dsDNA substrate library containing randomized PAM sequences. The cleaved substrates are separated from the uncleaved substrates and both populations are purified using magnetic beads. The cleaved substrates are then enzymatically treated to generate blunt ends and undergo dA-tailing to add an adenosine at each 3' end. This enables ligation of a dsDNA extender containing a complementary thymine overhang at the 3' end of the upper strand. Both cleaved and uncleaved substrates are subsequently amplified with barcoded primers and submitted for amplicon sequencing. Figure 2.7 depicts the main steps of the PAM assay, while a detailed description can be found in the next subsection.

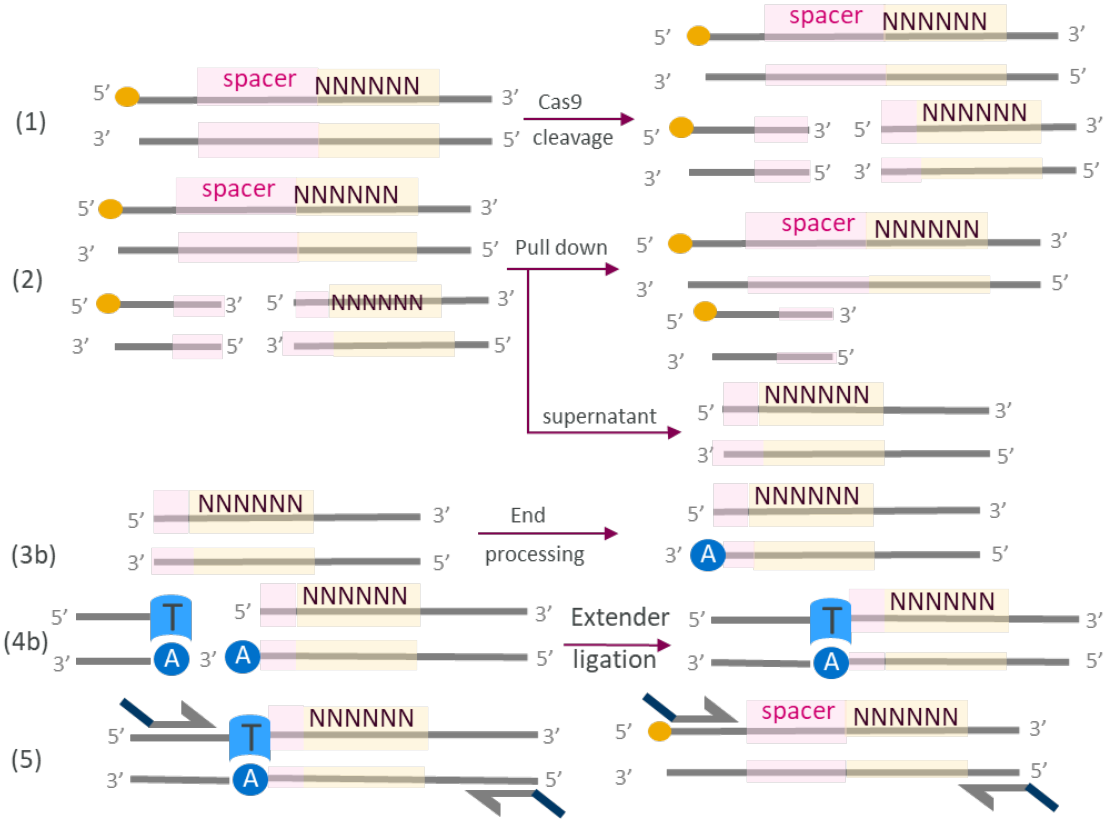


Figure 2.7: Schematic of the PAM assay workflow. The following steps of the PAM assay are depicted: (1) Cas9 RNP cleavage reaction of substrate, (2) Separation of cleaved and non-cleaved samples using magnetic beads, (2a) substrate products bound to the magnetic beads, (2b) substrate product in the supernatant, (3b) enzymatic end treatment (blunt ending, and dA-tailing) of the cleaved substrate, (4b) Extender ligation to the end treated cleaved substrate, and (5) amplification of substrates with barcoded primers.

2.4.2 The PAM Assay

Two different substrate libraries containing different protospacer sequences (EMX1b and PCSK9) were generated using a protospacer sequence from the *EMX1* gene (EMX1b) and a protospacer sequence from the *PCSK9* gene (PCSK9). The substrate libraries were generated by amplification using 2X KAPA HiFi HotStar ReadyMix (Roche, cat. #07958927001), an antisense oligonucleotide (oCN002 or oCN003, Tab. B.3), and a primer with a conjugated dual-biotin at the 5' end (pCN002 and pCN003, Tab. B.3). Following amplification, the reaction mixes were purified by gel extraction.

Three reaction mixtures were prepared according to table 2.2. The cleavage buffer consisted of HEPES pH7.5 (10 mM), NaCl (150 mM), MgCl (5 mM), and nuclease-free water. For experiments that used Cas9 enzyme expressed in cell lysate (see Subsection 2.4.3) 1 µL of cell lysate was used instead. For reactions containing the untreated substrate library, the Cas9 enzyme and sgRNA in mixture B was substituted with nuclease-free water.

Table 2.2: Contents of reaction mixtures in the cleavage reaction.

Mixture A	**Mixture B**	**Stop reaction**
Substrate library (50 nM)	Cas9 enzyme (100 nM)	Proteinase K (2 mg /mL)
Cleavage buffer (1x)	sgRNA (0.2 µM)	EDTA (50 mM)
Nuclease-free water	Nuclease-free water	Nuclease-free water
2.75 µL	2.25 µL	5 µL

The cleavage assay was performed according to the procedure shown in Table 2.3. The Cas9 enzyme and sgRNA first form a ribonucleoprotein (RNP) complex (step 1), which then cleaves substrates containing specific PAM sequences (step 2). Subsequently, proteinase K is added to digest the Cas9 enzyme (step 3) before being heat-inactivated (step 4).

Table 2.3: Procedure for performing the cleavage reaction.

Step	**Procedure**
1	Mixture A and B were incubated at 37^o for 15 min.
2	Mixture A added to mixture B and incubated at 37^o for 30 min.
3	5 µL of the Reaction stop buffer added to the reaction and incubated at 45^oC for 20 min.
4	Reaction incubated at 95^oC for 10 min

After the cleavage reaction, the reaction mixtures were incubated with 25 µg Dynabeads ™ Streptavidin Magnetic Beads (Thermo Fisher Scientific, cat. #11205D). The Magnetic Beads were prepared according to the manufacturer's instructions. The supernatant, containing the cleaved substrates, was transferred to another tube and purified using AMPure XP Beads (Beckman Coulter, cat. #A63880), while the streptavidin-bound substrates (the uncleaved substrates) were further purified using

the Dynabeads. Respective purification was performed according to each manufacturer's instruction.

Blunt ending, dA-tailing, and extender ligation of the cleaved substrates was performed using the NEBNext Ultra II DNA library prep kit for Illumina (New England Biolabs, cat. #E7645S). The Illumina adapter in the original protocol was replaced with the dsDNA extender. The extender annealing was carried out by mixing 1 nmol of each reverse complementary oligonucleotide (oZA080 and oZA081) at 95^{o}C for 5 min followed by a 4^{o}C hold step. After extender ligation, the samples were purified using AMPure XP beads.

Amplification of the cleaved and uncleaved samples was performed using the NEBNext Ultra II Q5 Master Mix, which is part of the NEBNext Ultra II DNA library prep kit, and barcoded primers. For the uncleaved substrates primers nCN002 (EMX1b) or nCN003 (PCSK9) with nZA111 were used and for the cleaved-extender ligated substrates nZA101 and nZA111 (Table B.3). After amplification, reaction mixtures purified using AMPure beads, eluted with Qiagen Elution Buffer (QIAGEN, cat. #19086) and submitted for amplicon sequencing with paired-end reads on an Illumina NextSeq500.

For the cleaved sample, the forward primer binds to the extender, while for the uncleaved and control sample, the forward primer binds upstream of the protospacer sequence. The binding sequence of the reverse primer is the same for all samples and binds downstream of the degenerative sequence. Unique combinations of barcodes from the forward and reverse allows for differentiation of samples from sequencing results when pooling samples together.

2.4.3 Cas9 Cell Lysate

HEK293T cells were seeded in 6-well plates at a density of 8.0 * 10^5 in 100 µL complete DMEM and incubated for 24 hours to reach approximately 70-80 % confluency. Cells were then transfected with mammalian Cas9 plasmids (1 µg/well) using Lipofectamine 3000 according to the manufacturer's protocol. 24 hours after transfection, cells were washed with PBS, incubated with 50 µL of lysis Buffer for 15 min, and stored at -80^{o}C. The lysis buffer was prepared according to table C.1 in Appendix C.2.

2.4.4 Bioinformatic Analysis

Paired-end amplicon sequencing was performed by the NGS facility at AstraZeneca. All reads were demultiplexed based on barodes using bcl2fastq (Illumina). Paired-end reads were merged using FLASH (https://github.com/ebiggers/flash). The merged reads were processed using custom python scripts (Python 3.12.12).

Merged reads were first filtered based on length, where only reads between 137-nt and 150-nt were kept for further analysis. A motif search was performed to identify the target sequence "ACCACAAACCCACGAGGGCAGA", and the six nucleotides

immediately upstream of this motif were extracted for PAM analysis. Below is an illustration of a general read sequence from a cleaved or uncleaved substrate:

5'-[38-nt or more]**NNNNNN**ACCACAAACCCACGAGGGCAGA[71-nt]-3'

The PAM nucleotide distribution heatmaps (log2 fold change) were plotted as follows: (1) extracting the first 2 nucleotides and the following 2 nucleotides from each PAM, (2) summing the frequency of the unique 4-nt combination, (3) repeat extraction and summing the frequency of the unique 4-nt combination from an untreated substrate library, (4) use the formula in equation 2.2 to calculate the $\log_2$ fold change, (5) visualize the log2 fold change values as heatmaps displaying 2-nt x 2-nt heatmaps. The same procedure was used to visualize the $\log_2$ fold change of the unique 6-nt nucleotide combination (3-nt x 3-nt heatmap). Plots were made using pandas (2.3.3) and matplotlib (3.10.8). Note that a pseudocount of $1*10^{-7}$ was added to avoid division by zero for absent sequences (eq. 2.2).

$$\log_2 \mathrm{FC} = \log_2 \left(\frac{f_{\mathrm{treated}} + 1 * 10^{-7}}{f_{\mathrm{control}} + 1 * 10^{-7}} \right) \tag{2.2}$$

2.5 Benchmark Experiment

Twelve genomic target sequences were selected from Walton et al. (2020), which engineered and characterized SpCas9 PAM-relaxed variants [48]. To assess editing efficiency across different PAM sequences, three target sites were selected for each of four PAM sequences: NGG, NTG, and NGC. Amplicon sequencing primers flanking each target site and sgRNA spacer sequences for SpCas9 were obtained from the original publication and are provided in Table B.4-B.5. For ePsCas9 experiments, the sgRNA spacer sequences were modified by adding 2 extra nt at the 5' end of the spacer sequence and pairing it with the ePsCas9 scaffold sequence (Tab. B.5.)

HEK293T cells were maintained as described in section 2.3.2 and passaged using Trypsin-EDTA (Gibco, cat. #15400054) every 2-3 days upon reaching 80-90 % confluency. Co-transfection with mammalian Cas9 plasmids and synthetic sgRNA was also performed as described in section 2.3.2. Genomic DNA was extracted from transfected cells after 72 hours using QuickExtract DNA extraction solution (Lucigen, cat. #QE09050). The targeted loci were amplified by PCR using barcoded primers for each targeted locus (Tab. B.4).

Amplicons were bead-purified using AMPure magnetic beads and sequenced with paired-end reads on an Illumina NextSeq500. Sequencing data were analyzed using an in-house software called RIMA4 to determine the Cas9 editing efficiency [46].

3
Results and Discussion

This chapter describes the results for the directed evolution, luciferase reporter assay, the PAM assay, and comparison with established SpCas9 PAM-relaxed variants. Due to confidentiality, no mutations obtained from the directed evolution will be disclosed.

3.1 First Generation of Evolution

A crucial step in directed evolution is to generate plasmid libraries that contain a large number of unique variants. To ensure a sufficient library diversity, the transformation efficiency was calculated according to equation 2.1 (see results in Table 3.1). The transformation efficiencies for Cas9 plasmid libraries L1-L4 was about 100 times lower than what the manufacturer claims is possible for about the same amount of plasmid. However, the benchmarking is performed using pUC19, a plasmid almost 4-times smaller than the Cas9 library plasmid used, which should reduce the possible maximum transformation efficiency [49]. Another approach for estimating the theoretical complexity of the plasmid library is by multiplying the number of colonies by the dilution and the number of transformations for each library. The theoretical complexity of the library, for example, L2 is 1.95 * 10^7 clones, which is about 10 times higher than reported by B. Kleinstiver, et al. (2015) to obtain evolved SpCas9 PAM-engineered variants [40]. This confirms a sufficient theoretical library complexity for L1-L4, following second generation libraries generated using the same methodology were assumed to have a sufficient library diversity. For more details on library generation see Method 2.1.1.

Figure 3.1 shows an overview of the first-generation library generation, bacterial selection, and validation. The Cas9 plasmid library was transformed into selection cells, six surviving colonies from each PAM selection were pooled, plasmid pools were extracted, and re-transformed into the same selection cells as a validation step,

Table 3.1: Number of colonies obtained from 20 000x dilution and the TE for one of four transformations from the generation of library L1, L2, L4. For L3 the average of five transformations are calculated.

Plasmid library:	**L1**	**L2**	**L3_avg**	**L4**
Colonies (20 000x dilution):	13	108	294	243.6
Transformation efficiency (cfu):	3.7e+6	3.1e+7	8.4e+7	7.14e+7

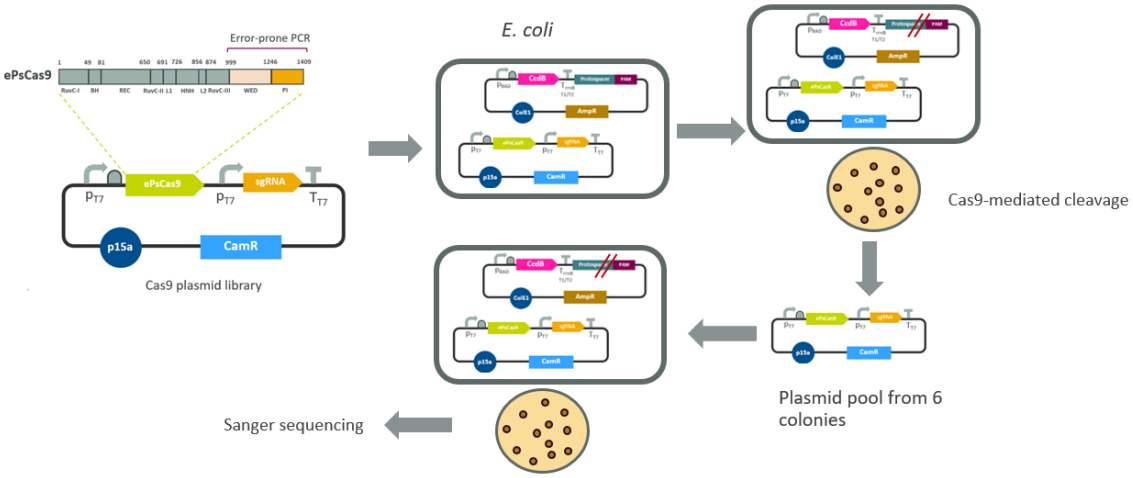


Figure 3.1: Overview of the first generation of evolution. Competent *E. coli* cells harboring a selection plasmid were transformed with Cas9 plasmid libraries. The surviving colonies were pooled into groups of six colonies from the same PAM selection, after which the plasmids were extracted and retransformed. Surviving colonies from the retransformation were then sequenced.

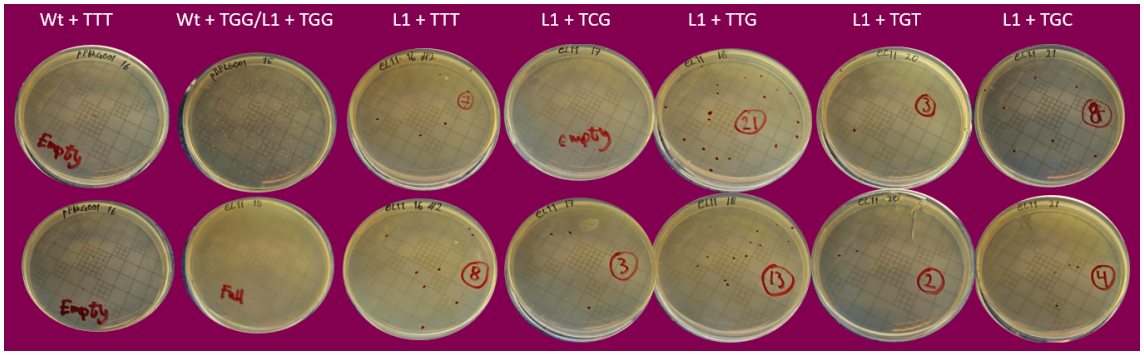
from which individual colonies were picked for further characterization. The Cas9 plasmid library encodes an sgRNA that targets a protospacer sequence, which is followed by a specific PAM sequence, on the selection plasmid. If the Cas9 variant encoded in the library plasmid confers recognition to a specific PAM sequence, the selection plasmid will be cleaved allowing the cell to survive.

Four different libraries targeting the same region of the WED and PI domain with different mutation rates were generated, transformed with the bacterial selection system, followed by re-transformation with selection-cells. Ten colonies surviving the re-transformation from each 6-plasmid pool were sequenced. Figure 3.2 illustrates the selection of ePsCas9 plasmid libraries L1-L4 transformed into TTT-, TGG-, TCG-, TTG-, TGT-, and TGC-selection cells. For each library, a plasmid expressing wild-type (wt) ePsCas9 was transformed into the selection-cells (first and second column in Fig.3.2a-d) to ensure the functionality of selection system. Each Cas9 plasmid library was also transformed into TGG-selection cells. It is worth noting that the wt control on the TGG-plates had fewer colonies for the L1-selection than expected and that a newer batch of control plasmids was used as for L2-L4 resulting in more surviving colonies. It is noticeable that L1 resulted in fewer surviving colonies than L2-L4, underlining the importance of ensuring a library with sufficient diversity in directed evolution.

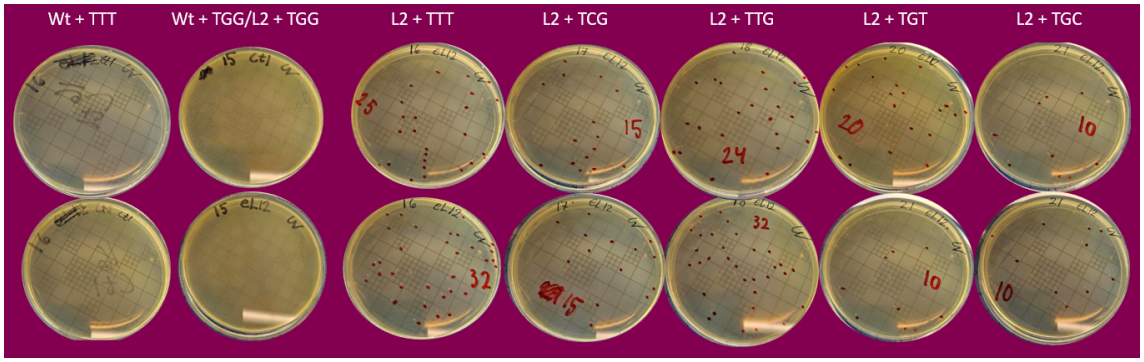
Approximately 40 out of 100 plasmid pools re-transformed into their respective selection cells resulted in surviving colonies, and plasmids from surviving colonies were extracted and sequenced. Sanger sequencing identified 39 unique variants from L2-L4 and 9 unique variants from L1. More than half of all identified variants originated from TTG-selection cells, and harbored a mutation interacting with the second PAM nucleotide. Eight TTG-variants (pCN1-pCN8) with 4-7 amino acid mutations and few consensus mutations were identified and cloned into mammalian

expression vectors. Previously established Cas9 PAM relaxed-variants contain 6-8 amino acid mutations [40, 27, 48].

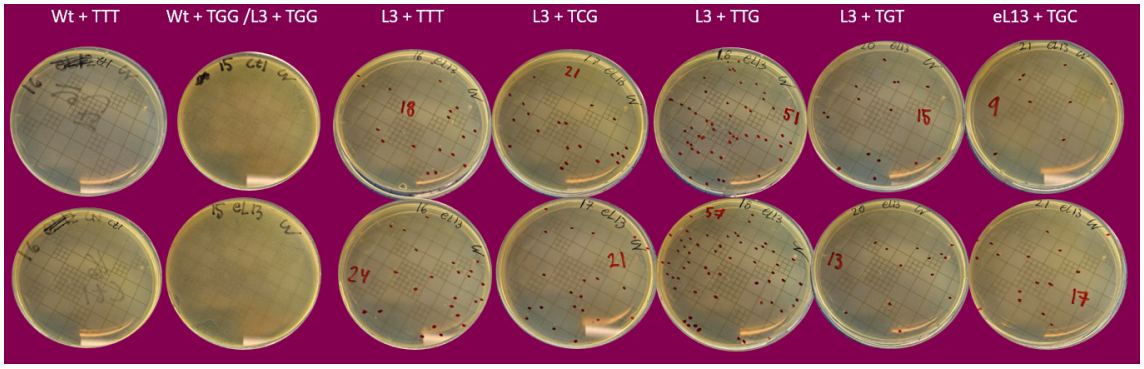
(a)



(b)



(c)



(d)

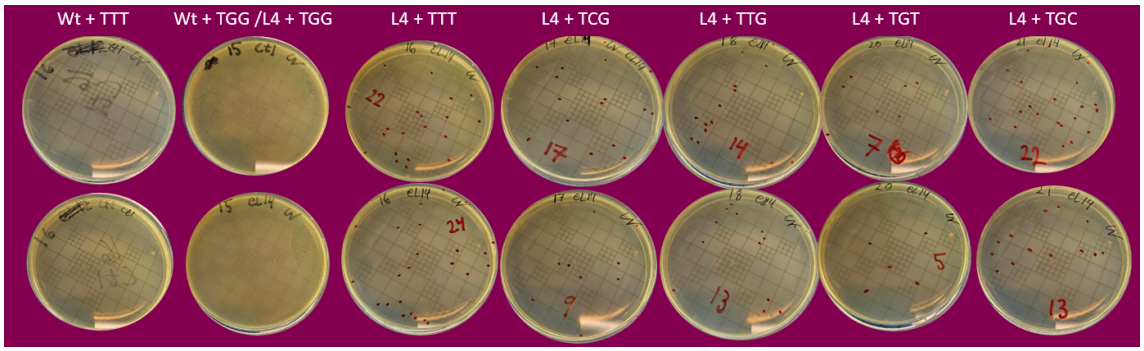


Figure 3.2: Selection of ePsCas9 plasmid libraries (a) L1, (b) L2, (c) L3, and (d) L4 on selection plasmids containing different PAMs.

3.2 Screening of Cas9 Variants in the Luciferase Reporter Assay

The bacterial survival selection system demonstrates high efficiency in isolating ePsCas9 variants capable of recognizing alternative PAM sequences. However, enhanced fitness conferred in *E. coli* does not invariably translate to equivalent benefits in mammalian cell systems. The eight selected ePsCas9 TTG-recognizing variants (pCN1-pCN8) cloned into a mammalian expression vector, where the Cas9 is driven by a strong CMV promoter, were co-transfected with synthetic sgRNA using HEK293T cells with the luciferase reporter system integrated at the *HBEGF* locus (see Fig. 2.6 for workflow) [25, 50].

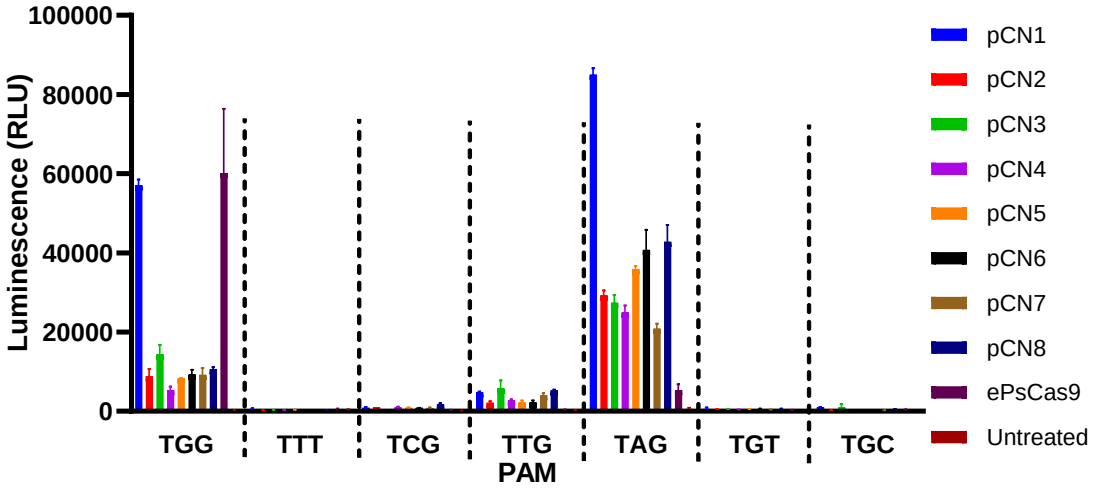
The luciferase reporter system (Fig. 2.5) is a gene cassette containing a strong EF1a promoter that drives expression of a coding sequence (CDS) harboring a premature STOP codon flanked by MMEJ microhomologies, with a downstream ribosomal skipping element and nanoluciferase (*NLuc*) reporter. A sgRNA spacer sequence was designed to span the upstream MMEJ microhomology and the STOP codon such that Cas9 variant recognition of the specific PAM sequence induces a double-strand break (DSB) between the two MMEJ microhomologies. As MMEJ microhomologies flank the DSB site, the MMEJ repair pathway is strongly favored over alternative repair mechanisms. MMEJ-mediated repair results in excision of the STOP codon, thereby enabling the translation of *NLuc*.

Figure 3.3 shows the luminescence output and total editing efficiency for ePsCas9 variants pCN1-pCN8 and wild-type ePsCas9 in the seven PAM-specific luciferase reporter cell lines. Untransfected cells served as a negative control. Luminescence was measured in relative light units (RLU), while total editing efficiency was determined by amplicon sequencing to quantify all indels. This screen was performed with one biological replicate and two technical replicates per PAM-reporter cell line. The PAM reporter cell lines used contained TGG, TTT, TCG, TTG, TAG, TGT, and TGC.

In the TGG-reporter cells, wild-type ePsCas9 demonstrated the highest mean luminescence activity of $6.0 * 10^4$ RLU while only the second highest mean total editing efficiency of 12%, followed by pCN1 in mean luminescence activity of $5.7 * 10^4$ and proceeded in mean total editing efficiency of 21%. In the TAG-reporter cells, pCN1 demonstrated the highest performance, achieving mean luminescence of $8.5 * 10^4$ RLU and 28% mean editing efficiency, followed by pCN8 (mean: $4.0 * 10^4$ RLU, 15% editing efficiency). In the TTG-reporter cells, pCN3 demonstrated the highest performance, achieving $5.7 * 10^3$ RLU and 7.8% editing efficiency, followed by pCN8 (mean: $5.2 * 10^3$ RLU, 6.1% editing efficiency), and pCN1 (mean: $4.8 * 10^3$ RLU, 7.8% editing efficiency). The luminescence activity measured below $1 * 10^3$ RLU for the remaining PAM-reporter cells, while the total editing efficiency below 3%. Wild-type ePsCas9 (mean range: 350-420 RLU, 0-3.1% editing efficiency) showed minimal activity across all PAM-reporter sequences, except for TGG, confirming the necessity of engineering mutations to relax the current PAM recognition.

Differences between luminescence output and total editing efficiency were observed across multiple variants and PAM-reporter cells, although the magnitude varied. In the TAG-reporter cells, pCN7 exhibited approximately half the mean luminescence of pCN8 (mean: $2.1 * 10^4$ RLU vs. $4.3 * 10^4$) despite achieving similar editing efficiencies (mean: 12% vs. 15%). These differences likely reflect differences in the indel spectrum generated by each variant at different target sequences, as the luciferase reporter system responds to deletions that excise the STOP codon through MMEJ repair or NHEJ, whereas insertions or other indel types that maintain reading frame cause no luminescence signal.

(a)



(b)

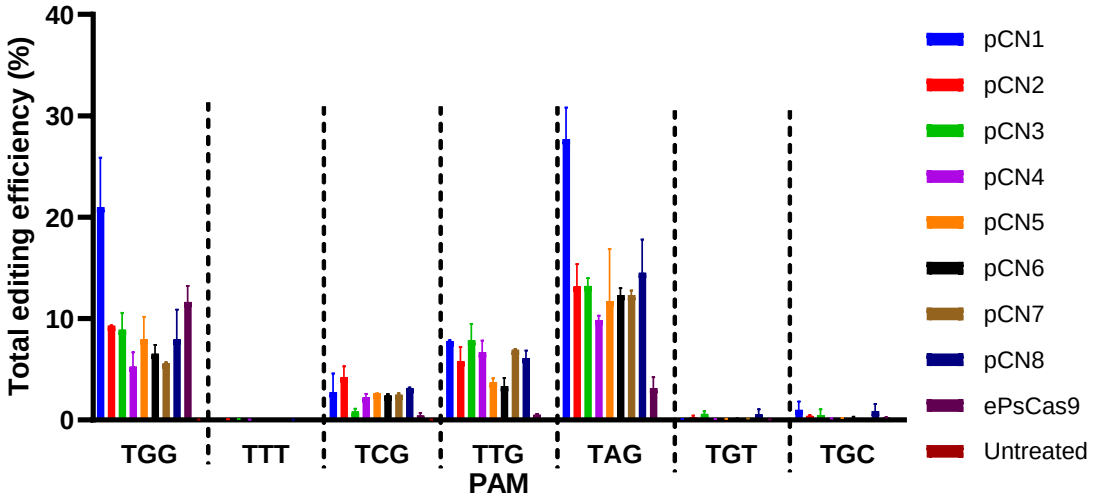


Figure 3.3: pCN1-pCN8 luciferase reporter assay: (a) luminescence output and (b) total editing efficiency. Bars represent mean ± standard deviation (SD) from 2 technical replicates.

3.3 Second Generation of Evolution

The bacterial survival selection system identifies Cas9 variants capable of recognizing and cleaving specific PAM sequences on the selection plasmid. However, a key limitation is that the system is not designed to select variants based on their degree of activity. This limitation could potentially be overcome through iterative enrichment (see Fig. 3.4): pooling all variants from surviving colonies, re-transforming the resulting plasmid library, and repeating the selection process multiple times. Each successive round would favor variants with higher cleavage efficiency, as more efficient variants would generate larger surviving populations. This enrichment strategy will subsequently enable the selection and mutagenesis of second-generation variants with optimized activity at specific PAM sequences.

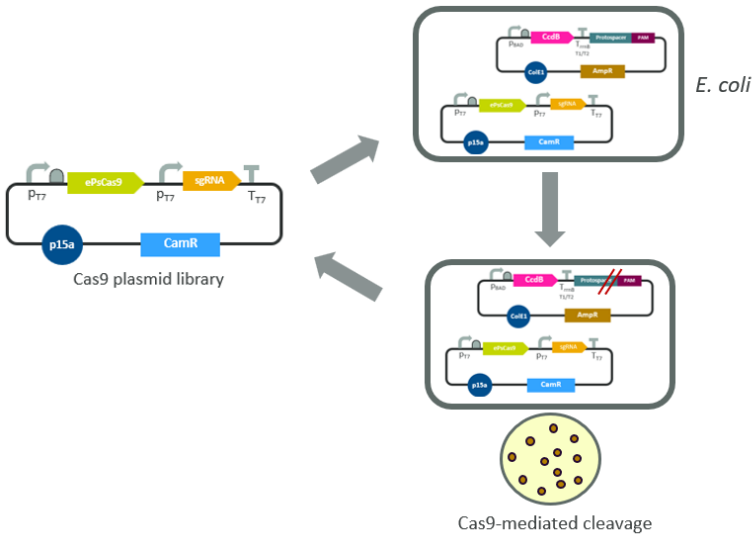


Figure 3.4: Principle of enrichment strategy. The Cas9 plasmid libraries were transformed into the bacterial selection system. The surviving colonies were pooled into a single mixture, after which the plasmids were extracted and retransformed. Multiple rounds of transformation, colony pooling, and retransformation were performed to enrich for the most efficient variants.

To evaluate this iterative selection strategy, second-generation mutagenesis plasmid libraries underwent four successive rounds of transformation and selection using four distinct approaches. Approach A used a single TTG-selected variant containing a conserved amino acid mutation found in most TTG variants, followed by Wedge and PI domain mutagenesis. Approach B pooled two TGC-selected variants from the first generation and performed a second round of Wedge and PI domain mutagenesis. Approach C pooled all plasmids from the first-generation of evolution and re-mutagenized the Wedge and PI domains. Approach D retained the Wedge and PI domains from variants pCN1-pCN8 while mutagenizing the remaining upstream coding sequence. Details for each library generation can be found in Method 2.1.1.

Approach A was designed specifically to enhance the efficiency for recognizing TTG PAM and approach D to enhance overall efficiency of the Cas9 enzyme, both used

TTG-selection cells exclusively throughout all four rounds. Approach B focused on TGC PAM recognition and utilized TGC-selection cells across the entire iterative selection process. The library generated using approach C was initially transformed into all PAM-selection cells, but surviving colonies were only recovered from TTG-plates after the first round. This indicates that the pooled library from approach C had converged toward TTG specificity. The remaining three rounds for approach C therefore employed only TTG-selection cells.

Ten colonies resulting from each approach (A-C) were selected and sequenced. All unique variants were cloned into a mammalian expression vector (see Method 2.1.1). For approach A, most variants mainly contained the conserved mutation from the first-generation evolution but one variant fulfilled the requirement of 4 mutations (pCN29). Suggesting that while the iterative selection for the conserved beneficial mutation, the stringency was insufficient to consistently drive accumulation of additional beneficial mutations. For approach B and C, four variants (pCN24-pCN28) and five variants (pCN19-pCN23, and pCN29), respectively, reached the requirement. No variants from approach D were sequenced at this stage, instead undetermined variants (pCN9-pCN18) from 10 colonies were directly cloned into mammalian expression vectors (see Method 2.1.1).

Figure 3.5 shows the luminescence output of second-generation variants (pCN9-PCN29) derived from Approaches A-D and SpCas9 and ePsCas9, evaluated across the PAM-specific reporter cell lines, with untransfected HEK293T cells as negative control. This screen was performed with one biological replicate and three technical replicates per variant and reporter cell line. Second-generation variants showed diverse outcomes with a high dependency of which approach was used for generating the Cas9 plasmid library. In the TAG-reporter cells, variants generated using approach D demonstrated the highest mean luminescence activity of $1.2 * 10^5$ RLU for pCN12 and pCN14 and $1.1 * 10^5$ RLU for pCN15. In comparison, the highest activity in TAG-reporter cells for variants from approaches A-B was measured for pCN23 with mean luminescence of $7.8 * 10^4$ RLU. The highest activity in TTG-reporter cells was also measured for variants from approach D: pCN9, pCN12, and pCN14 with a mean luminescence of $1.7 * 10^4$, $1.6 * 10^4$, $8.9 * 10^3$ RLU, respectively. In the TGC-reporter cells, pCN24 and pCN27 generated using approach B, clearly demostrated the highest activity reaching $1.2 * 10^5$ RLU and $2.3 * 10^5$ RLU. In the TGG-reporter cell, all variants generally demonstrated a similar or higher activity than the wild-type ePsCas9 (mean: $2.3 * 10^5$ RLU).

The best performing variants (Fig. 3.5) in the TTG-, TAG-, and TGC-reporter cells from the second-generation were compared to the first-generation variants (Fig. 3.3) in a following comparison experiment. pCN23 demonstrating the highest activity (mean: $8.9 * 10^3$ RLU) in TTG-reporter cells among the variants generated from approaches A-C was also included in the same comparison experiment. Prior to the following experiment, the complete coding sequences of pCN9, pCN12, pCN14 and pCN15 were determined by Sanger sequencing, revealing that each variant represented a unique variant with a distinct mutation profile.

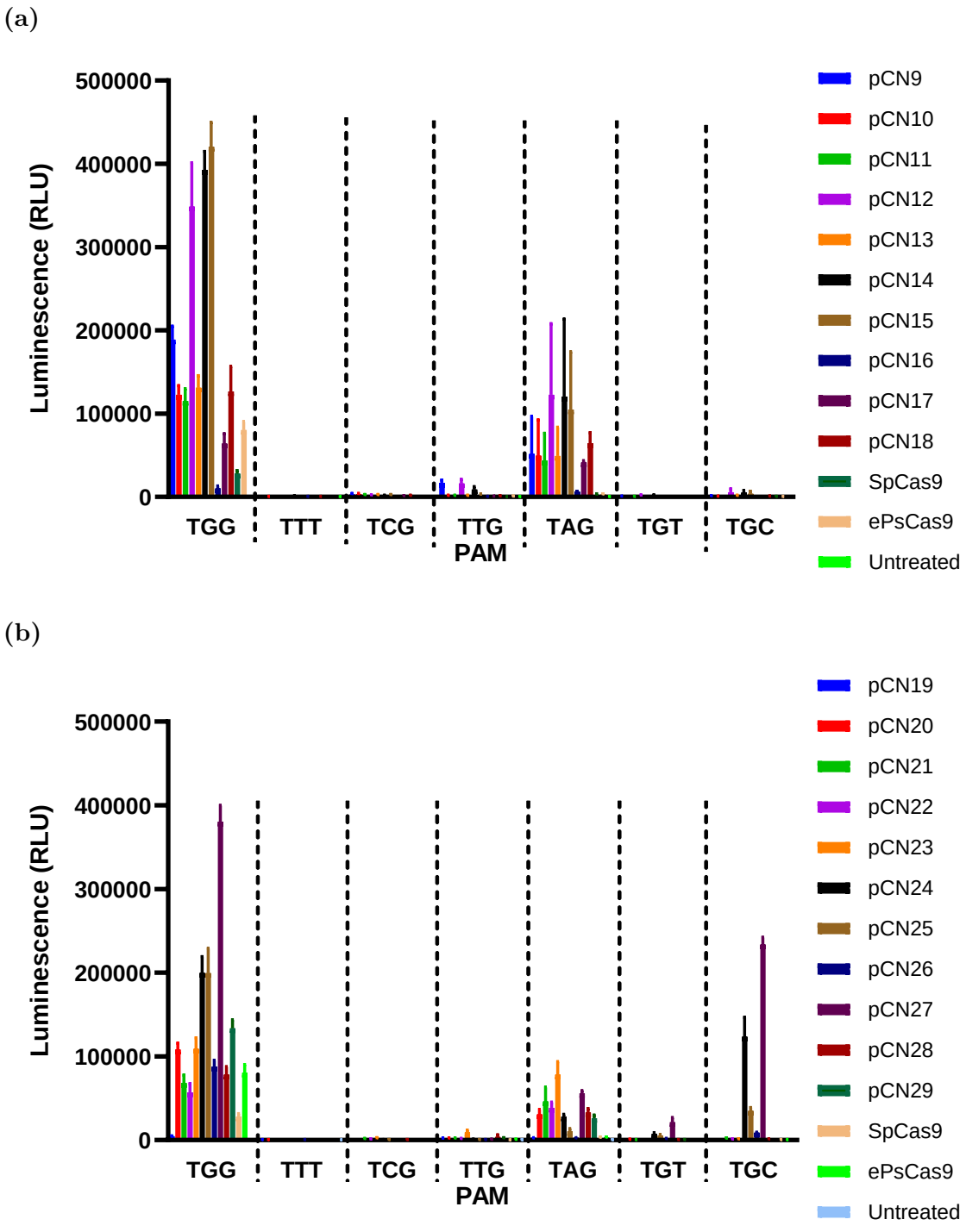


Figure 3.5: Luminescence output from the luciferase reporter assay for the second-generation ePsCas9 variants (a) pCN9-pCN18 and (b) pCN19-pCN29. Bars represent mean ± SD from 3 technical replicates.

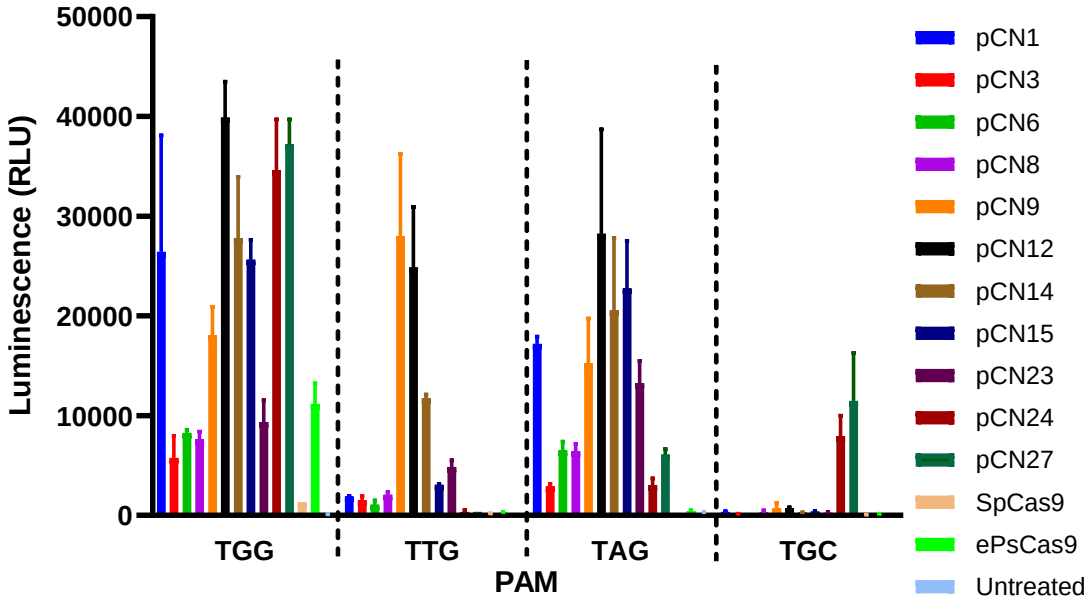
3.4 First and Second Generation variants

To directly compare the performance of the best first and second-generation variants, a subset of lead candidates from each generation was evaluated in the luciferase reporter assay in three selected PAM-specific reporter cells: TGG-, TAG-, TTG-, and TGC-reporter cells. Figure 3.6 presents the luminescence output and total editing efficiency for first-generation variants (pCN1, pCN3, pCN6, and pCN8), and second-generation variants (pCN9, pCN12, pCN14, pCN15, pCN23, pCN24, and pCN27), along with wild-type ePsCas9 and SpCas9 as additional comparison, and untransfected HEK293T as negative control. This experiment was performed with one biological replicate and two technical replicates.

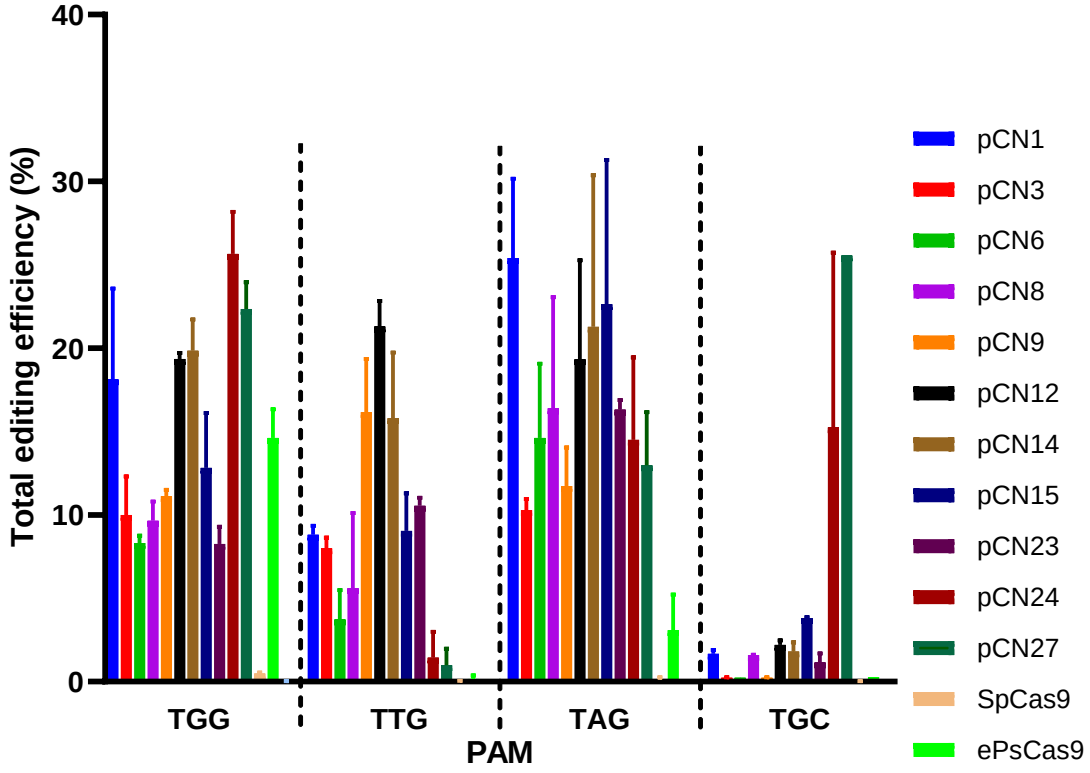
In the TTG-reporter cells, second-generation variants pCN9 (mean values: $2.8 * 10^4$ RLU and 16 % total editing efficiency) , pCN12 (mean: $2.5 * 10^4$ RLU, 21 %), and pCN14 (mean: $1.2 * 10^4$ RLU, 16 %) consistently outperformed all first-generation variants with a difference of 7-12 percentile points to the best performing first-generation variant (pCN1). In the TAG-reporter cells, pCN12 and pCN14 maintained their performance advantage over most first-generation variants, achieving a mean editing efficiency of 19 % and 21 %, suggesting that the wedge and PI domain mutations obtained by these variants result in a general relaxation of the second nucleotide in the PAM.

In contrast, pCN15, which had the second highest mean total editing efficiency of 22 % in TAG-reporter cells, demonstrated considerably lower editing efficiency in TGG- and TTG-reporter cells (13 % and 9 % , respectively). Similarly, pCN1 achieved a 25 % mean total editing efficiency in TGG, and 8.8 % in TTG-reporter cells. For pCN1 and pCN15, these patterns hint at a change in PAM recognition rather than a relaxation at the second position. However, pCN1 still achieved 18 % editing efficiency in TGG-reporter cells. In the TGC-reporter cell lines, pCN24 and pCN27 demonstrated a mean activity of $8.0 * 10^3$ RLU and $1.1 * 10^4$ RLU, respectively, as well as editing efficiencies up to 25 % for both variants. However, for pCN27 the total editing efficiency was only measured for one of two samples, increasing the uncertainty for that sample.

Although these results are based on limited biological and technical replication, there is a strong correlation between the initial screening experiment (Fig. 3.5) of second-generation variants and the comparison experiment (Fig. 3.6) showing that combining the enrichment strategy and the approach D obtained variants with improved efficiency in recognizing TTG and TAG as PAM sequences. In addition, combining the enrichment strategy and approach B led to the discovery of two variants (pCN24 and pCN27) with the ability to recognize TGC PAM sequences with an efficiency comparable to variants recognizing TTG or TAG PAM sequences.



(a)



(b)

Figure 3.6: Comparison between the best performing variants from the first-generation and second-generation of evolution in the luciferase reporter assay: (a) Luminescence output (b) Total editing efficiency. Bars represent mean ± standard deviation (SD) from 2 technical replicates.

3.5 Protein Engineering

A recent article crystallized the structure of the wild-type PsCas9 using cryo-electron microscopy (cryo-EM) [51]. In the same article, Shen, et al. (2025), identified two amino acids (S1223 and I1222) in the wedge domain of the PsCas9 enzyme interacting with the PAM sequence. Shen, et al. (2025), then showed that the S1223A and I1222A mutations conferred increased recognition of TAG in an *in vitro* cleavage assay. Neither S1223A nor I1222A were present in any of the ePsCas9 PAM-relaxed variants generated from the directed evolution system in this Master's thesis. Thus, I hypothesized that introducing these mutations into selected ePsCas9 PAM-relaxed variants could confer additional PAM-relaxation. In Table 3.2, the mutations introduced and which ePsCas9 variants the mutations were introduced into can be seen. By analyzing the ePsCas9 protein structure in PyMOL, I further hypothesized that only mutating S1223A could prove beneficial for relaxing the second PAM nucleotide while keeping more of the interactions with the third PAM-nucleotide. All mutations were introduced into the variants using site-directed mutagenesis (see Method 2.2.2).

To determine whether the newly introduced mutations confered a beneficial effect on relaxing the PAM sequence, the variants with either S1223A or S1223A together with I1222A introduced (pCN30-pCN39) were evaluated alongside selected variants without the new mutations, wild-type SpCas9 and wild-type ePsCas9 in the luciferase reporter assay. The PAM reporter cell lines used contained TGG, TTT, TCG, TTG, TAG, TGT, TGC, and CTG. In Figure 3.7, the luminescence and editing efficiency of the luciferase reporter assay can be seen. This experiment included one biological replicate and two technical replicates.

Due to low bioluminescence activity for the TTT-, TCG-, and TGT-reporter cell lines, these were excluded in the amplicon sequencing submission. In this experiment, a CTG-reporter cell line was additionally included, allowing for comparison with the TTG-reporter cell line. In general, all variants obtained by site-mutagenesis (pCN30-pCN39) indicate decreased or similar efficiency to the variants obtained only from directed evolution (pCN1, pCN9, pCN12, pCN15, pCN24, and pCN27). Therefore, it can be concluded that S1223A and I1222A did not confer any additional benefit for the best performing ePsCas9 variants. One potential explanation for the decreased activity is that the mutations may have disrupted critical contacts with the PAM sequence, reducing the PAM-binding affinity. In addition, similar activities could be seen for the CTG- and TTG-reporter cell lines, with a slight increase for cytosine at the first PAM position.

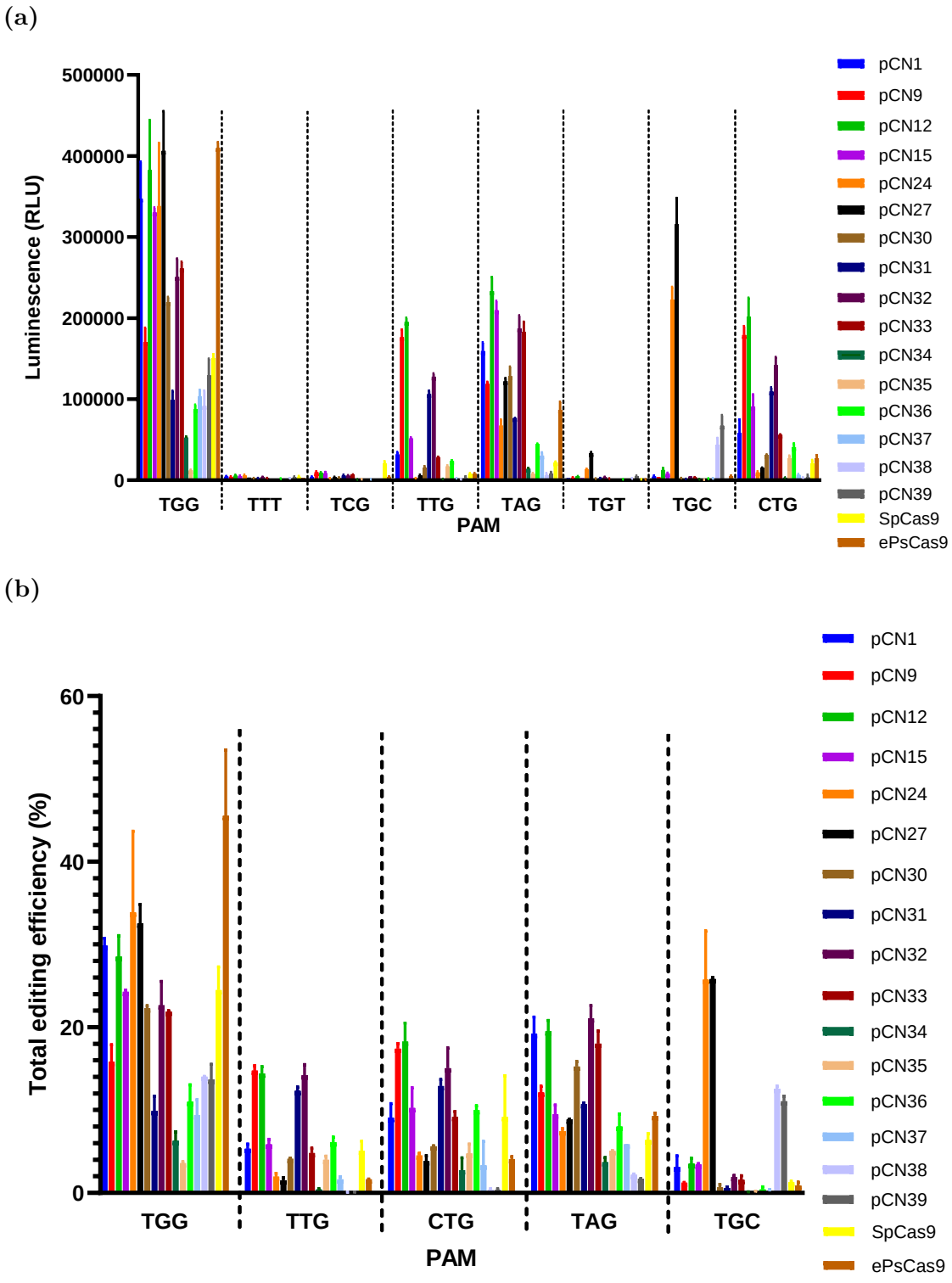


Figure 3.7: ePsCas9 variants with introduced S1223A or I1222A and S1223A mutations in the luciferase reporter assay: (a) Luminescence output (b) Total editing efficiency. Bars represent mean ± standard deviation (SD) from 2 technical replicates.

Table 3.2: ePsCas9 variants generated by introducing either S1223A or S1223A and I1222A into selected variants obtained form the directed evolution system.

ePsCas9 variant	S1223A	S1223A & I1222A
pCN1 (TTG)	pCN30	pCN34
pCN9 (TTG)	pCN31	pCN35
pCN12 (TTG)	pCN32	pCN36
pCN15 (TTG)	pCN33	pCN37
pCN24 (TGC)	-	pCN38
pCN27 (TGC)	-	pCN39

3.6 PAM Assay Development

The PAM assay consists of an *in vitro* Cas9-mediated cleavage reaction using a dsDNA substrate library with a protospacer sequence and 6-nt variable PAM sequences. Amplicon sequencing of both cleaved substrates (cleavage sample) and uncleaved substrates (depletion sample) enables characterization of recognized PAM sequences by their presence, in the cleavage sample, and their absence in the depletion sample. For more details about the PAM assay see Method 2.4.

Initially, a substrate library containing a protospacer of the EMX1a sequence from the *EMX1* gene was used in the cleavage assay together with purified ePsCas9 protein. However, the frequency of the first 4-nt PAM sequences in the cleavage sample had a similar distribution to the untreated substrate library (see Fig. A.2). Although, a small increase for GG at the second position and third PAM position can be observed.

To investigate the cause of the observed similarity in PAM distribution between the cleavage sample and the untreated library, cleaved substrates were aligned to an uncleaved substrate reference using Smith-Waterman alignment [52]. Prior to aligning the reads from the cleaved substrate, the extender sequence was trimmed off (Fig. 2.7). Figure 3.8 shows a subsample of 100 000 reads aligned to the reference sequence. The data were normalized using maximum absolute scaling, where the read count at each cleavage position was divided by the maximum read count observed across all cleavage positions. Note that the cleavage position in Figure 3.8 is assumed to be the same as where the extender is ligated, which in reality could be affected by the end processing, such as during blunt-ending. The bar in Figure 3.8 indicates the cleavage position and the sequence shown on the x-axis is the 22-nt spacer followed by the variable PAM sequence.

The alignment indicated that the substrates were unexpectedly cleaved at -13 from the PAM position (circled in red in Fig. 3.8a). As ePsCas9 usually confers a dsDNA break at -6 or -9 from the PAM position, it was hypothesized that the consecutive guanines at the end of the EMX1a spacer (see Fig 3.8a) could be recognized as an alternative PAM sequence by the Cas9 enzyme. Thus, potentially moving the cleavage site further upstream.

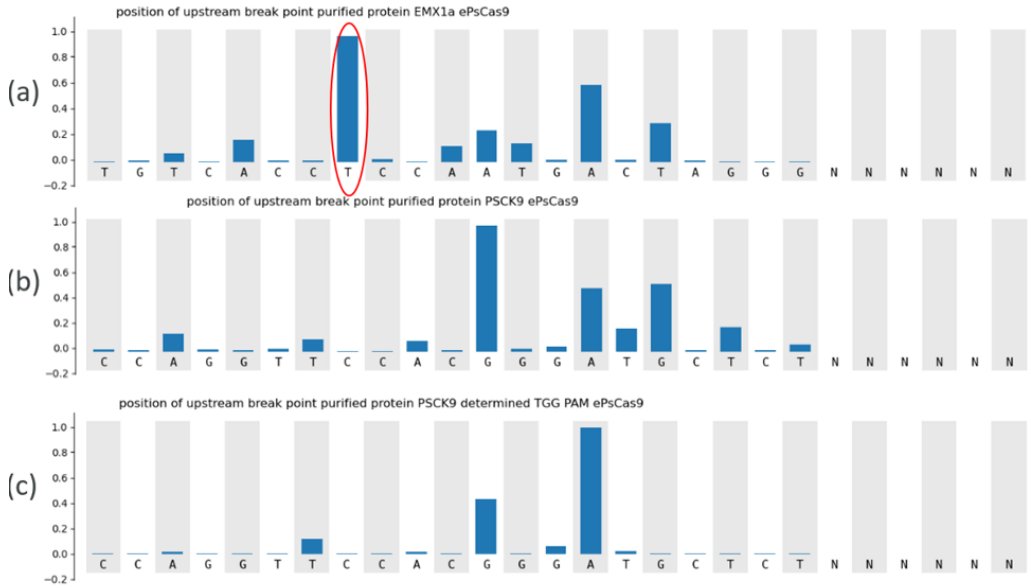


Figure 3.8: Cleavage position of a subsample of 100 000 reads from three different cleaved samples in the PAM assay. The reads from the cleaved samples are aligned to their respective uncleaved substrate library sequence. The substrate libraries consist of (a) the EMX1a protospacer sequence followed by six randomized nucleotides, (b) the PCSK9 protospacer sequence followed by six randomized nucleotides, and (c) the PCSK9 protospacer sequence followed by a determined TGG(TTT) sequence.

3.6.1 Investigating the EMX1a Spacer Sequence

To investigate whether the EMX1a spacer sequence was potentially cleaved independently of adjacent PAM, a fluorescence-based cleavage assay was designed (see Fig. A.3). Two pairs of reverse complementary oligonucleotides, each 5'-labeled with either ATTO647 or FAM, were designed containing the EMX1a sequence followed by TGG or CCC at the PAM position. These oligonucleotides were annealed to form double-stranded substrates, used in cleavage reactions, denatured, and resolved on TBE-Urea gels.

As expected, the fluorophore-bound substrates containing TGG at the position of the PAM were efficiently cleaved by both SpCas9 and ePsCas9, while the control reactions without Cas9 remained uncleaved (see Fig. A.4). Surprisingly, ePsCas9 generated a faint band at the expected size for the cleaved ATTO647-labeled upper strand of the CCC-PAM substrate, while SpCas9 showed no detectable cleavage. This indicates that a substrate with the EMX1a protospacer sequence could be cleaved, at least for ePsCas9, despite not containing a canonical NGG PAM sequence, but potentially using the GGG-motif within the spacer. No obvious band could be seen for the FAM-bound bottom strand of the CCC containing substrate, but it was reasoned that this might be due to the much lower sensitivity of the FAM fluorophore compared to ATTO647 fluorophore.

This suggests that the EMX1a sequence is not ideal for a PAM discovery assay, new substrate libraries carrying a protospacer of either the PCSK9 sequence or the

EMX1b sequence followed by a 6-nt degenerative sequence, as well as two substrates containing a determined TGG were used in the PAM assay. Both spacer sequences lacked consecutive guanines in the last three nucleotides. Figure 3.8b-c shows the PCSK9 spacer sequence upstream of the 6-nt PAM, displayed as either (b) a random sequence or (c) the determined TGG(TTT) sequence. The cleavage reactions using the PCSK9 protospacer sequence (Fig. 3.8b-c) were processed bioinformatically using the methodology previously described. The cleavage position in Figure 3.8c occurs at -9 and -6 nt from the PAM. While Figure 3.8b shows the same cleavage positions at -9 and -6 nt, it also displays additional cleavage sites flanking the expected positions. Interestingly, there is no noticeable cleavage at -13 for PCSK9 as was the case for EMX1a, further supporting the hypothesis that the GGG-motif in the spacer sequence adjacent to the PAM might impact the PAM recognition and potentially the cleavage position.

3.6.2 Interpreting PAM Preference

In Figures 3.9-3.10, heatmaps of the PAM log2 fold change where the first two PAM nucleotides adjacent to the spacer sequence are plotted against the third and fourth PAM nucleotides. Cleaved (extender-ligated) reads below 137 nt or above 150 nt were excluded (Fig. A.5a for EMX1b). To extract the 6-nt PAM sequences, a motif search was performed to locate a known 22-nt sequence positioned downstream of the PAM, and the 6 nucleotides immediately upstream of this sequence were extracted as the PAM. Each unique occurence of the 6-nt PAM was then counted and normalized by dividing each unique 6-nt combination by the total number of 6-nt PAM sequences to calculate the frequency. To calculate the frequency of the first 4-nt PAM, the frequency of all 6-nt PAM sequences that had the same first 4-nt were added. The 4-nt PAM frequency for the untreated substrate library was calculated in the same way and the log2 fold change was calculated by performing a log2 transformation of the 4-nt PAM frequency of the cleavage experiment divided by the 4-nt PAM frequency of the untreated substrate library.

Since SpCas9 and ePsCas9 recognize the canonical NGG-PAM, the heatmap was plotted to display all combinations of the first four nucleotides in the PAM (Fig. 3.10-3.9). As expected, the PAM log2 fold change for SpCas9 (Fig. 3.10) and ePsCas9 (Fig. 3.9) on the EMX1b and the PCSK9 substrate library both have a high log2 fold change for NGG. There is additionally, a vertical and a horizontal line for GGNN and NNGG indicating that some PAM recognition is conferred by two consecutive guanines in the vicinity of the second and third position. A similar pattern of consecutive recognized guanines can be seen in HT-PAMDA [47]. It is possible that this effect could be enhanced by the *in vitro* setting of the assay and the low ratio of NGGNNN-substrate. Looking at the log2 fold change, it appears as if NNGGNN-substrate is generally cleaved at a higher ratio than GGNNNN-substrate.

Furthermore, a weak PAM recognition of NAG is also observed for both ePsCas9 and SpCas9 (Fig. 3.9-3.10), as well as an increase in NNAGNN-substrate. This is likely the same "consecutive-recognized-nt" phenomenon that is seen for two consecutive

guanines. In addition to NAG, for SpCas9 (Fig. 3.10) there also seems to be a weak recognition for NGA, NCG, and NTG with vertical lines at NNGA, NNCG, NNTG likely due to the secondary consecutive-recognized-nt. ePsCas9 seems to be more restrictive where a weaker recognition can be seen at NTG and NCG with corresponding consecutive-recognized-nt at NNTG and NNCG.

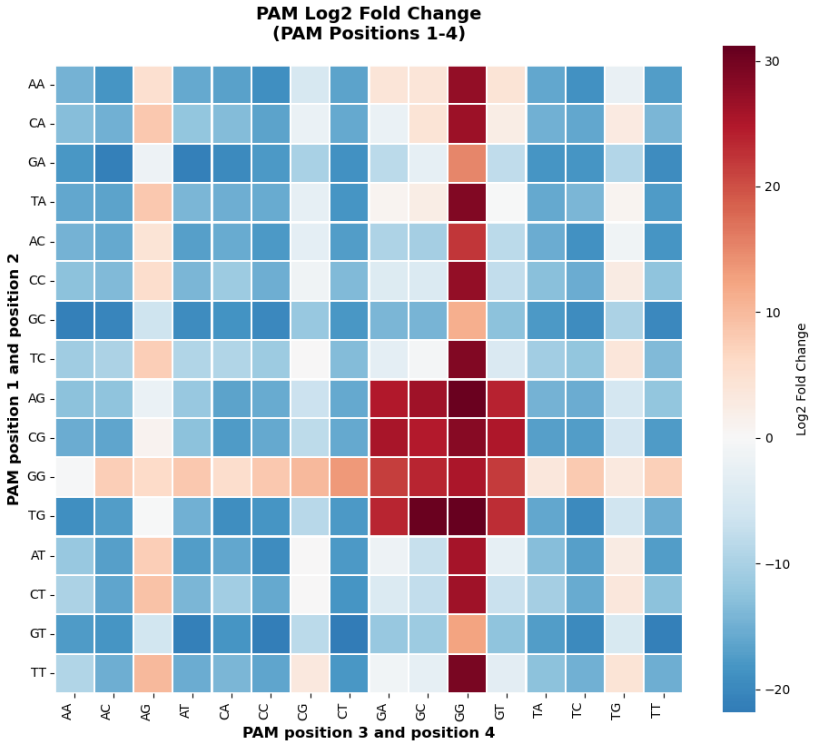
To increase throughput, the PAM assay was performed using ePsCas9 and SpCas9 from mammalian cell lysate rather than purified protein. Both PCSK9 and EMX1b substrate libraries were tested (Fig. A.6-A.7), yielding comparable results to purified protein (Fig. 3.9-3.10). Table 3.3 shows experimental conditions: number of raw reads, number of filtered reads, library type, and Cas9 source (cell lysate/purified protein). In experiments using Cas9 from cell lysate a decrease in the number of good reads after filtering can been seen (Tab. 3.3). This is likely due to endogenous nucleases or exonucleases in the cell lysate that cleave or degrade the substrate library non-specifically. Since any cleavage event results in loss of the biotin label, cleaved substrates remain in the supernatant during streptavidin bead purification and are excluded from sequencing, thereby reducing the total number of filtered reads regardless of whether cleavage was Cas9-mediated.

Table 3.3: The number of total reads obtained from the amplicon sequencing and the number of filtered reads (the reads from which the PAMs are extracted from) for ePsCas9 tested in the PAM cleavage assay.

Purified/lysate	**Substrate library**	**Filtered reads (k)**	**Total reads (k)**	**% filtered reads**
Purified	EMX1b	955	1440	66
Purified	PCSK9	392	568	69
Cell lysate	EMX1b	510	1500	34
Cell lysate	PCSK9	828	1600	52

Depletion samples using purified ePsCas9 and SpCas9 proteins with the EMX1b library (Fig. A.8) produced results consistent with those observed in the cleavage samples. Similarly, depletion samples using ePsCas9 and SpCas9 from cell lysate with the PCSK9 library (Fig. A.9) yielded comparable results to the cleavage experiments. However, two experimental conditions failed: depletion samples using purified ePsCas9 and SpCas9 proteins with the PCSK9 library, and depletion samples using ePsCas9 and SpCas9 from cell lysate with the EMX1b library. Due to time constraints, these experiments could not be repeated.

(a)



(b)

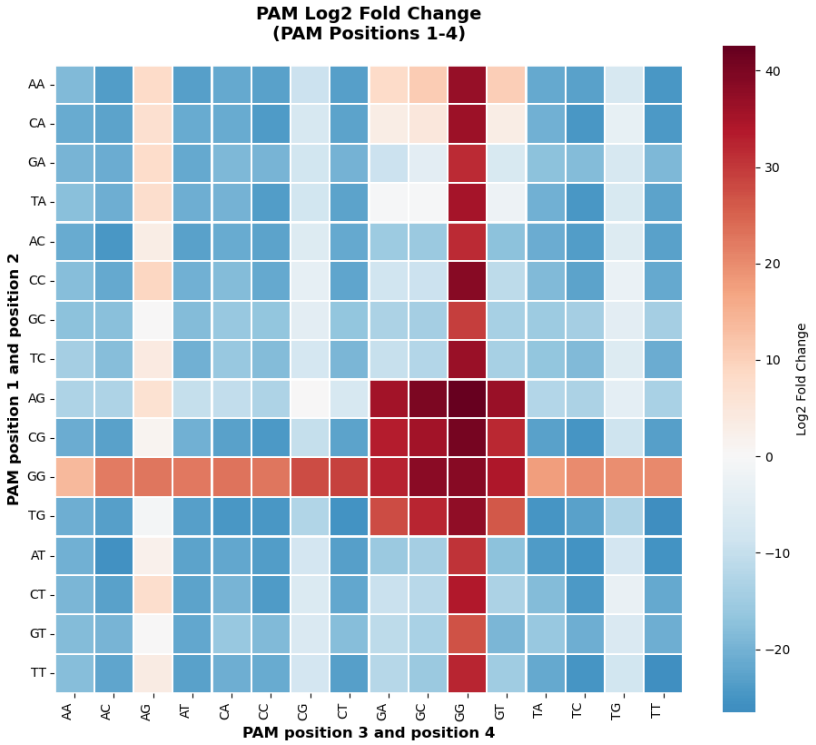
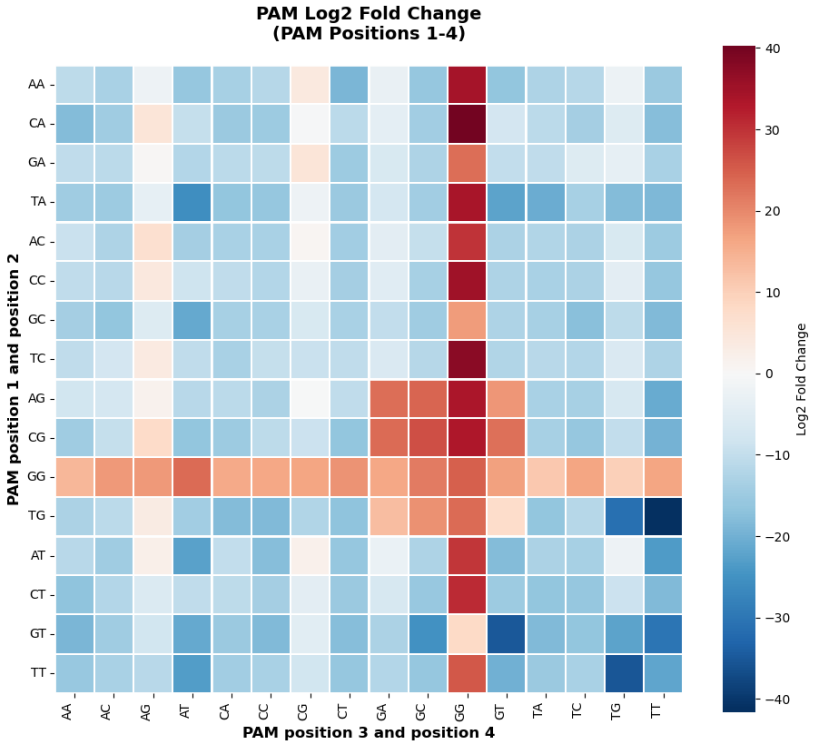


Figure 3.9: The PAM Log_2 Fold Change from the cleavage samples using purified ePsCas9 protein (a) PCSK9 substrate library (b) EMX1b substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

(a)



(b)

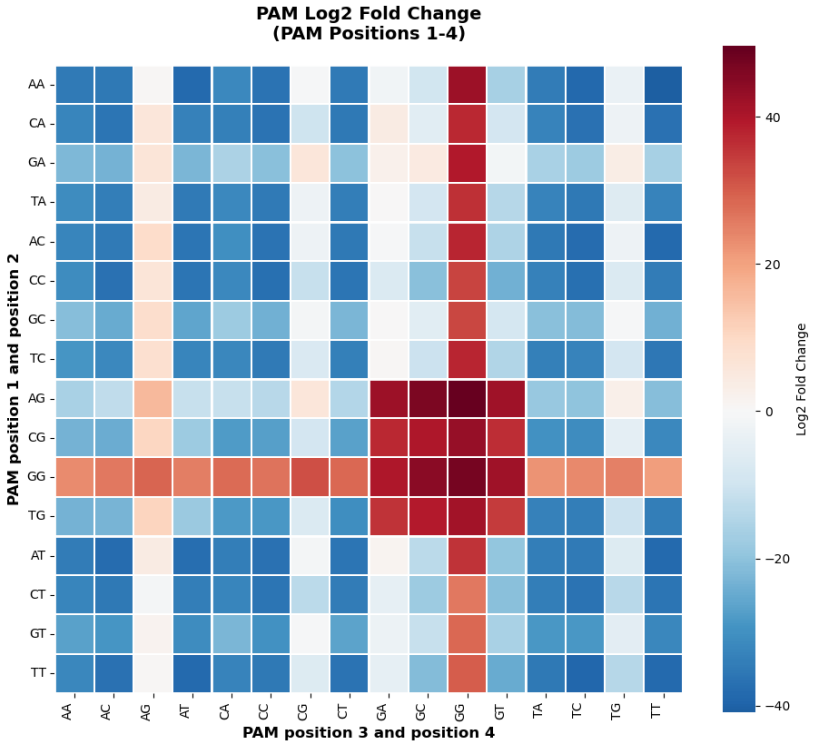


Figure 3.10: The PAM Log_2 Fold Change from the cleavage samples using purified SpCas9 protein (a) PCSK9 substrate library (b) EMX1b substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

3.7 PAM Landscape of a New Cas9 Enzyme

To further evaluate the capability of the PAM assay, an AZ propriatary Cas9 enzyme suspected of recognizing a 6-nt PAM sequence was tested using Cas9 protein expressed in cell lysate. The PAM log2 fold change from the cleavage samples using the EMX1b-substrate library and PCSK9-substrate library (see Fig. 3.11-3.14) was extracted as previously described and calculated by log2 transforming the 6 nt PAM experiment frequency by the 6 nt PAM untreated substrate library. In Figures 3.11-3.14 a 3-nt x 3-nt heatmap of the PAM log2 fold change, where the y-axis is the three first nucleotides adjacent to the spacer sequence and the x-axis the last three nucleotides.

In Figures 3.11 and 3.12 from the cleavage samples, the highest log2 fold change can be observed at NNNGAA with a lower log2 fold change for NNNGRV, where R is A or G and V is A, C, or G. There also seems to be a pattern of possible secondary consecutive-recognized-nt as there is an increase in the log2 fold change of the x-axis at NNGRVN and NGRVNN. The same pattern can be seen for depletion samples in Figures 3.13 and 3.14. However, the depletion samples (Fig. 3.13-3.14) exhibited lower absolute ranges in log2 fold change compared to cleavage samples, indicating reduced sensitivity. Consequently, these samples provide less comprehensive information about the full spectrum of PAM recognition and instead primarily highlight the most efficiently recognized PAM sequences, which appear to be NNNGAA.

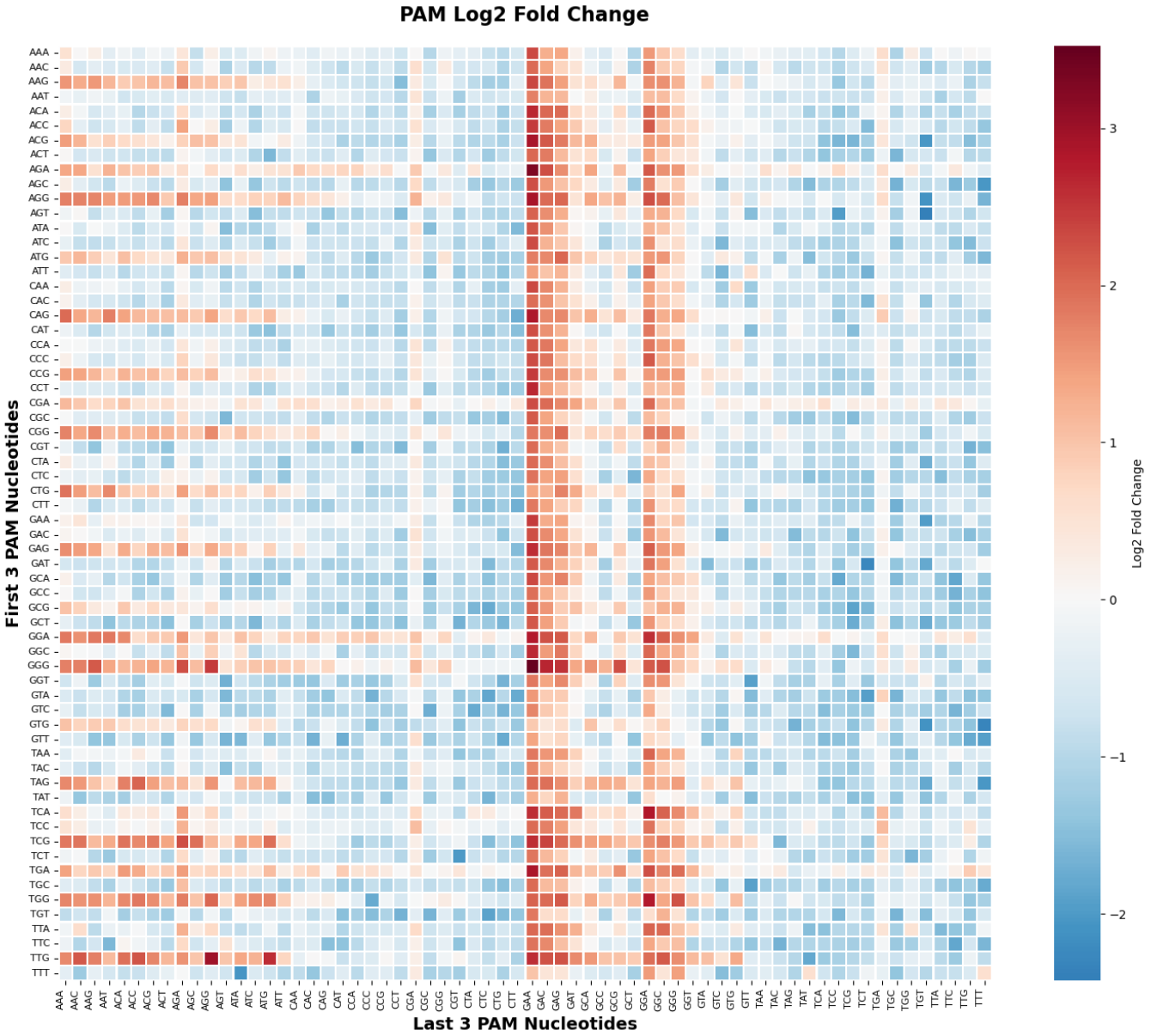


Figure 3.11: The PAM Log_2 fold change from the cleavage assay for a new AZ proprietary Cas9 enzyme expressed in cell lysate and the EMX1b substrate library. The first, second, and third PAM nucleotides are plotted on the y-axis, and the fourth, fifth, and sixth PAM nucleotides are plotted on the x-axis.

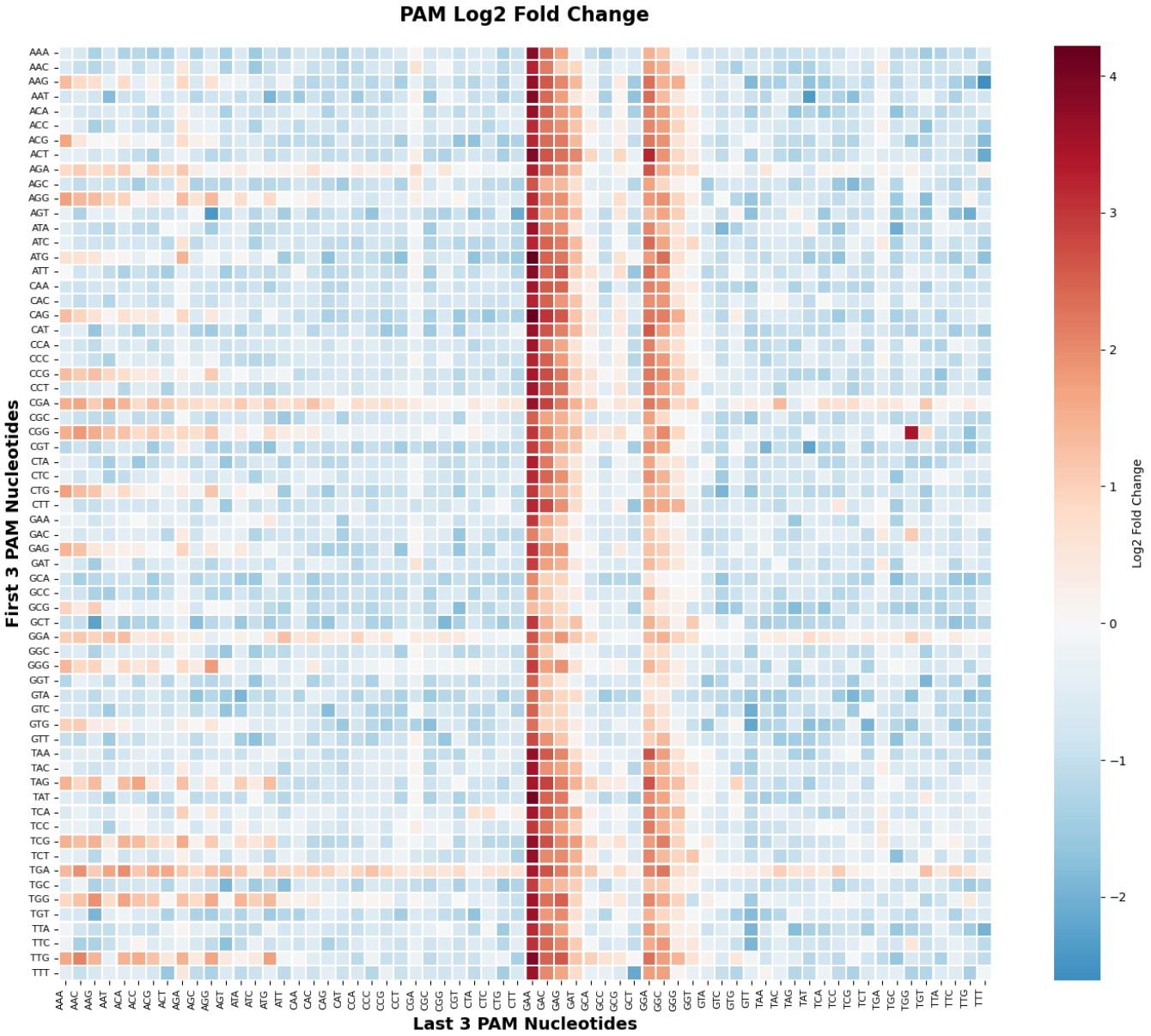


Figure 3.12: The PAM Log_2 fold change from the cleavage assay for a new AZ proprietary Cas9 enzyme expressed in cell lysate and the PCSK9 substrate library. The first, second, and third PAM nucleotides are plotted on the y-axis, and the fourth, fifth, and sixth PAM nucleotides are plotted on the x-axis.

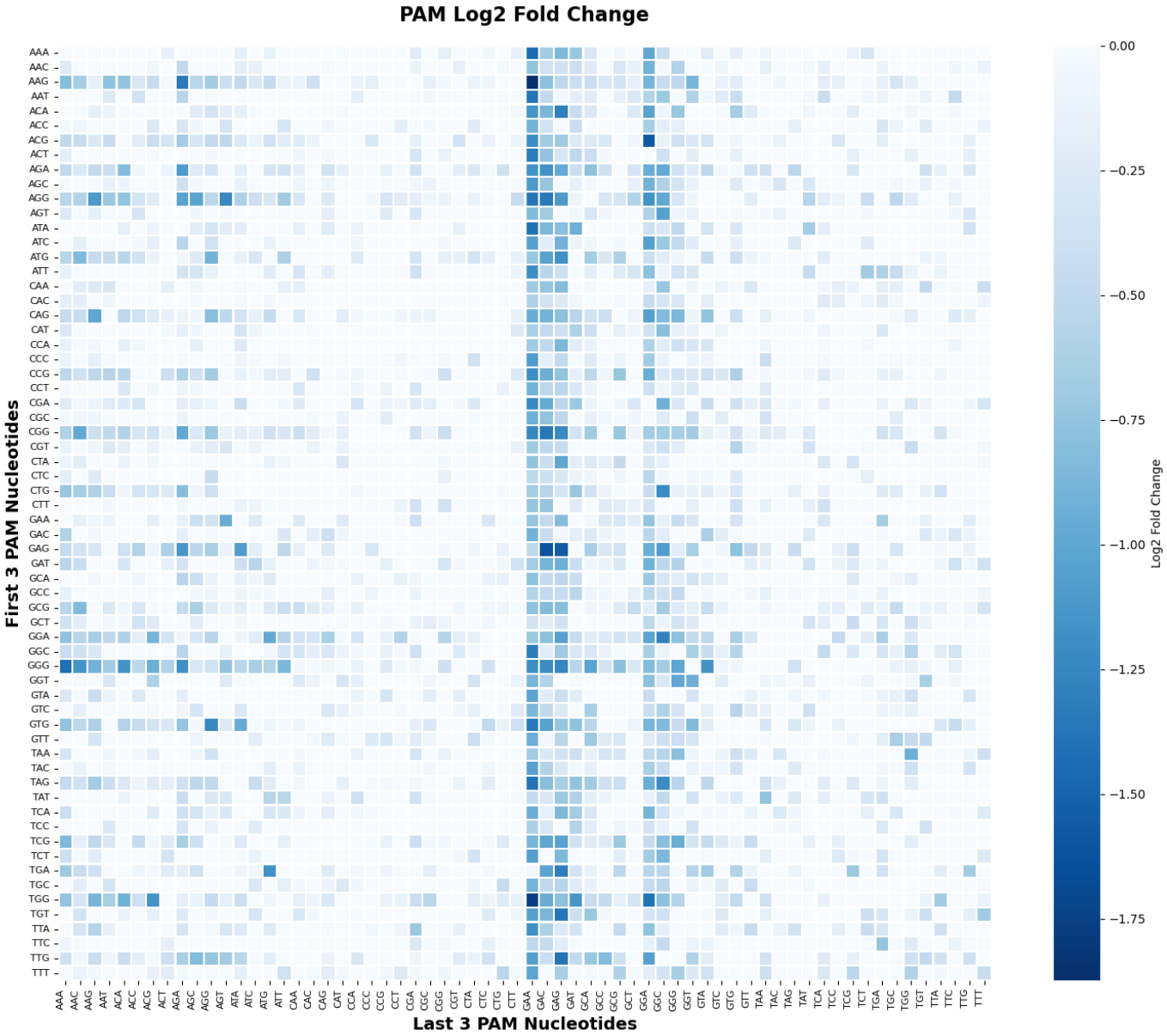


Figure 3.13: The PAM Log_2 fold change from the depletion sample for a new AZ proprietary Cas9 enzyme expressed in cell lysate and the EMX1b substrate library. The first, second, and third PAM nucleotides are plotted on the y-axis, and the fourth, fifth, and sixth PAM nucleotides are plotted on the x-axis.

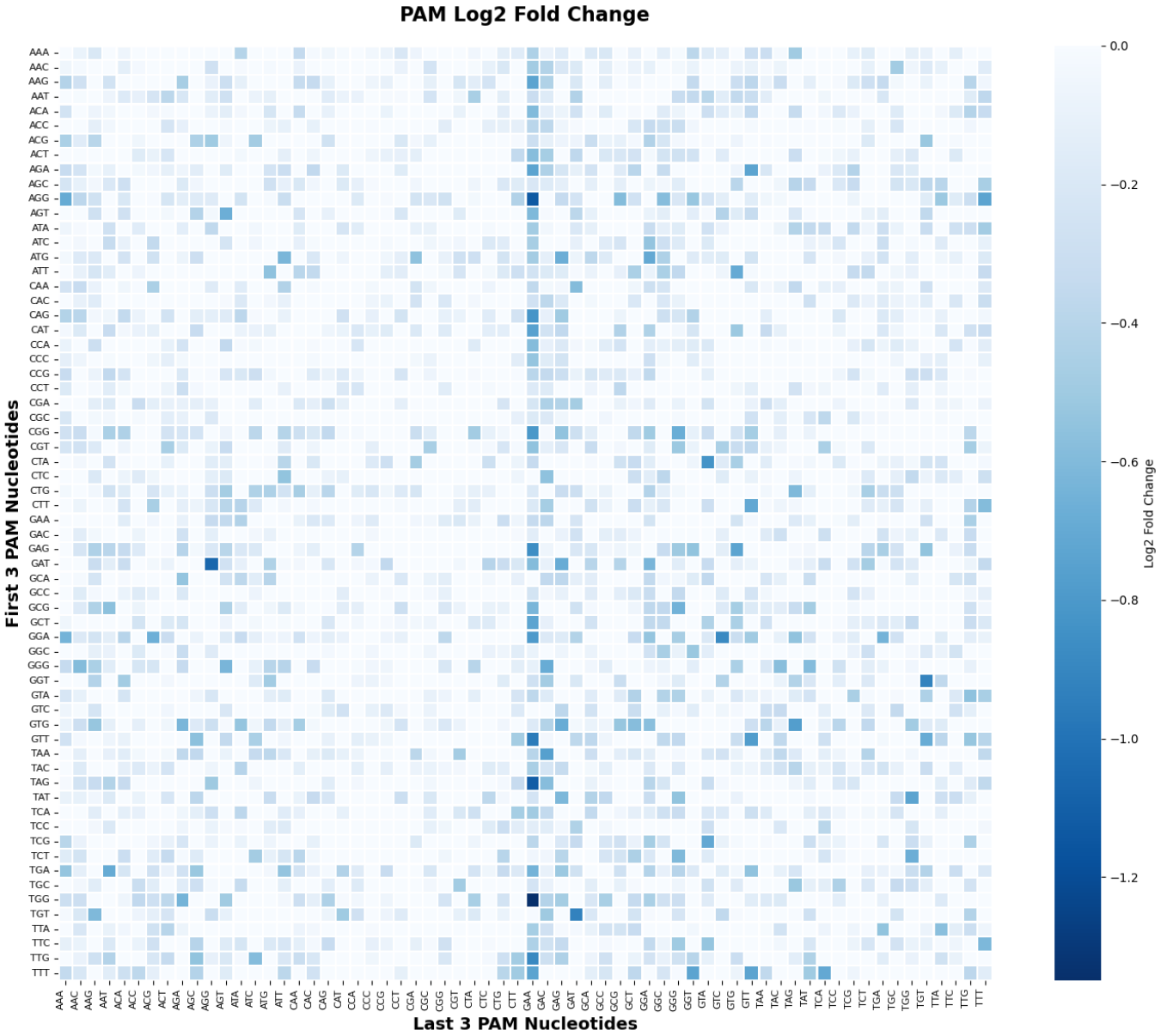


Figure 3.14: The PAM Log_2 fold change from the depletion sample for a new AZ proprietary Cas9 enzyme expressed in cell lysate and the PCSK9 substrate library. The first, second, and third PAM nucleotides are plotted on the y-axis, and the fourth, fifth, and sixth PAM nucleotides are plotted on the x-axis.

3.8 PAM Recognition of ePsCas9 Variants

After identifying a lead set of efficient ePsCas9 PAM-relaxed variants from the luciferase assay, the PAM assay was applied to determine the entire PAM recognition. Four lead ePsCas9 PAM-relaxed variants were choosen: pCN1, pCN12, pCN15, and pCN27. ePsCas9 variant pCN1 and pCN15 were indicated to mostly recognize TRG, where R is A or G, pCN12 was indicated to be able to recognize TDG, where D is A, T, or G, and pCN27 was indicated to be able to recognize TGS, where S is either G or C. The experiments used cell lysate from Cas9-expressing cells and the EMX1b-substrate library. All Cas9 variants were expressed as Cas9 protein in cell lysate using the EMX1b-substrate library.

The PAM log2 fold change for the cleavage assay for ePsCas9 variants pCN1, pCN12, pCN15, pCN27, and ePsCas9 control using the EMX1b substrate library can be seen in Figures 3.15-3.19, where (a) the y-axis is first and second PAM position and the x-axis is the third and fourth PAM position, and where (b) the y-axis is second and third PAM position and the x-axis is the fourth and fifth PAM position. The PAM log2 fold change for heatmap (b) in Figures 3.15-3.19 was calculated as previously described but the PAM frequency of the experiment and PAM frequency of the untreated substrate library were calculated by summing 6 nt PAM frequency from position 2-5 instead of position 1-4 adjacent to the spacer sequence. In addition, the ePsCas9 variant pCN1 was also tested using the PCSK9 substrate library and the PAM log2 fold change for the cleavage assay can be seen in Figure 3.20. The ePsCas9 control reaction can be seen in Figure 3.19.

As described earlier, only the reads above or equal to 137 nt and below or equal to 150 nt are used to extract the PAM sequence if the downstream 22 nt sequence is an exact match, referred to as filtered reads. In Table 3.4, the true read count (k), the total read count, the percentage of true read count compared to the total read count, Cas9 used, substrate library used, and the cleavage time (min) for each experiment. The percentage filtered reads seem to be generally higher for a cleavage reaction time of 8 min compared to 32 min. The PAM log2 fold change for the cleavage assays in Figures 3.15-3.19 were therefore plotted for a cleavage reaction time of 8 min. Furthermore, the percentage of filtered reads for pCN1 and the ePsCas9 control reaction were higher for PCSK9-experiment compared to the EMX1b-experiment. This could indicate that the cleavage reaction was less efficient during the EMX1b experiment than for the PCSK9 experiment, potentially due to a variation in cell lysate quality. More importantly, there is a clear correlation between number of filtered reads and the sensitivity of the assay. This would explain, for example, why the ePsCas9 variant pCN1 is indicated to confer recognition to NTG in Figure 3.20 while this is less noticeable in Figure 3.15. However, since the interest for the ePsCas9 variants is to determine the main PAM recognition, the filtered reads from the EMX1b experiment should be enough to indicate such recognition as long as the ePsCas9 control is good.

Although, the sensitivity is lower for the EMX1b-experiment, since the number of filtered reads and the percentage filtered reads are comparable across the experi-

Table 3.4: The number of total reads obtain from the amplicon sequencing and the number of filtered reads (the reads from which the PAMs are extracted from) for selected ePsCas9 variants tested in the PAM cleavage assay. Additionally, the substrate library and the cleavage reaction time for each experiment are indicated.

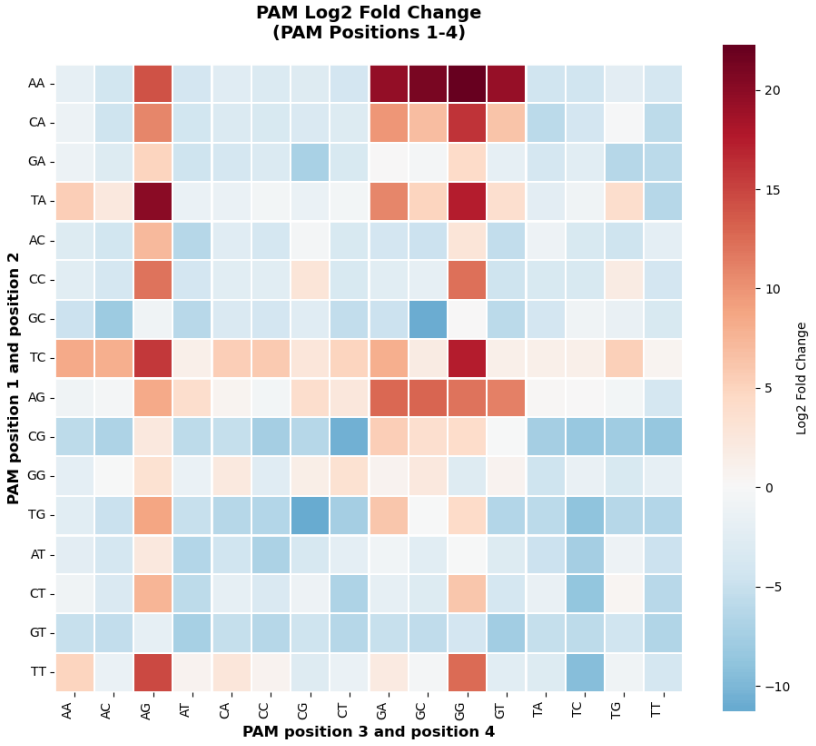
Cas9	Substrate	Cleavage time	Filtered reads (k)	Total reads (k)	% filtered reads
pCN1	EMX1b	8 min	250	1000	25
pCN1	EMX1b	32 min	199	950	21
pCN12	EMX1b	8 min	266	830	32
pCN12	EMX1b	32 min	230	930	25
pCN15	EMX1b	8 min	270	750	36
pCN15	EMX1b	32 min	290	980	29
pCN27	EMX1b	8 min	275	944	29
pCN27	EMX1b	32 min	254	1000	25
ePsCas9	EMX1b	8 min	312	959	33
ePsCas9	EMX1b	32 min	250	1100	23
pCN1	PCSK9	30 min	790	1940	40
ePsCas9	PCSK9	30 min	800	1600	50

ment, the sensitivity should be comparable between variants in the same experiment. Therefore, it is reasonable to assume that pCN12 (Fig. 3.16) has more of a NDG recognition, than pCN1 and pCN15 (Fig. 3.15 and 3.17) that seems to confer recognition to NRG.

There is no obvious pattern in Figure 3.18a and surprisingly there is no increase in the PAM log2 fold change for NGC. However, in Figure 3.18b there is an increase in the PAM log2 fold change for both NNGCNN and for NNNGCN, which could be due to the previously observed secondary consecutive-recognized-nt phenomena. Although, high activity was seen for TGC in the luciferase reporter assay, there is no knowledge if there is any recognition to TGA, as no such cell line currently exists. Additionally, in the luciferase reporter assay a weaker recognition for TAG could also be observed. If pCN27 recognizes a broader spectrum of PAMs, then a higher true read depth would be needed compared to the ePsCas9 control and the other ePsCas9 variants. Therefore, another experiment with a higher true read depth should be performed before the exact PAM-recognition of pCN27 can be determined.

The depletion samples for all ePsCas9 variants, including the control, failed for EMX1b and due to time constraints this experiment could not be repeated. However, in Figure 3.21 the depletion sample for pCN1 and wild-type ePsCas9 can be seen.

(a)



(b)

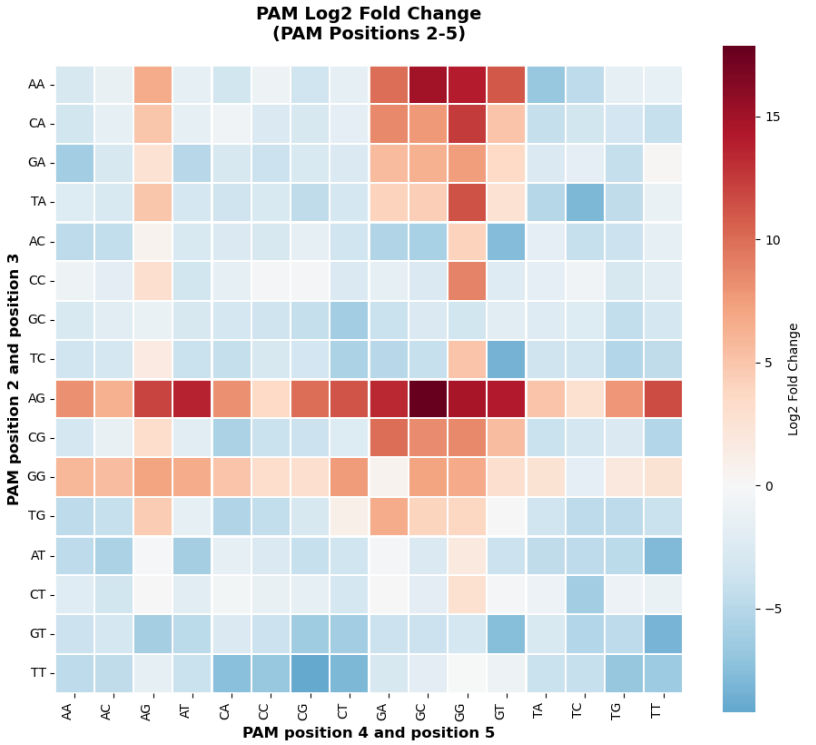
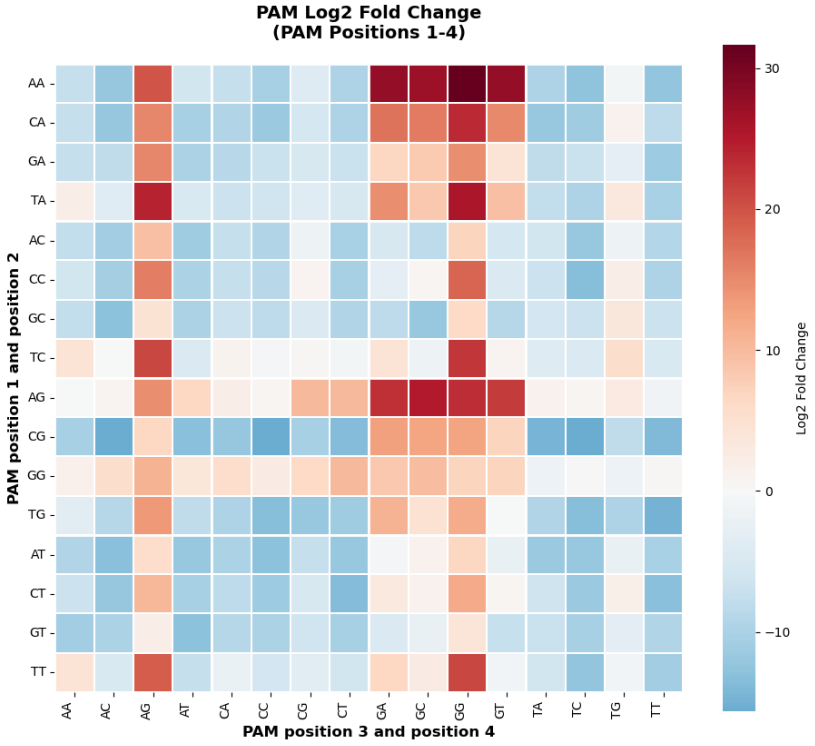


Figure 3.15: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 variant pCN1 expressed in cell lysate and the EMX1b substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

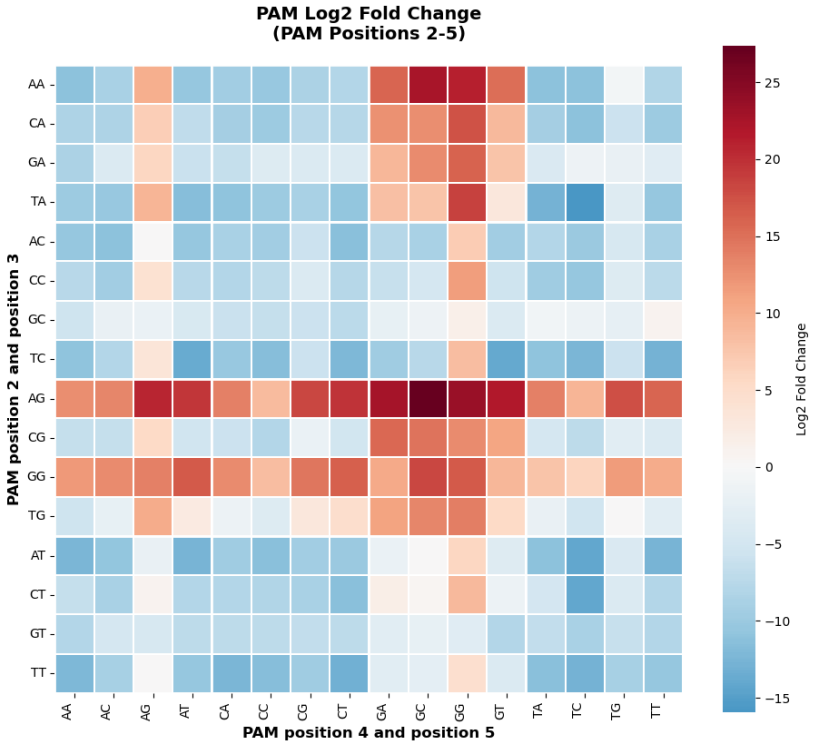
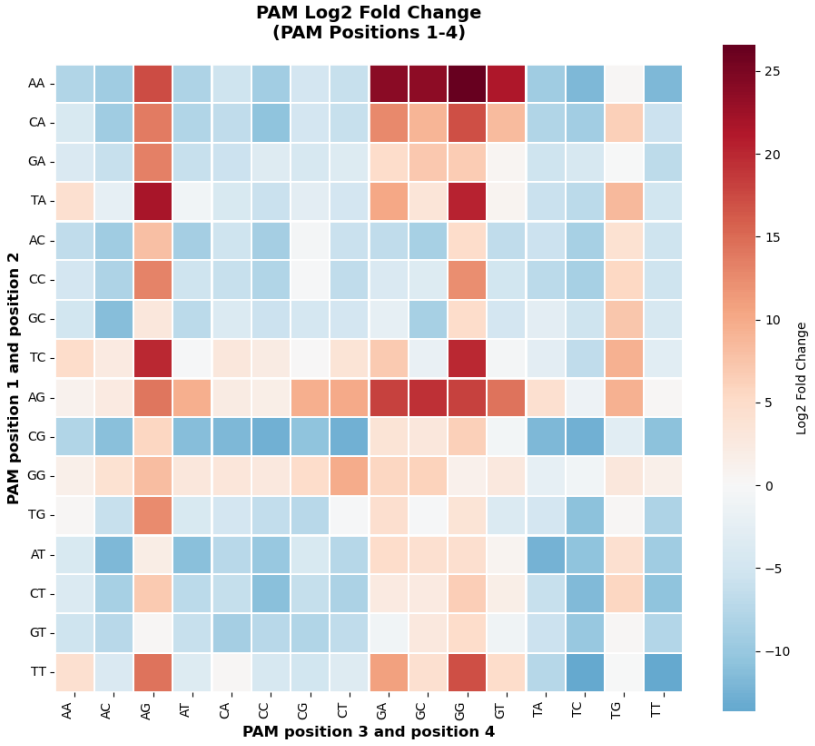


Figure 3.16: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 variant pCN15 expressed in cell lysate and the EMX1b substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

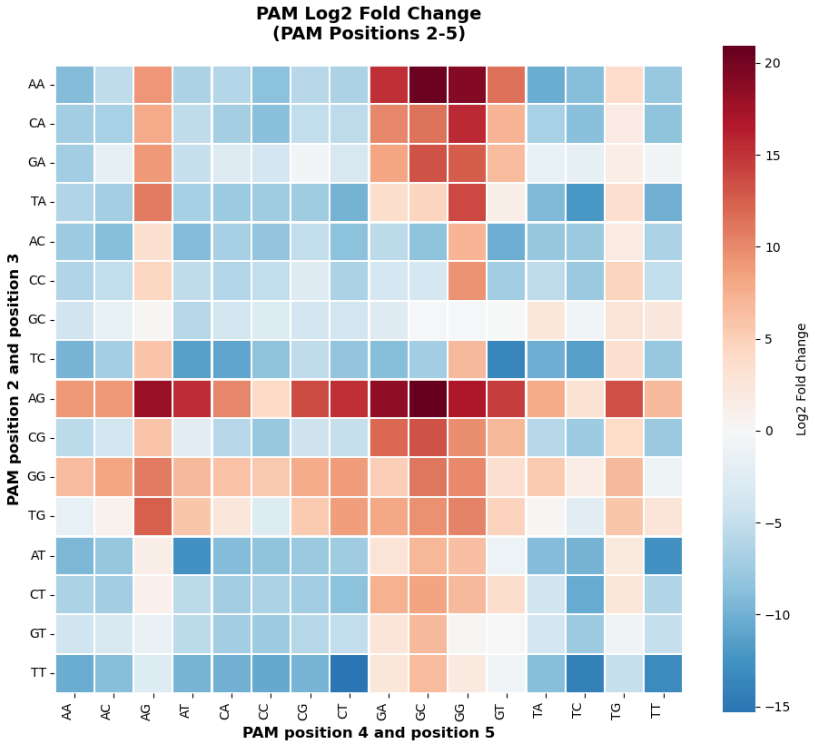
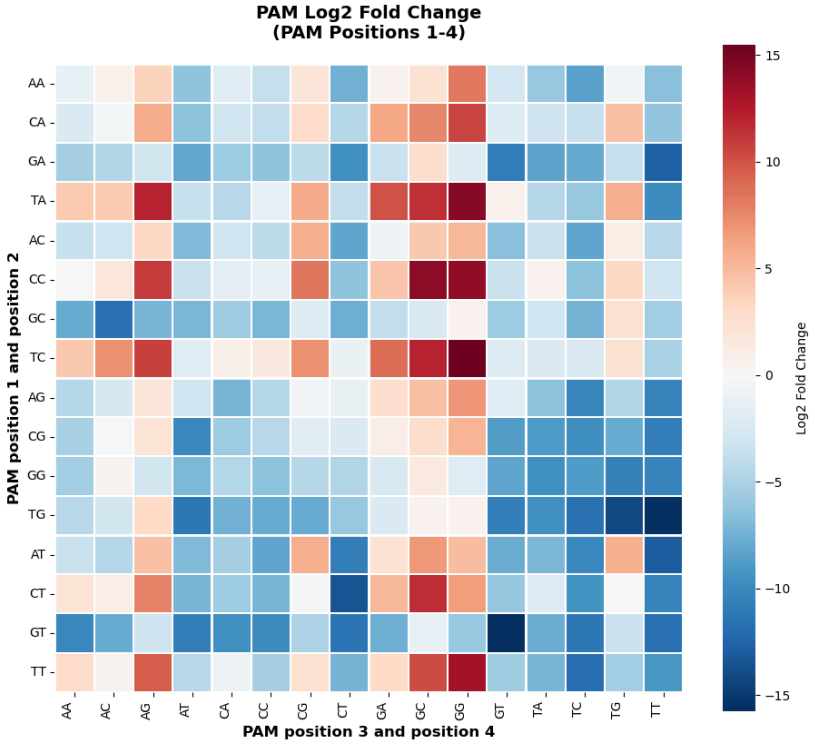


Figure 3.17: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 variant pCN12 expressed in cell lysate and the EMX1b substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

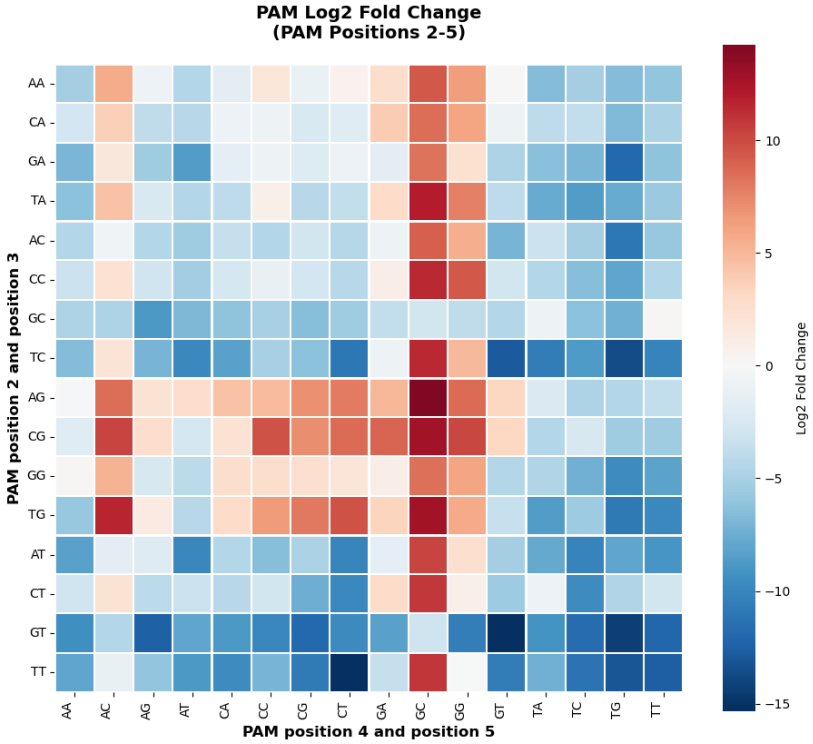
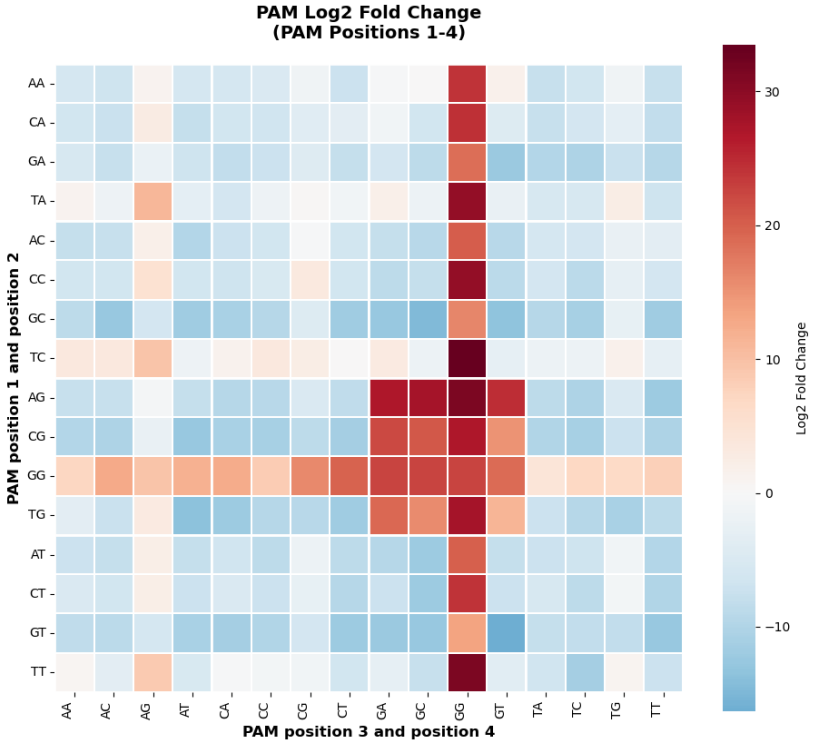


Figure 3.18: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 variant pCN27 expressed in cell lysate and the EMX1b substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

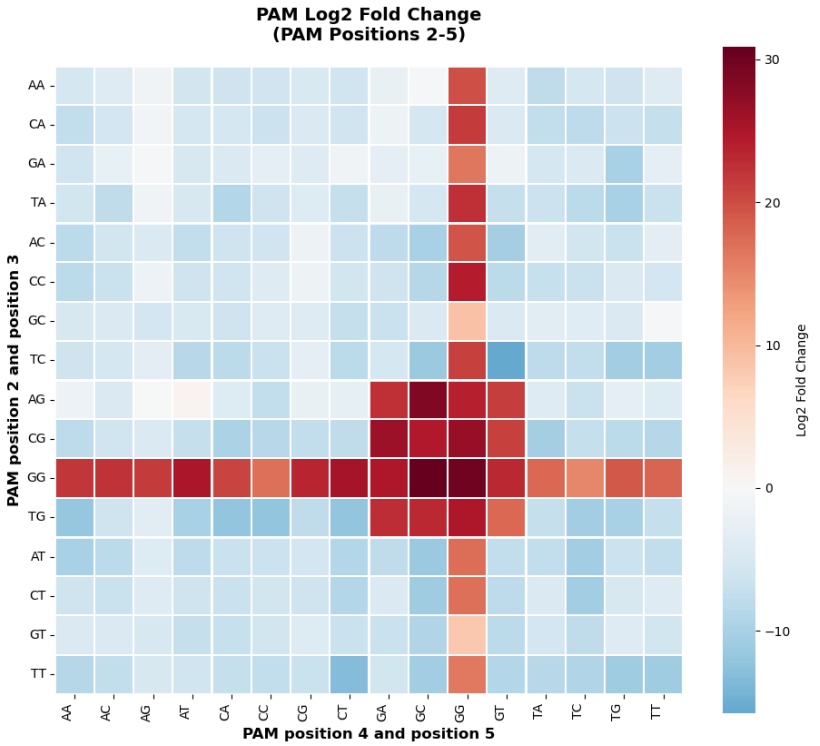
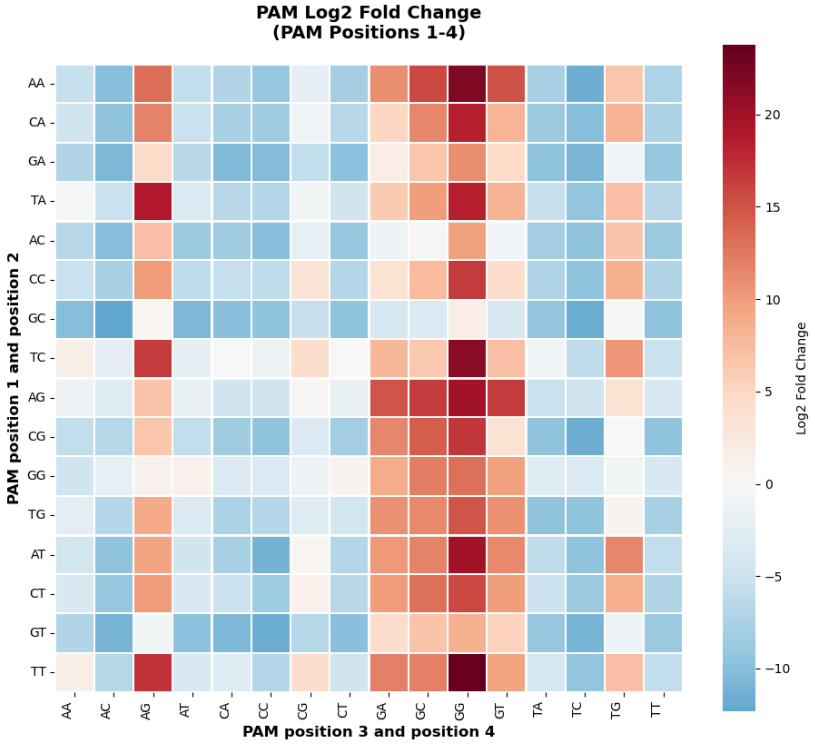


Figure 3.19: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 expressed in cell lysate and the EMX1b substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

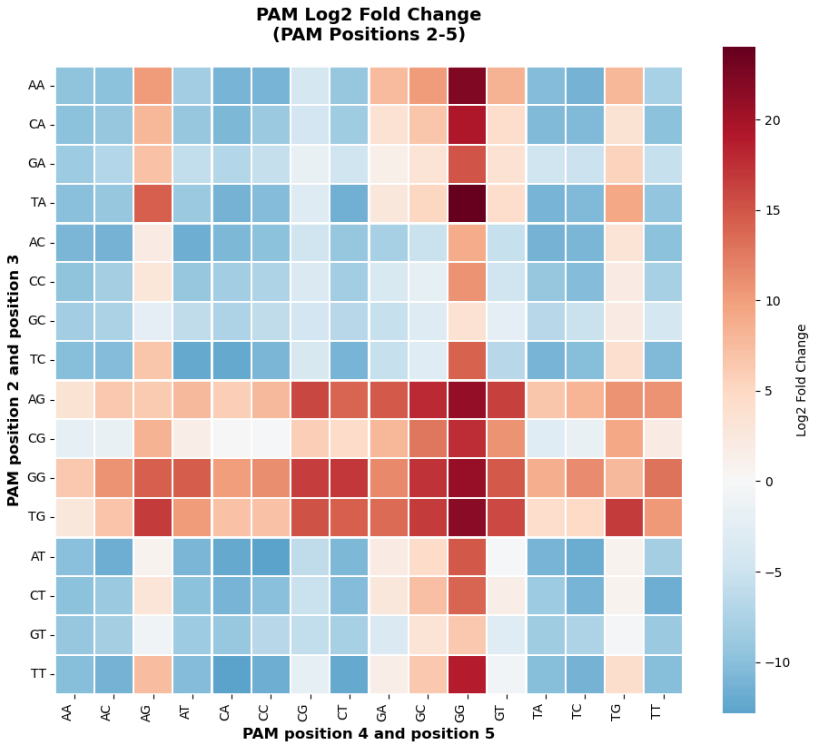
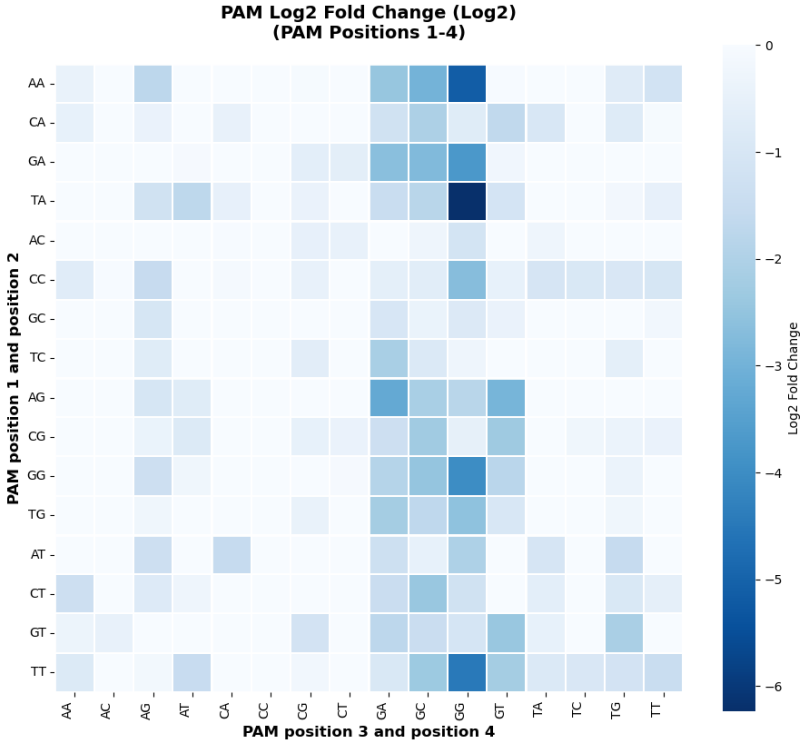


Figure 3.20: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 variant pCN1 expressed in cell lysate and the PCSK9 substrate library. (a) The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide. (b) The second and third PAM nucleotide is plotted against the forth and fifth PAM nucleotide.

(a)



(b)

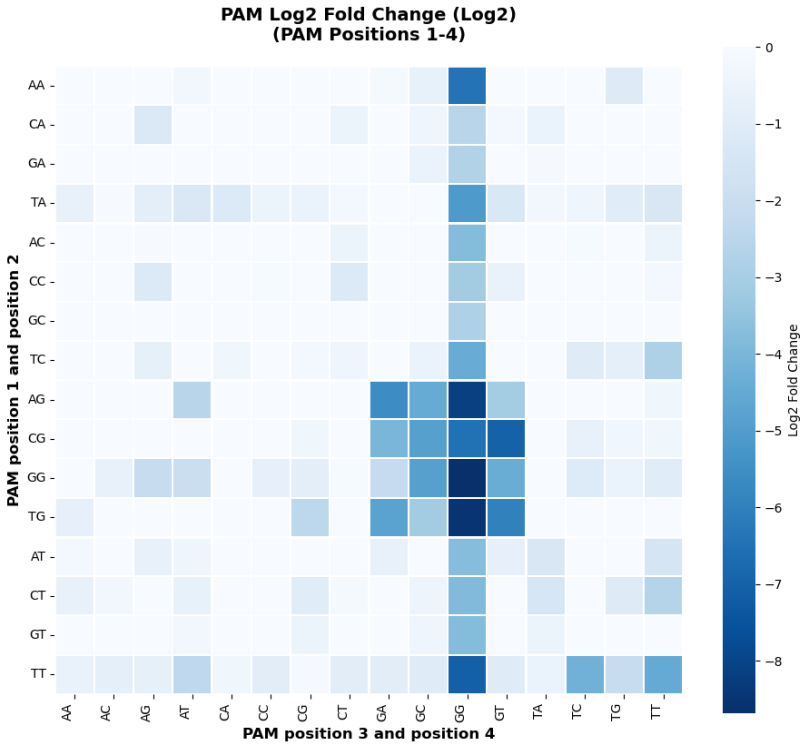


Figure 3.21: (a) The PAM Log_2 Fold Change from the depletion samples for ePsCas9 variant pCN1 expressed in cell lysate and the PCSK9 substrate library. (b) The PAM Log_2 Fold Change from the depletion samples for wild-type ePsCas9 expressed in cell lysate and the PCSK9 substrate library. The first and second PAM nucleotide is plotted against the third and forth PAM nucleotide.

3.9 Comparing ePsCas9 PAM-relaxed Variants with SpCas9 PAM-relaxed Variants

To further evaluate the efficiency of the four lead ePsCas9 variants (pCN1, pCN12, pCN15, and pCN27), twelve genomic loci encompassing four distinct PAM sequences (NAG, NGG, NTG, and NGC) were selected as target sequences. These variants were compared with three established SpCas9 PAM-relaxed variants (SpRY, SpG, and xCas9), wild-type SpCas9 and ePsCas9, with untransfected HEK293T cells serving as negative control [48, 27]. SpRY, which recognizes NRN PAMs (with reduced activity on NYN sequences), provides an appropriate benchmark for assessing ePsCas9 variant efficiency on NAG and NTG PAMs. Furthermore, SpG recognizes NGN PAMs, and xCas9 NGN, GAA, and GAT PAMs, providing a benchmark for the ePsCas9 variants, which recognizes NGC PAM sequences. This experiment was carried out with two biological replicates, each of which contained three technical replicates. All Cas9 constructs (Fig. 2.4) were delivered as plasmid DNA together with sgRNA, and amplicon sequencing data were analyzed using the AZ internal software RIMA4. The total editing efficiency (%) for each Cas9 at the different target sequences is shown in Figure 3.22.

The results of the total editing efficiency (Fig. 3.22) revealed differences both in activity between the Cas9 variants and between Cas9 orthologs. At the canonical NGG-PAM sites, SpCas9 (wild-type and variants) demostrated a higher activity than ePsCas9 (wild-type and variants). The difference in activity is likely attributed to distinct spacer preferences between the two Cas9 orthologs, as SpCas9 and ePsCas9 have evolved in different bacterial species. Previous studies have documented that Cas9 orthologs, in addition to their PAM requirement may display variation in their preference for specific nucleotide sequences within the protospacer sequence [53].

Notably, at NGG-PAM sites, both SpCas9 PAM-relaxed variants and ePsCas9 PAM-relaxed variants generally exhibited lower activity compared to their respective wild-type enzymes. For SpCas9, this reduction in editing activity has previously been documented for SpRY and SpG, and the similar observation for ePsCas9 indicates that this phenomenon may extend to ePsCas9 [48]. The structural basis for this activity reduction at NGG-sites likely involves less optimal interaction with the PAM sequence and mutated amino acids of the Cas9 variants.

However, at non-canonical PAM sites (NAG, NTG, and NGC), a substantial increase in editing activity was observed for the Cas9 variants compared to their wild-type counterparts. At NAG sites, ePsCas9 variants pCN12 and pCN15 achieve mean total editing efficiencies of 42% and 44%, respectively, performing equally well or better than SpRY at two of three loci. Notably, the second-generation variants pCN12 and pCN15 also exhibit higher editing efficiency than the first-generation variant pCN1 at NAG-sites (Fig. 3.22). Notably, this differs from the luciferase reporter assay results (Fig. 3.6), where pCN1, pCN12, and pCN15 demonstrated comparative activity on TAG PAM. This contradiction could result from limited statistical power

during the screening. In addition, differences in chromatin structures at different genomic loci can also influence editing efficiency [54, 55].

At the NGC-PAM sites, SpG consistently reached the highest total editing efficiency at all three target sequences with a mean total editing efficiency of approximately 15-40%, followed by SpRY (mean: 5-20%), and pCN27 (mean: 0-10%), while no other variants exhibit detectable activity. This indicates that although the engineering of pCN27 improved NGC-PAM recognition compared to wild-type ePsCas9, its activity at NGC sites could be further enhanced. Interestingly, xCas9 did not exhibit detectable editing efficiencies despite its increased recognition for NGN-PAMs. However, xCas9 has been reported to confer poor tolerance to cytosine at the third PAM position, which explained the low activity at the NGC-PAM sites[27].

The editing efficiencies were lowest at the *FANCF* and *MECP2* NTG-PAM sites compared to all other target sites. At *FANCF*, SpRY demonstrated the highest activity with 5% mean total editing efficiency, followed by pCN12 at 2%. The enhanced PAM-relaxation at the second PAM nucleotide position for pCN12 was indicated in the PAM assay, which showed higher log2 fold change values for NTG sequences compared to the other ePsCas9 lead variants. These findings are further supported by the results in Figure 3.22. However, the *EMX1* NTG-PAM site demonstrated an editing efficiency of approximately 30 % for SpRY, with detectable activities observed across all SpCas9 variants but not for any ePsCas9 variants. The increased activity at the *EMX1* NTG-PAM site compared to the other NTG-PAM sites is likely attributed to spacer preferences.

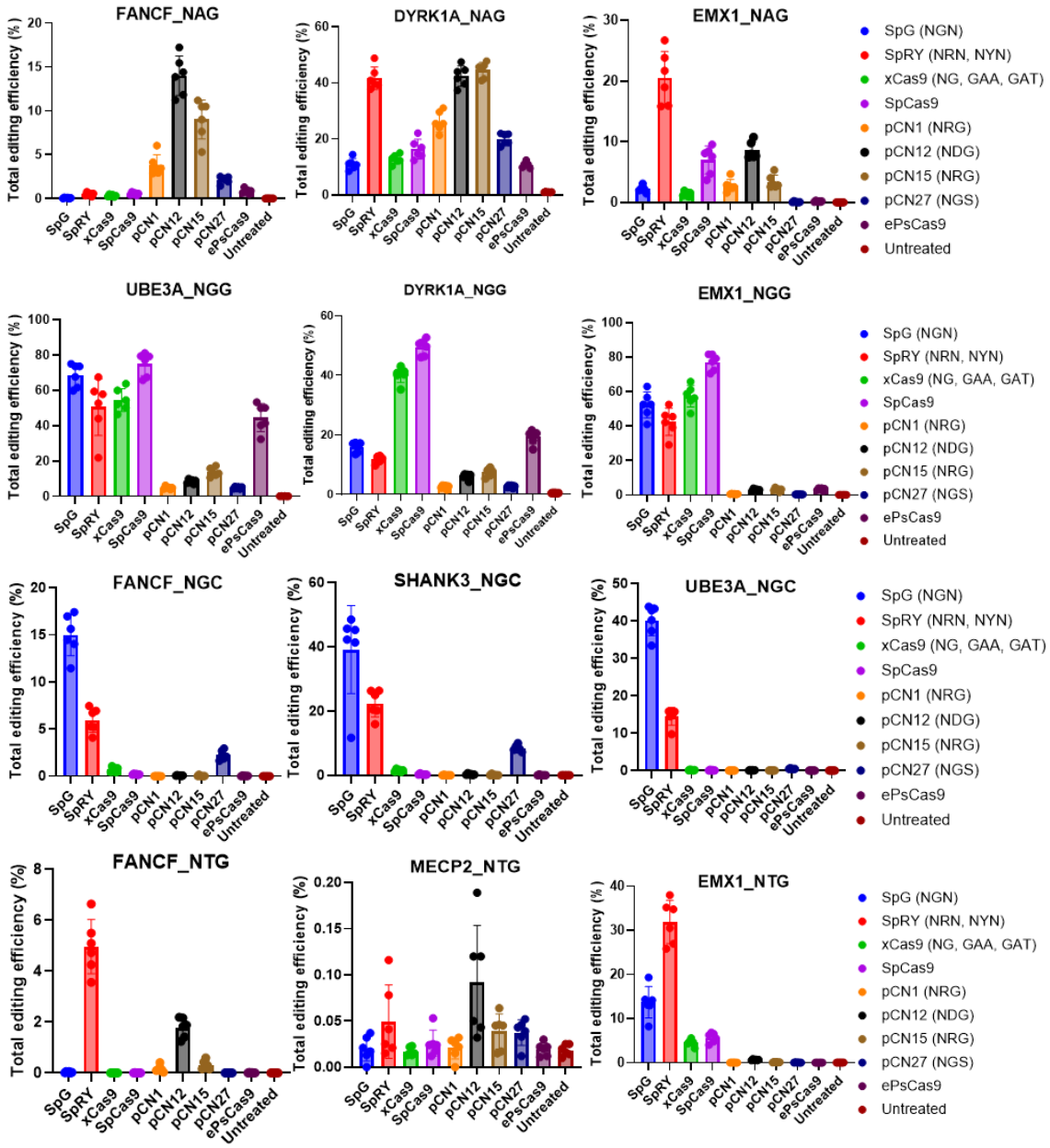


Figure 3.22: Total editing efficiency (%) of SpCas9 variants and ePsCas9 variants on twelve genomic loci. For each genomic loci, the gene and PAM sequence at that target sequence is shown. The PAM sequence that each Cas9 variant recognizes is shown in the parenthesis (right).

4
Conclusion

In this master's thesis, a directed evolution system was employed to generate ePsCas9 variants with altered PAM recognition capabilities. Multiple approaches for library generation and selection strategies were systematically explored. Selected variants were subsequently screened using a luciferase reporter system, which identified four lead candidates. The PAM preferences of these lead ePsCas9 variants were characterized through a novel PAM determination assay. Finally, comparative analysis against established SpCas9 PAM-relaxed variants revealed that pCN12 and pCN15 exhibit equal or higher activity at two of three target sequences containing NAG-PAM sequences. Additionally, pCN27 (selected in TGC) demonstrated enhanced activity on NGC-PAM sequences compared to wild-type ePsCas9.

The novel PAM assay developed in this work provides an additional tool for evaluating the PAM preferences of Cas9 enzymes. This assay was successfully employed to determine the relaxed PAM preferences (NRG or NDG) for pCN1, pCN12, and pCN15. Furthermore, the assay was applied to characterize a novel proprietary AstraZeneca Cas9 enzyme suspected of recognizing a 6-nt long PAM sequence, which was determined to be NNNGAA.

However, both the PAM assay results and comparative analysis with SpCas9 variants demonstrate that the evolved ePsCas9 variants retain a more stringent PAM recognition profile than SpCas9 PAM-relaxed variants, thus limiting their targeting range to fewer genomic sequences. Notably, increased PAM flexibility in Cas9 variants has been associated with elevated off-target effects in genome editing applications [48, 56]. Consequently, a moderately PAM-relaxed Cas9 variant may exhibit reduced off-target activity. Given that ePsCas9 is already recognized as a high-fidelity Cas9 enzyme due to its enhanced sequence specificity, these novel variants could represent a safer alternative for genome editing applications by balancing the expanded targeting capabilities with maintained precision.

The novel ePsCas9 variants expand the CRISPR-Cas9 genome editing toolbox by enabling increased targeting capability at sequences harboring NAG or NGC PAM sequences. Previous studies have demonstrated that SpRY and SpG can be successfully fused to deaminases to function as efficient base editors [48]. Given their comparable activity at NAG-PAM sites, pCN12 and pCN15 show strong potential as base editor platforms. Prime editing applications should also be explored for pCN12 and pCN15, wherein ePsCas9 would be fused to a reverse transcriptase with the Cas9 enzyme mutated into a nickase variant.

Future research should therefore aim to characterize the off-target effects of these ePsCas9 variants. Additionally, increased activity was observed for variants (e.g., pCN12 and pCN15) in which the non-PAM-interacting coding sequence was mutagenized. This approach could be applied to enhance the activity of pCN27, which exhibits NGC recognition but currently demonstrates lower activity than SpG. Furthermore, this mutagenesis strategy could potentially improve the activity of pCN12 at NTG-PAM sequences.

References

1. Pacesa M, Pelea O, and Jinek M. Past, present, and future of CRISPR genome editing technologies. Cell 2024; 187:1076–100. DOI: 10.1016/j.cell.2024.01.042
2. Carroll D. Genome Engineering With Zinc-Finger Nucleases. Genetics 2011; 188:773–82. DOI: 10.1534/genetics.111.131433
3. Sun N and Zhao H. Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. Biotechnology and Bioengineering 2013; 110:1811–21. DOI: 10.1002/bit.24890
4. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, and Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 2012 Aug; 337:816–21. DOI: 10.1126/science.1225829
5. Doudna JA and Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science 2014; 346:1258096. DOI: 10.1126/science.1258096
6. Wang JY and Doudna JA. CRISPR technology: A decade of genome editing is only the beginning. Science 2023; 379:eadd8643. DOI: 10.1126/science.add8643
7. Food U and Administration D. FDA Approves First Gene Therapies to Treat Patients with Sickle Cell Disease. Press Release. U.S Food and Drug Administration, 2023 Dec. Available from: https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapies-treat-patients-sickle-cell-disease
8. Komor AC, Kim YB, Packer MS, Zuris JA, and Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 2016; 533:420–4. DOI: 10.1038/nature17946
9. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, and Liu DR. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019; 576:149–57. DOI: 10.1038/s41586-019-1711-4
10. Bhaya D, Davison M, and Barrangou R. CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. Annual Review of Genetics 2011; 45:273–97. DOI: 10.1146/annurev-genet-110410-132430
11. Gottesman S. Dicing defence in bacteria. Annual Review of Genetics 2011; 471:588–9. DOI: 10.1038/471588a
12. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D, Shah SA, Siksnys V, Terns MP, Venclovas Č, White MF, Yakunin AF,

Yan W, Zhang F, Garrett RA, Backofen R, Oost J van der, Barrangou R, and Koonin EV. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nature Reviews Microbiology 2019 Dec; 18:67–83. DOI: 10.1038/s41579-019-0299-x

13. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, and Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 2011; 471:602–7. DOI: 10.1038/nature09886
14. Jiang W, Bikard D, Cox D, Zhang F, and Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnology 2013; 31:233–9. DOI: 10.1038/nbt.2508
15. Sciences TRSA of. Nobel Prize in Chemistry 2020. Press release. The Royal Swedish Academy of Sciences, 2020 Oct. Available from: https://www.nobelprize.org/prizes/chemistry/2020/press-release/
16. Wu CC, Lin JL, and Yuan HS. Structures, Mechanisms, and Functions of His-Me Finger Nucleases. Trends in Biochemical Sciences 2020; 45:935–46. DOI: 10.1016/j.tibs.2020.07.002
17. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, and Nureki O. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell 2014; 156:935–49. DOI: 10.1016/j.cell.2014.02.001
18. Jiang F, Zhou K, Ma L, Gressel S, and Doudna JA. A Cas9–guide RNA complex preorganized for target DNA recognition. Science 2015; 348:1477–81. DOI: 10.1126/science.aab1452
19. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, and Doudna JA. Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. Science 2014; 343:1247997. DOI: 10.1126/science.1247997
20. Lieber MR. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annual Review of Biochemistry 2010; 79:181–211. DOI: 10.1146/annurev.biochem.052308.093131
21. Wright WD, Shah SS, and Heyer WD. Homologous recombination and the repair of DNA double-strand breaks. Journal of Biological Chemistry 2018; 293:10524–35. DOI: 10.1074/jbc.tm118.000372
22. Sfeir A and Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends in Biochemical Sciences 2015; 40:701–14. DOI: 10.1016/j.tibs.2015.08.006
23. Bestas B, Wimberger S, Degtev D, Madsen A, Rottner AK, Karlsson F, Naumenko S, Callahan M, Touza JL, Francescatto M, Möller CI, Badertscher L, Li S, Cerboni S, Selfjord N, Ericson E, Gordon E, Firth M, Chylinski K, Taheri-Ghahfarokhi A, Bohlooly-Y M, Snowden M, Pangalos M, Nuttall B, Akcakaya P, Sienski G, and Maresca M. A Type II-B Cas9 nuclease with minimized off-targets and reduced chromosomal translocations in vivo. Nature Communications 2023; 14. DOI: 10.1038/s41467-023-41240-7

24. Hibshman GN and Taylor DW. Structural basis of a dual-function type II-B CRISPR–Cas9. Nucleic Acids Research 2025 Jun; 53. DOI: 10.1093/nar/gkaf585
25. Degtev D, Bravo J, Emmanouilidi A, Zdravković A, Choong OK, Liz Touza J, Selfjord N, Weisheit I, Francescatto M, Akcakaya P, Porritt M, Maresca M, Taylor D, and Sienski G. Engineered PsCas9 enables therapeutic genome editing in mouse liver with lipid nanoparticles. Nature Communications 2024 Nov; 15. DOI: 10.1038/s41467-024-53418-8
26. Acharya S, Ansari AH, Kumar Das P, Hirano S, Aich M, Rauthan R, Mahato S, Maddileti S, Sarkar S, Kumar M, Phutela R, Gulati S, Rahman A, Goel A, Afzal C, Paul D, Agrawal T, Pulimamidi VK, Jalali S, Nishimasu H, Mariappan I, Nureki O, Maiti S, and Chakraborty D. PAM-flexible Engineered FnCas9 variants for robust and ultra-precise genome editing and diagnostics. Nature Communications 2024 Jun; 15:5471. DOI: 10.1038/s41467-024-49233-w
27. Johnny H. H, Shannon M. M, Maarten H. G, Weixin T, Liwei C, Ning S, Christina M. Z, Xue G, Holly A. R, Zhi L, and Liu DR. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature 2018
28. S. Vidal L, Isalan M, Heap JT, and Ledesma-Amaro R. A primer to directed evolution: current methodologies and future directions. RSC Chemical Biology 2023 Jan; 4:271–91. DOI: 10.1039/d2cb00231k
29. Packer MS and Liu DR. Methods for the directed evolution of proteins. Nature Reviews Genetics 2015; 16:379–94. DOI: 10.1038/nrg3927
30. Wang Y, Xue P, Cao M, Yu T, Lane ST, and Zhao H. Directed Evolution: Methodologies and Applications. Chemical Reviews 2021; 121:12384–444. DOI: 10.1021/acs.chemrev.1c00260
31. McLure RJ, Radford SE, and Brockwell DJ. High-throughput directed evolution: a golden era for protein science. Trends in Chemistry 2022 May; 4:378–91. DOI: 10.1016/j.trechm.2022.02.004
32. Zeymer C and Hilvert D. Directed Evolution of Protein Catalysts. Annual Review of Biochemistry 2018; 87:131–57. DOI: 10.1146/annurev-biochem-062917-012034
33. Fujii R. RAISE: a simple and novel method of generating random insertion and deletion mutations. Nucleic Acids Research 2006; 34:e30–e30. DOI: 10.1093/nar/gnj032
34. Popa SC, Inamoto I, Thuronyi BW, and Shin JA. Phage-Assisted Continuous Evolution (PACE): A Guide Focused on Evolving Protein–DNA Interactions. ACS Omega 2020; 5:26957–66. DOI: 10.1021/acsomega.0c03508
35. Kipnis Y, Dellus-Gur E, and Tawfik DS. TRINS: a method for gene modification by randomized tandem repeat insertions. Protein Engineering Design and Selection 2012; 25:437–44. DOI: 10.1093/protein/gzs023
36. Chen Dawn X, Chen Z, Wythes G, Zhang Y, Orr BC, Sun G, Chao YK, Navarro T. A, Thao K, Vallurupalli M, Sun J, Borji M, Tkacik E, Chen H, Bernstein BE, and Chen F. Helicase-assisted continuous editing for programmable mutagenesis of endogenous genomes. Science 2024; 386:eadn5876. DOI: 10.1126/science.adn5876

37. Chen H, Liu S, Padula S, Lesman D, Griswold K, Lin A, Zhao T, Marshall JL, and Chen F. Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. Nature Biotechnology 2019; 38:165–8. DOI: 10.1038/s41587-019-0331-8
38. Cravens A, Jamil OK, Kong D, Sockolosky JT, and Smolke CD. Polymerase-guided base editing enables in vivo mutagenesis and rapid protein engineering. Nature Communications 2021; 12. DOI: 10.1038/s41467-021-21876-z
39. Mengiste AA, McDonald JL, Nguyen Tran MT, Plank AV, Wilson RH, Butty VL, and Shoulders MD. MutaT7GDE: A Single Chimera for the Targeted, Balanced, Efficient, and Processive Installation of All Possible Transition Mutations In Vivo. ACS Synthetic Biology 2024; 13:2693–701. DOI: 10.1021/acssynbio.4c00316
40. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales APW, Li Z, Peterson RT, Yeh JRJ, Aryee MJ, and Joung JK. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 2015 Jun; 523:481–5. DOI: 10.1038/nature14592
41. Badran AH and Lui D. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. Nature Communications 2015 Jul; 6:8425. DOI: 10.1038/ncomms9425
42. Agilent. GeneMorph II Random Mutagenesis Kit. https://www.agilent.com/en/product/mutagenesis-cloning/mutagenesis-kits/random-mutagenesis-kits/genemorph-ii-233115, [Accessed 29-06-2025]
43. Daffern N, Francino-Urdaniz IM, Baumer ZT, and Whitehead TA. Standardizing cassette-based deep mutagenesis by Golden Gate assembly. Biotechnology and Bioengineering 2023; 121:281–90. DOI: 10.1002/bit.28564
44. Shendure J and Ji H. Next-generation DNA sequencing. Nature Biotechnology 2008 Oct; 26:1135–45. DOI: 10.1038/nbt1486
45. Li S, Akrap N, Cerboni S, Porritt MJ, Wimberger S, Lundin A, Möller C, Firth M, Gordon E, Lazovic B, Sieńska A, Pane LS, Coelho MA, Ciotta G, Pellegrini G, Sini M, Xu X, Mitra S, Bohlooly-Y M, Taylor BJM, Sienski G, and Maresca M. Universal toxin-based selection for precise genome engineering in human cells. Nature Communications 2021 Jan; 12. DOI: 10.1038/s41467-020-20810-z
46. Taheri-Ghahfarokhi A, Taylor BJM, Nitsch R, Lundin A, Cavallo AL, Madeyski-Bengtson K, Karlsson F, Clausen M, Hicks R, Mayr LM, Bohlooly-Y M, and Maresca M. Decoding non-random mutational signatures at Cas9 targeted sites. Nucleic Acids Research 2018 Jul; 46:8417–34. DOI: 10.1093/nar/gky653
47. Walton RT, Hsu JY, Joung JK, and Kleinstiver BP. Scalable characterization of the PAM requirements of CRISPR–Cas enzymes using HT-PAMDA. Nature Protocols 2021 Feb; 16:1511–47. DOI: 10.1038/s41596-020-00465-2
48. Walton R, Christie K, Whittaker M, and Kleinstiver B. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science 2020; 368:290–6. DOI: 10.1126/science.aba8853.
49. Biolabs NE. NEB 10-beta Electrocompetent E.coli. https://www.neb.com/en/products/c3020-neb-10-beta-electrocompetent-e-coli, [Accessed 13-03-2026]

50. Hall MP, Unch J, Binkowski BF, Valley MP, Butler BL, Wood MG, Otto P, Zimmerman K, Vidugiris G, Machleidt T, Robers MB, Benink HA, Eggers CT, Slater MR, Meisenheimer PL, Klaubert DH, Fan F, Encell LP, and Wood KV. Engineered Luciferase Reporter from a Deep Sea Shrimp Utilizing a Novel Imidazopyrazinone Substrate. ACS Chemical Biology 2012; 7:1848–57. DOI: 10.1021/cb3002478
51. Lu X, Wang C, Huang JW, Min J, Luo J, Sun Y, Chen CC, and Guo RT. Structural and Mechanistic Investigations of a High-Fidelity Cas9 from Parasutterella secunda. American Chemical Society Catalysis 2025; 15. DOI: 10.1021/acscatal.5c02993
52. Smith T and Waterman M. Identification of common molecular subsequences. Journal of Molecular Biology 1981; 147:195–7. DOI: 10.1016/0022-2836(81)90087-5
53. Esvelt KM, Mali P, Braff JL, Moosburner M, Yaung SJ, and Church GM. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nature Methods 2013 Sep; 10:1116–21. DOI: 10.1038/nmeth.2681
54. Jung WJ, Park SJ, Cha S, and Kim K. Factors affecting the cleavage efficiency of the CRISPR-Cas9 system. Animal Cells and Systems 2024; 28:75–83. DOI: 10.1080/19768354.2024.2322054
55. Liu X, Homma A, Sayadi J, Yang S, Ohashi J, and Takumi T. Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. Scientific Reports 2016 Jan; 6. DOI: 10.1038/srep19675
56. Wu Y, Ren Q, Zhong Z, Liu G, Han Y, Bao Y, Liu L, Xiang S, Liu S, Tang X, Zhou J, Zheng X, Sretenovic S, Zhang T, Qi Y, and Zhang Y. Genome-wide analyses of PAM-relaxed Cas9 genome editors reveal substantial off-target effects by ABE8e in rice. Plant Biotechnology Journal 2022 May; 20:1670–82. DOI: 10.1111/pbi.13838

A

Supplementary Figures

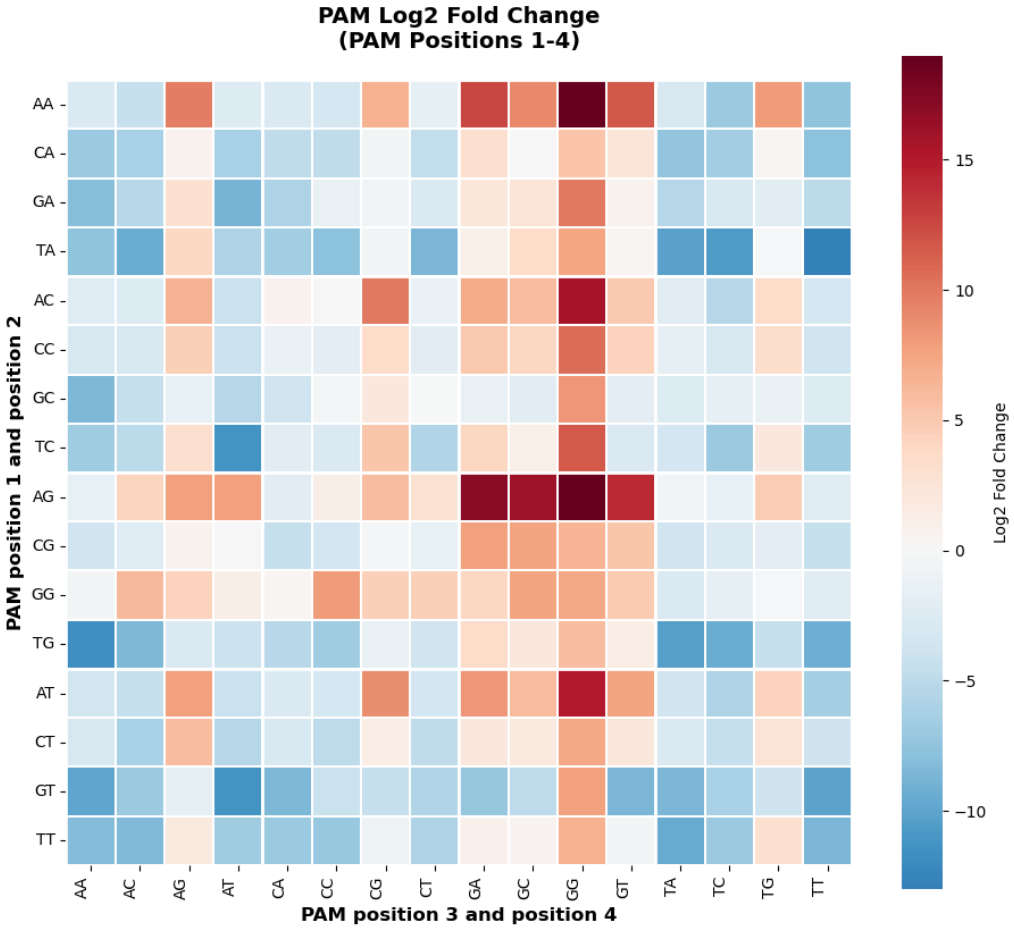
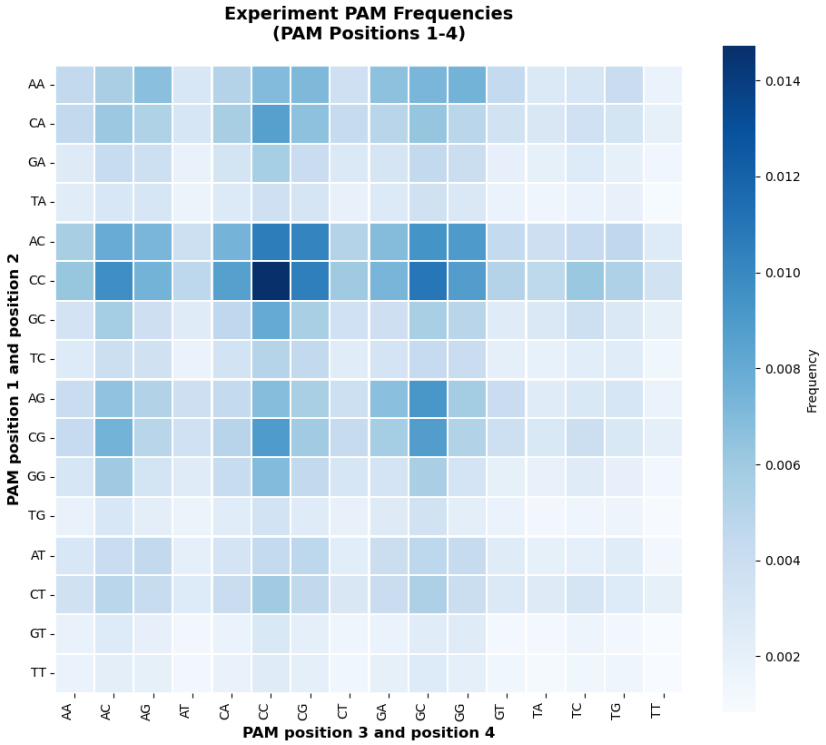


Figure A.1: PAM Log2 Fold Change from cleavage assay for ePsCas9 using purified protein and a substrate library containing the EMX1a protospacer sequence followed by a 6 nt degenerative sequence. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

(a)



(b)

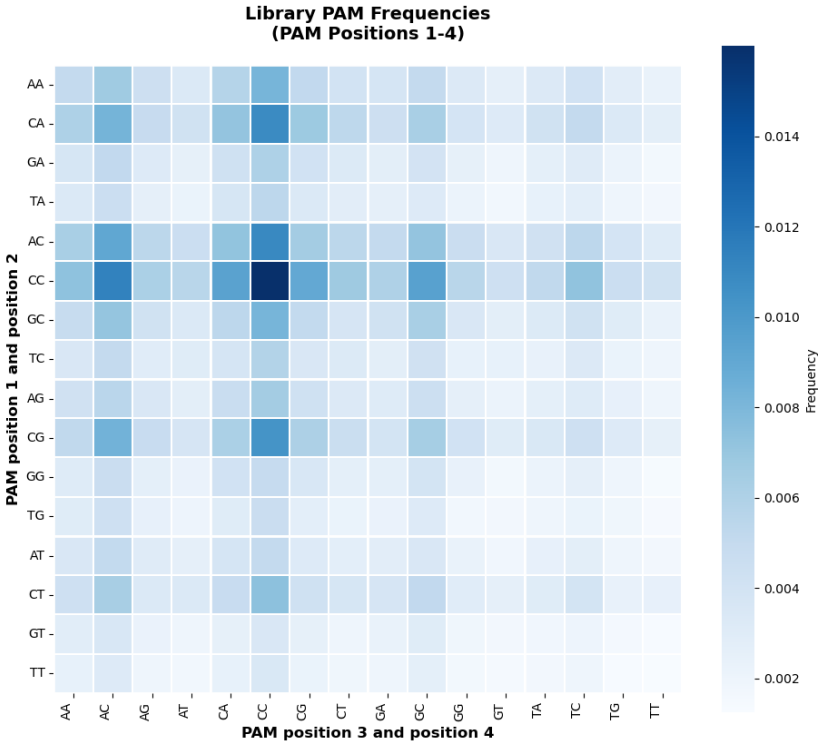


Figure A.2: The frequency of the first 4 nt downstream of the EMX1a protospacer sequence in (a) the cleavage experiment using purified ePsCas9 enzyme and (b) the untreated substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

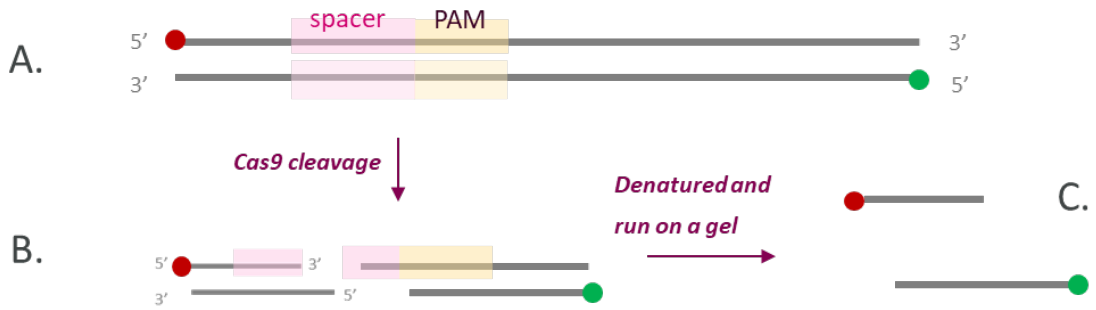


Figure A.3: Principle of cleavage assay with fluorophore-tagged substrate with a determined PAM sequence of either TGG or CCC. The dsDNA substrate, fluorophore-tagged at either 5' end, is incubated with the Cas9 RNP, during which substrates containing a recognized PAM sequence are cleaved. The cleaved substrate is then denatured and run on a 15 % UREA gel.

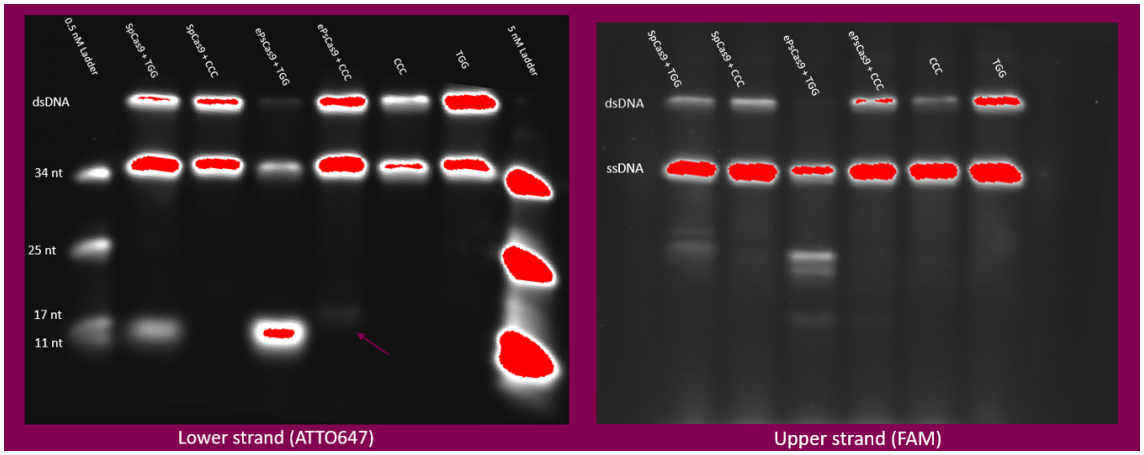
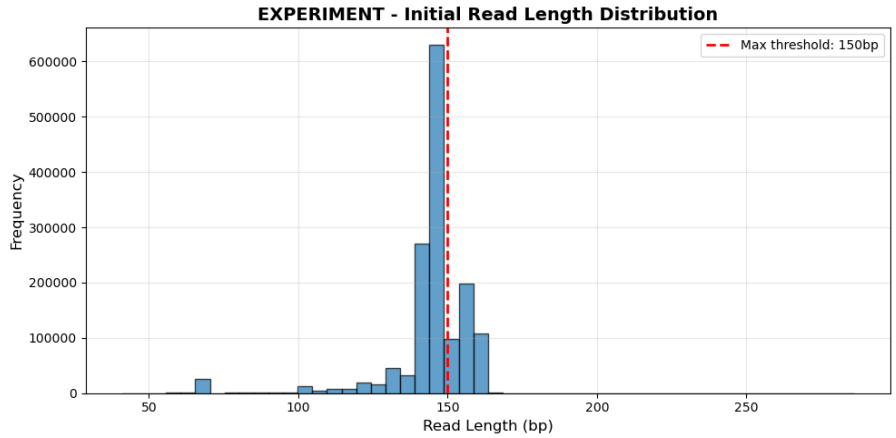
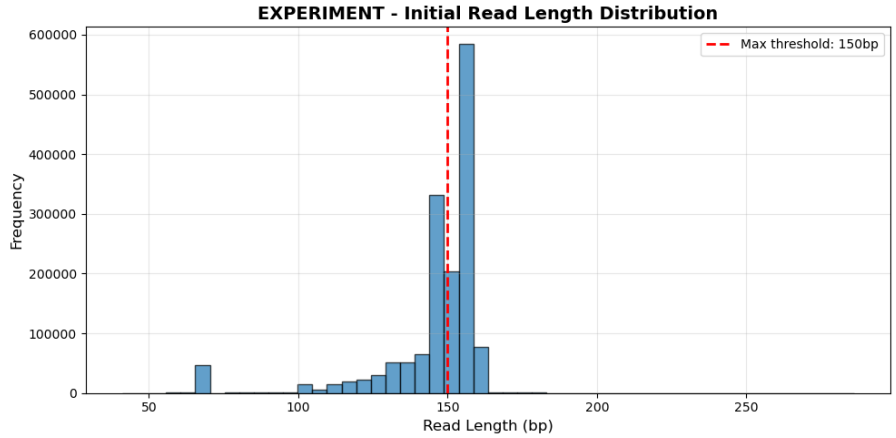


Figure A.4: Fluorescent detection of Cas9-mediated dsDNA cleavage products. Denatured cleavage products of dsDNA substrates labeled with ATTO647 or FAM on opposite 5' ends were separated on a 15 % UREA gel. Fluorescence imaging was performed independently in the ATTO647 and FAM channels to visualize strand-specific cleavage products.

(a)



(b)



(c)

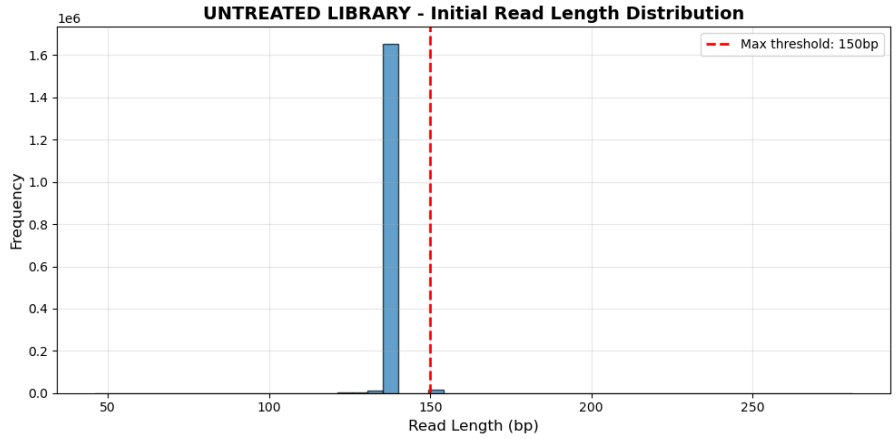
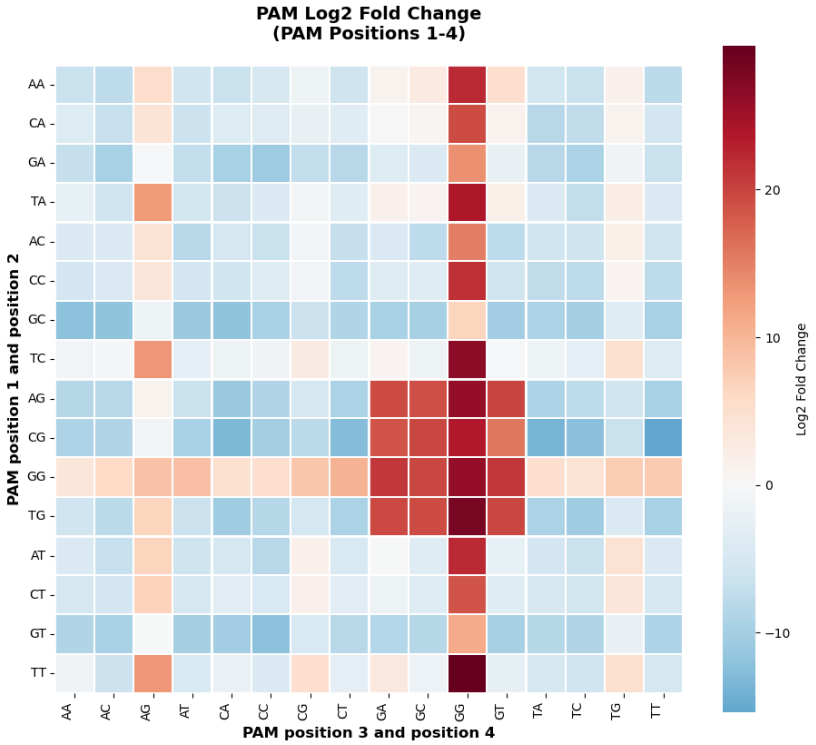


Figure A.5: Read length of the cleavage assay using the EMX1b substrate library and (a) purified ePsCas9 protein and (b) cell lysate containing ePsCas9 protein. (c) Read length of the untreated EMX1b substrate library.

(a)



(b)

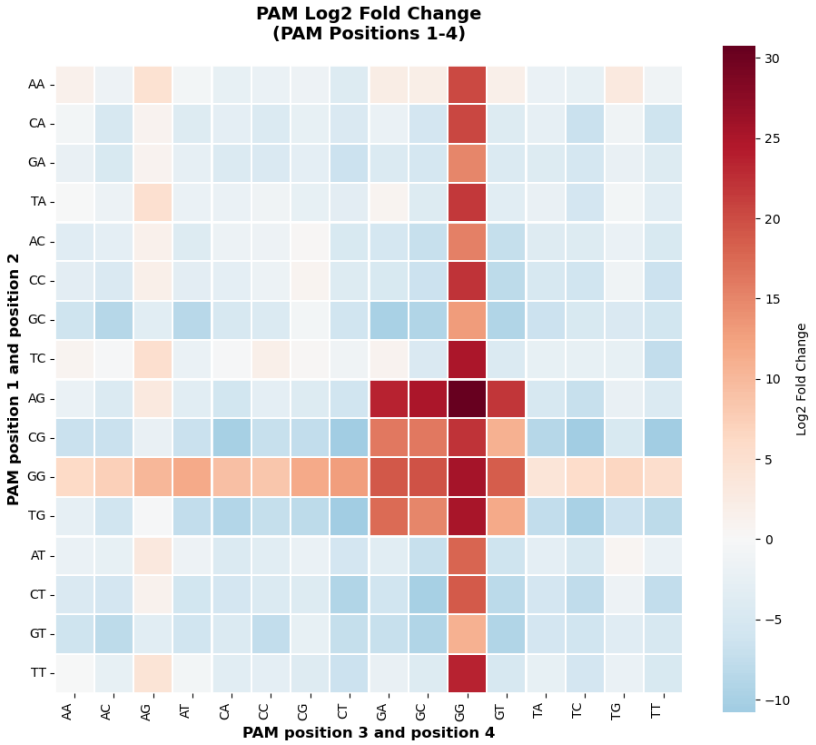
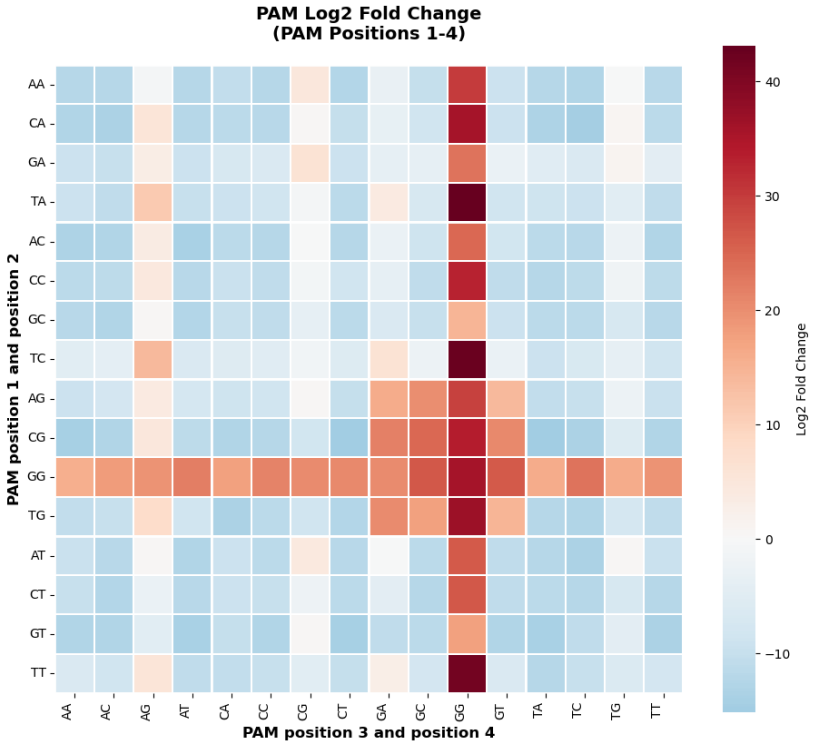


Figure A.6: The PAM Log_2 Fold Change from the cleavage samples for ePsCas9 expressed in cell lysate and (a) The PCSK9 or (b) the EMX1b substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

(a)



(b)

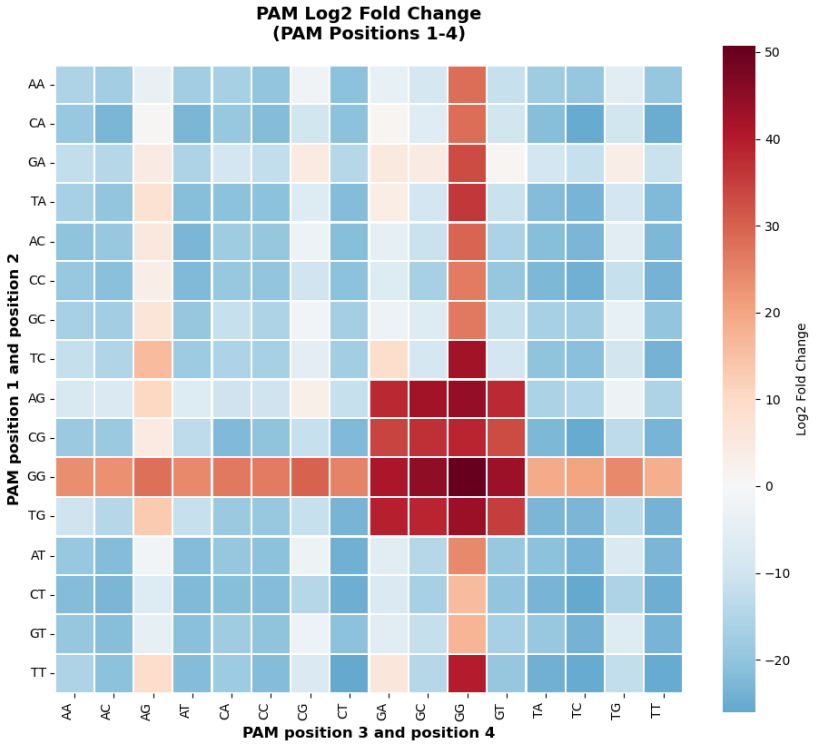
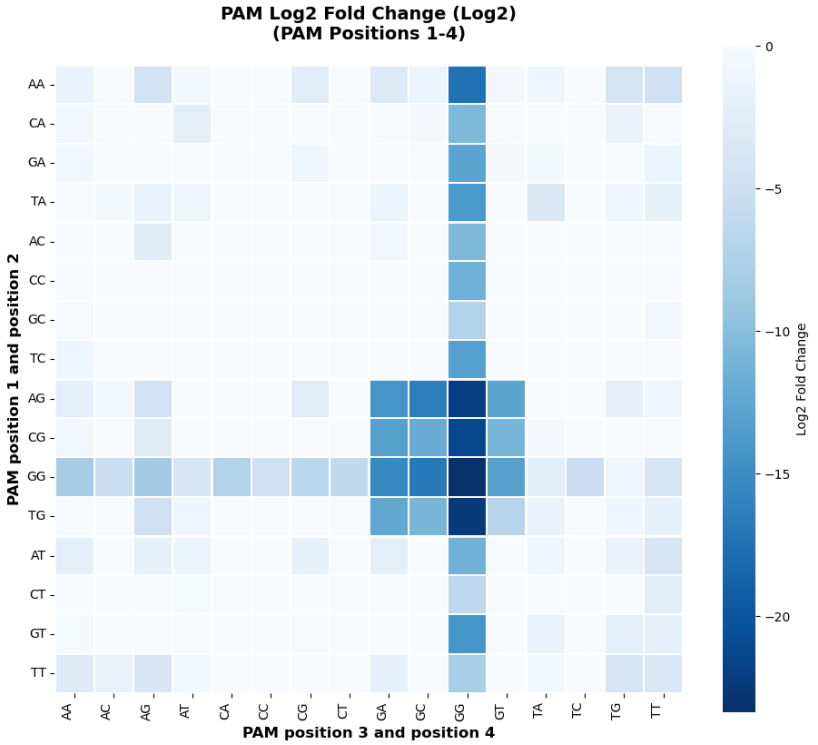


Figure A.7: The PAM Log_2 Fold Change from the cleavage samples for SpCasp expressed in cell lysate and (a) The PCSK9 or (b) the EMX1b substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

(a)



(b)

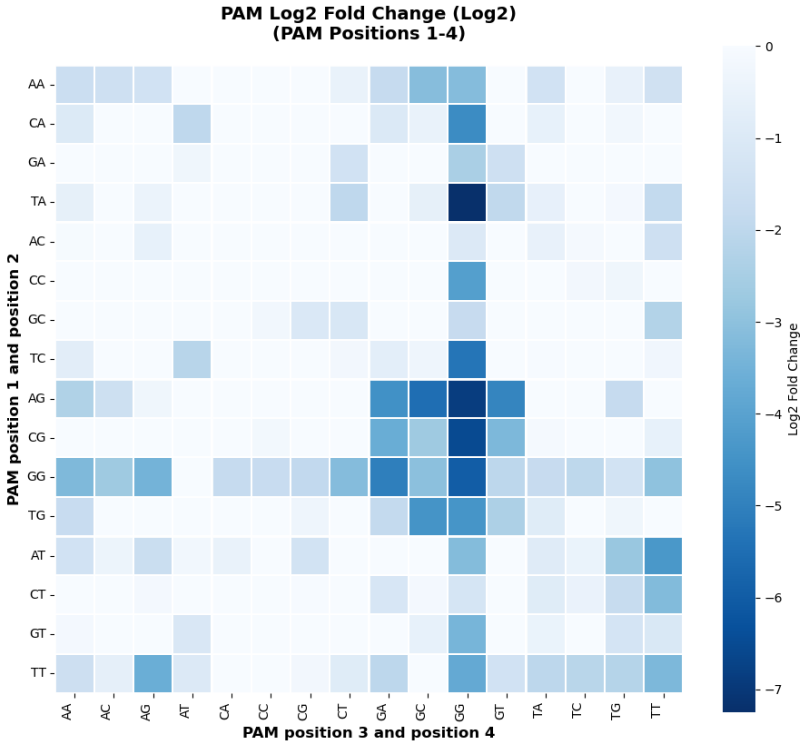
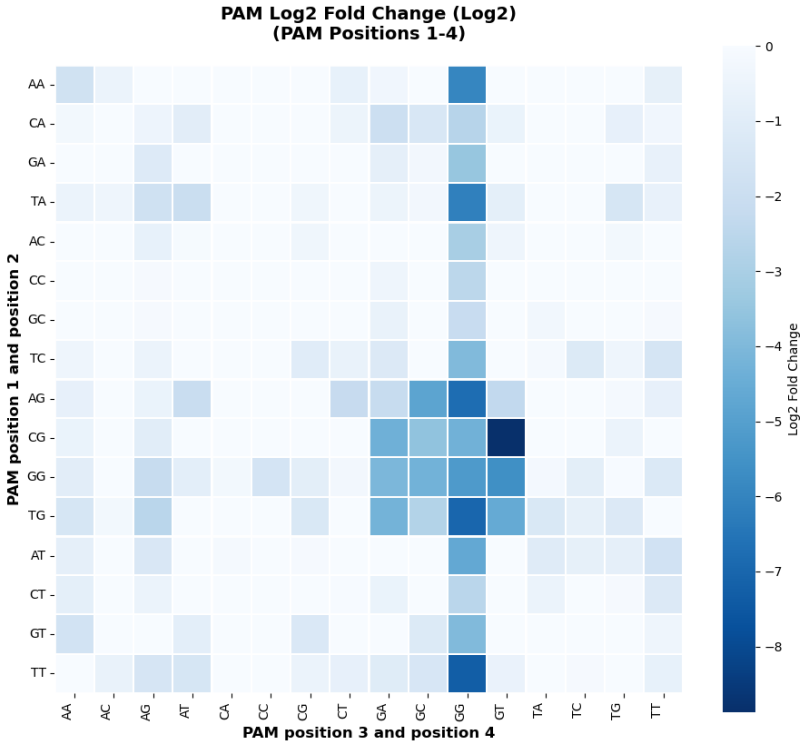


Figure A.8: The PAM Log_2 Fold Change from the depletion samples for (a) purified SpCas9 protein and (b) purified ePsCas9 protein and the EMX1b substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

(a)



(b)

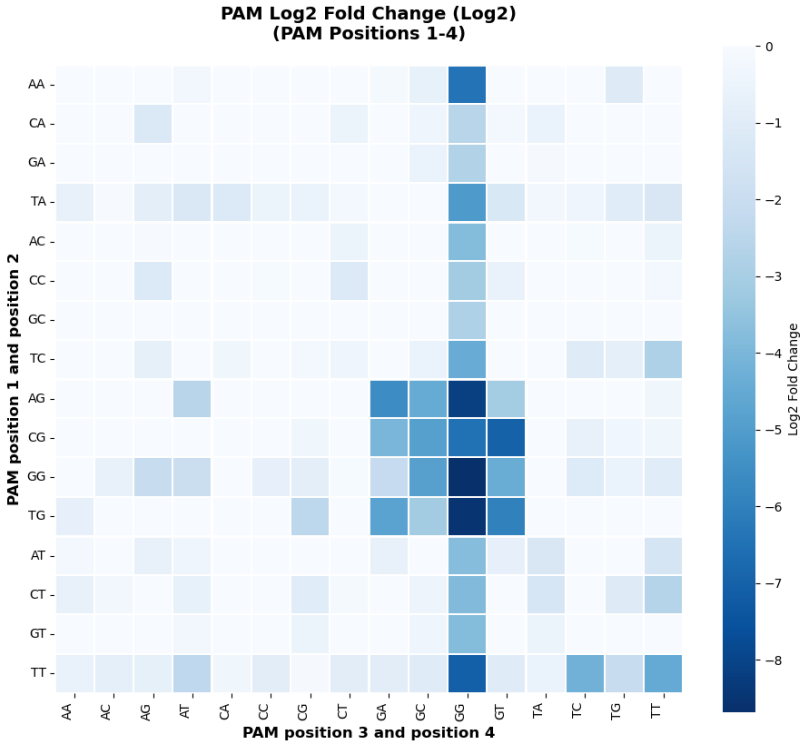


Figure A.9: The PAM Log_2 Fold Change from the depletion samples for (a) SpCas9 expressed in cell lysate and (b) ePsCas9 expressed in cell lysate and the PCSK9 substrate library. The first and second PAM nucleotides are plotted on the y-axis, and the third and fourth PAM nucleotides are plotted on the x-axis.

B

Primers and Oligonucleotide Sequences

In this section the primers, oligonucleotides, and spacer sequences used in this project are listed. "INDEX-" indicates the annealing sequence of indexing amplification performed by the NGS facility at AstraZeneca. "XXXX" indicates the location of the 4-nucleotide barcode.

Table B.1: Primer names and explanations of primers.

Primers	**Sequences or Explanations**
CN9	Forward primer annealing outside the start of the WED and PI domain
CN10	Reverse primer annealing at the end of the WED and PI domain
CN9G	CN9 with overhangs that contain BsaI recognition sites
CN10G	CN10 with overhangs that contain BsaI recognition sites
CN11	Reverse primer annealing outside the start of the WED and PI domain
CN13	Forward primer annealing outside the end of the WED and PI domain
CN38	Forward primer annealing at the start of the ePsCas9 coding sequence
CN39	Reverse primer annealing outside the start of the WED and PI domain
CN12G	CN38 with overhangs that contain BsaI recognition sites
CN11G	CN39 with overhangs that contain BsaI recognition sites
CN14G	Reverse primer annealing at the start of the ePsCas9 coding sequence with BsaI recognition sites.
CN30	Forward primer annealing at the start of the ePsCas9 coding sequence.
CN31	Reverse primer annealing at the end of the ePsCas9 coding sequence.
CN40	Reverse primer annealing at the start of the ePsCas9 coding sequence
CN33	Forward primer annealing at the end of the ePsCas9 coding sequence.
CN35	Forward primer introducing a silent mutation to remove BsaI recognition site.
CN36	Reverse primer containing a 15-nucleotide overlap with CN35 and the silent mutation.
CN45	Forward primer for introducing S1223A
CN47	Reverse primer for introducing S1223A
CN46	Forward primer for introducing I1222A and S1223A.
CN48	Reverse primer for introducing I1222A and S1223A.

Table B.2: Primers and sequences used for amplicon sequencing in the luciferase assay.

Primers	**Sequences or Explanations**
Fwd-reporter	INDEX-XXXXGAGGAGCAGGATGATGGCAC
Rev-reporter	INDEX-XXXXGTTGAGAAACTATTGCTTCCT

Table B.3: Primers and sequences used in the PAM assay.

Primers or oligonucleotides	**Sequences**
oCN006-EMX1b-6N	AAAGCCTGGCCAGGGAGTGGCCAGAG TCCAGCTTGGGCCCACGCAGGGGCCT GGCCAGCAGCAAGCAGCACTCTGCCCT CGTGGGTTTGTGGTNNNNNNTTCTTC TTCTGCTCGGACTCAGCGATGTCC
oCN007-EMX1b-TGG	AAAGCCTGGCCAGGGAGTGGCCAGA GTCCAGCTTGGGCCCACGCAGGGGC CTGGCCAGCAGCAAGCAGCACTCTGC CCTCGTGGGTTTGTGGTGGGCCATTCT TCTTCTGCTCGGACTCAGCGATGTCC
oCN008-PCSK9-6N	AAAGCCTGGCCAGGGAGTGGCCAGAG TCCAGCTTGGGCCCACGCAGGGGCCTG GCCAGCAGCAAGCAGCACTCTGCCCTC GTGGGTTTGTGGTNNNNNNAGAGCA TCCCGTGGAACCTGGACGATGTCCA
oCN009-PCSK9-TGG	AAAGCCTGGCCAGGGAGTGGCCAGAG TCCAGCTTGGGCCCACGCAGGGGCCTG GCCAGCAGCAAGCAGCACTCTGCCCTC GTGGGTTTGTGGTGGGCCAAGAGCA TCCCGTGGAACCTGGACGATGTCCA
oOK001-EMX1a-6N	AAAGCCTGGCCAGGGAGTGGCCA GAGTCCAGCTTGGGCCCACGCAG GGGCCTGGCCAGCAGCAAGCAGCACT CTGCCCTCGTGGGTTTGTGGTNNNNNN CCCTAGTCATTGGAGGTGACATCGATGTCC
pCN002-EMX1b-biotin	/52-Bio/GGACATCGCTGAGTCCGAGCA
pCN003-PCSK9-biotin	/52-Bio/TGGACATCGTCCAGGTTCCACG
pCN001-EMX1a-biotin	/52-Bio/GGACATCGATGTCACCTCCAATGAC
oZA080	CAATCGGGCTGGACATCGGAACTAACTCACT
oZA081	GTGAGTTAGTTCCGATGTCCAGCCCGATTGC
nZA101	INDEX-XXXXCAATCGGGCTGGACATCGGAAC
nZA111	INDEX-XXXXAAAGCCTGGCCAGGGAGTGG
nCN002	INDEX-XXXXGGACATCGCTGAGTCCGAGCA
nCN003	INDEX-XXXXTGGACATCGTCCAGGTTCCACG
oCN002-sense CCC-PAM	/56-FAMN/AGTCGATGTCACCTCCAATG ACTAGGGCCCCCCAC
oCN003-antisense CCC-PAM	/5ATTO647/GTGGGGGGCCCTAGTCATT GGAGGTGACATCGACT
oCN004-sense TGG-PAM	/56-FAMN/AGTCGATGTCACCTCCAATG ACTAGGGTGGCCCAC
oCN005-antisense TGG-PAM	/5ATTO647/GTGGGCCACCCTAGTCA TTGGAGGTGACATCGACT

Table B.4: Primers and sequences used in the benchmarking experiment.

Gene and PAM	**Forward primer**	**Reverse primer**
MECP2-NTG	INDEX-XXXX GCCTCTTGGTTG TAATATGCAGGCG	INDEX-XXXX CATCAGAGAGCATT GATCACAGATCACC
EMX1-NTG	INDEX-XXXX CCTGGGACCAC TTGGCCTTCTC	INDEX-XXXX AGCTGGATGCCC GTGTCATTAAGAG
FANCF-NTG	INDEX-XXXX GGGACTCAGTTC CAACCCAAATGC	INDEX-XXXX GGCATCCACAAA TCACCTGGAGAG
EMX1-NAG	INDEX-XXXX GGCACCACTGTAG TTTAGTGATCCC	INDEX-XXXX ACCCTTGACCC CCTCCACCAG
FANCF-NAG	INDEX-XXXX CTCTCCAGGTGA TTTGTGGATGCC	INDEX-XXXX GGAGGACTCTC TGATGAAGACCCA
DYRK1A-NAG	INDEX-XXXX CAAATAATGAGGG TTACAGTTGCCCTC	INDEX-XXXX AACATCACTGAGT ATACACTGCACGC
SHANK3-NGC	INDEX-XXXX CGCCGACCCT TCCAGCAGAA	INDEX-XXXX GTGACGCCC AGCTCCACGA
FANCF-NGC	INDEX-XXXX CCACCTCCTG CAGACGCTCC	INDEX-XXXX GGTGCAGCAA CTCTTTCCCGG
UBE3A-NGC	INDEX-XXXX GGGATAAGTGGTT TTCGACAATCCAGG	INDEX-XXXX GGGAAGTGGACAC AAATCACAATGAAG
EMX1-NGG	INDEX-XXXX CAGAACCGGAGGA CAAAGTACAAACG	INDEX-XXXX GCAGCAAGCA GCACTCTGCC
DYRK1A-NGG	INDEX-XXXX CCCATTGCAACT TCCAGTCCTGC	INDEX-XXXX CGCTAGACGGTA GAGCCTACTGC
UBE3A-NGG	INDEX-XXXX GGGAAATCCAGT GAGAAGACAGCAG	INDEX-XXXX CTGCCCCTCTTT CACTAGCTTCATGG

Table B.5: sgRNA spacer sequences used in the benchmarking experiment for SpCas9 and ePsCas9.

sgRNA spacer	**Sequences**
MECP2-NTG-SpCas9	GATACTGAGTTTTTAACAAA
EMX1-NTG-SpCas9	GGGATCACTAAACTACAGTG
FANCF-NTG-SpCas9	GTGCAAACGTTGACCCCAGT
EMX1-NAG-SpCas9	GGATGCCCGTGTCATTAAGA
FANCF-NAG-SpCas9	GCAGGAGGTGGGGAAGGCCG
DYRK1A-NAG-SpCas9	GGTTTGCAGCCTAAGAGCAG
SHANK3-NGC-SpCas9	GGACCATGAGATAGAAGGCG
FANCF-NGC-SpCas9	GAGACACTCCAAGAGAGCCT
UBE3A-NGC-SpCas9	GCCCATCCCTGAGTCCAGCG
EMX1-NGG-SpCas9	GAGTCCGAGCAGAAGAAGAA
DYRK1A-NGG-SpCas9	GTTGCCCTCATTATTCAGCA
UBE3A-NGG-SpCas9	GTACAGTTAGTACTCAGCAG
MECP2-NTG-ePsCas9	CAGATACTGAGTTTTTAACAAA
EMX1-NTG-ePsCas9	TGGGGATCACTAAACTACAGTG
FANCF-NTG-ePsCas9	TAGTGCAAACGTTGACCCCAGT
EMX1-NAG-ePsCas9	CTGGATGCCCGTGTCATTAAGA
FANCF-NAG-ePsCas9	CTGCAGGAGGTGGGGAAGGCCG
DYRK1A-NAG-ePsCas9	CAGGTTTGCAGCCTAAGAGCAG
SHANK3-NGC-ePsCas9	GAGGACCATGAGATAGAAGGCG
FANCF-NGC-ePsCas9	AGGAGACACTCCAAGAGAGCCT
UBE3A-NGC-ePsCas9	GAGCCCATCCCTGAGTCCAGCG
EMX1-NGG-ePsCas9	CTGAGTCCGAGCAGAAGAAGAA
DYRK1A-NGG-ePsCas9	CAGTTGCCCTCATTATTCAGCA
UBE3A-NGG-ePsCas9	GTGTACAGTTAGTACTCAGCAG

C

Experimental protocols

C.1 Electroporation protocol

Prior to electroporation pre-warm SOC and agar plates at 37^{o}C and place electroporation cuvettes (1mm) and labelled microcetrifuge tubes on ice.

1. Thaw 50 µL electrocompetent BW25141 (DE3) *E. coli* containing desired selection plasmids.
2. Add 1 µL of plasmid DNA/plasmid library and transfer the cell/DNA to a labelled microcentrifuge tube.
3. Electroporate the cuvette at 2.0 kV. Time constant should be 4-5 ms.
4. Immediately add 950 µL of SOC medium and transfer to labelled microcentrifuge tube.
5. Shake vigoorsly (250 rpm) at 37^{o}C for 1 hour.
6. Streak the 1 mL transformation on agar plates containing arabinose and chloramphenicol.
7. Incubate plates overnight at 37^{o}C.

C.2 Preparation of lysis buffer (1X)

The following reagents were used to prepare 10 mL of lysis buffer.

Table C.1: Reagents and volumes for preparing 10 mL of lysis buffer 1X.

Lysis buffer (1X)	**Volume**
Hepes pH 7.5	200 µL
KCl	1000 µL
$MgCl_2$	50 µL
Glycerol	500 µL
SIGMAFAST Protease Inhibitor Cocktail tablet	1 tablet
DTT	10 µL
Triton X-100	10 µL
Nuclease-free water	up to 10 mL
Total	10 mL

C.3 Preparation of agar plates

Dissolve 37 g/L LB Agar (Thermfisher scientific, cat. #22700025) in Mille-Q water and autoclave the contents. Add chloramphenicol (Sigma-Aldrich, cat. #c0378), carbinicillin (Sigma-Aldrich, cat. #C3416) or L-arabinose (Sigma-Aldrich, cat. #A3256-25G) to a final concentration of 25 µg/mL, 100 µg/mL, and 100 mg/mL, respectively. Then mix and pour into the desired plate. Store prepared agar plates at -20^{o}C.

C.4 Preparation of TB medium

Reagents in table C.2 were added together and autoclaved to produce 1 L of TB medium. The TB medium can be stored at 4^{o}C.

Table C.2: Terrific Broth (TB) medium composition.

Component	**Amount or volume**
Tryptone	12 g
Yeast extract	24 g
Glycerol	4 mL
KH_2PO_4	2.3 g
K_2HPO_4	12.54 g
Milli-Q H_2O	up to 1 L

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY