Stabilizing Multiplex crRNA Arrays via High-Throughput Mutation for the CRISPR-Cas12a Endonuclease

W. Baksht, S. Biggio, C. Bruce, B. Dulitsky, S. Eichorn, J. Geffner, T. Jayaraj, B. King, I. Kulstad, L. Li, M. Lorino, R. Mohan, T. Nguyen, N. Prus, A. Smith, V. Sutharsan, S. Trubin, L. Woelfle, N. Sandoval, Ph.D.

1: Introduction

Genome editing is a method of genetic engineering where DNA can be freely inserted, replaced, or deleted from a living organism's genome. Currently, genome editing is widely used for cutting-edge research in biomedicine, biotechnology, and synthetic biology (Paul and Montoya, 2020). To edit a DNA strand, a double stranded break (DSB) must be created at a specific locus. One recently introduced technique for creating a DSB is the use of a CRISPR-Cas complex, which allows for more versatility than the existing engineered nucleases. CRISPR complexes use altering nucleic acid sequences in their guide RNA to determine the location of attachment in the operon, while other nuclease complexes are protein-driven. The significantly smaller scale of nucleic acid alteration has solidified CRISPR-Cas complexes as the more commonly used method (Leitão et al., 2017). One of the most utilized and well-known CRISPR-Cas complexes is CRISPR-Cas9 (Xu and Li, 2020). It has gained recognition in the news and the media in recent years due to its ability to fundamentally alter the genetic code. As a result, harvesting and manipulating the CRISPR-Cas complex opens a wide range of possibilities.

The hereinafter described project was designed to improve the capabilities of the CRISPR-Cas12a enzyme. This protein-nucleic acid complex has gained popularity because of its potential to free humans from genetic disease by activating genes that lead to beneficial phenotypes and removing genes that elicit harmful ones. For this project, CRISPR-Cas12a, formerly known as Cpf1, was chosen over CRISPR-Cas9 because it only requires one RNA molecule called crRNA, while the CRISPR-Cas9 system requires two RNA molecules, tracrRNA and crRNA. (Paul and Montoya, 2020). This is because CRISPR-Cas12a processes its own crRNA in its ribonuclease site (Paul and Montoya, 2020). Another important feature of Cas12a is that the "sticky ends" it creates allow for a more efficient insertion of DNA at the cut site (Swarts and Jinek, 2018). Lastly, the Cas12a system also possesses a PAM sequence upstream of the target sequence, unlike the Cas9 system. The presence of PAM sequences, protospacer adjacent motifs, allow the Cas endonuclease to distinguish between the bacteria's own DNA and the DNA of an invading virus. The upstream PAM sequences enable the Cas enzyme to locate other PAM sequences more quickly and edit faster (Sanders, 2018). Because of these characteristics, we decided to utilize CRISPR-Cas12a as it appears to be a more versatile tool for genome editing compared to CRISPR-Cas9.

Homologous recombination is commonly understood to be a mechanism by which cells maintain genome plasticity, such as by repairing DSBs. However, homologous recombination sometimes results in errors like the deletion of entire DNA sequences. This presents a challenge to scientists looking to genetically modify crops used for biofuels. The hairpin-like structure of the scaffolding region of the crRNA is prone to homologous recombination due to the frequency of the repeated segments - the 19 base pair sequences are the same repeatedly in multiplexed versions. This experiment seeks to reduce homologous recombination, i.e. improve the fitness, of the CRISPR-Cas12a system by introducing mutations into the scaffolding region. The enzyme's overall fitness was then calculated using a logarithmic formula to categorize mutations as beneficial or deleterious. Beneficial mutations reduced homologous recombination, thereby increasing the efficacy of the CRISPR-Cas12a system. As this system has various uses in agriculture and biofuels, this research has important implications for developing more efficient crops.

2: Methods

2.1: General

All *in vivo* experiments were run using the NEB5α strain of *E. coli*. The plasmids used in the study were derived from pCRJ004 (Joseph and Sandoval, 2022). GeneWiz performed both Sanger sequencing and NGS. Each primer used in this study was synthesized by integrated DNA technologies.

2.2: Plasmid Construction of *Fn*Cas12a Scaffold Stem Loop Sort-Seq Identified Validations

To introduce mutations, insertions, and deletions, the Q5 High-Fidelity 2X Master Mix (New England BioLabs [NEB]) was used as per the manufacturer's guidelines. The specified primers and the NEB5α plasmid served as the initial template for 10-µl volume PCR reactions. Visualization of the PCR products was done on 0.8% agarose to confirm the size and successful amplification. Once confirmed, the products were purified using the Qiagen QIAquick PCR Purification Kit. The 3' ends of PCR products were phosphorylated with T4 Polynucleotide Kinase (NEB) and then ligated with Instant Sticky-end Ligase Master Mix (NEB) before incorporation into chemically competent NEB5α cells (NEB). Transformants were selected on LB agar containing 100 µg/ml ampicillin (Amp 100). Select colonies underwent overnight cultures for plasmid miniprep and were subsequently sent for Sanger sequencing.

2.3: Scaffold Library Generation and Assembly

Site-directed mutagenesis was used to build a mutant library by constructing oligonucleotides with incorporated random point mutations. In other words, these sites had degenerate mutations. These forward and reverse primers were added to the pCRJ004 plasmid via PCR. The products of this PCR were linear DNA amplicons with edited crRNA scaffolds. The plasmids coded for "N," any

of the four nucleotide bases, and generated endless mutated combinations of CRISPR tandem repeats to test which mutations resulted in a successful CRISPR—Cas12a system.

2.4: Library Sorting

Two to four frozen glycerol library stocks were inoculated into 50 ml LB Amp100 and grown in 250 ml baffled flask (200 rpm, 32°C) to start the library seed cultures. The cultures reached an OD600 of ~0.20 (8-9 hr) and the library was induced with 0.1% lactose with CRISPR construct and incubated for 16 hr (250 rpm, 30°C; Thermo MaxQ). Library cultures were normalized to an OD600 of ~0.1 and put in phosphate buffered saline with pH 7.4 (ThermoFisher) to dilute to a factor of 4, then iced before they were sorted.

Cells were sorted into two bins based on GFP expression at a rate of ~9,000 events per second on the iSort Automated Cell Sorter with a blue solid-state laser (488 nm, 165 mW), optical filters 525/50 BP for GFP and 488/10 SCC, 85 µm ceramic nozzle with a fixed sample flow rate of 23 µl/min (ThermoFisher). The low GFP bin contained 35% (~9,000 events) and the high GFP bin contained 65% (~170,000 events) of the population. The recovered sorted samples were then plasmid miniprepped (Qiagen QIA Spin Miniprep Kit) for NGS sample preparation.

2.5: Deep Sequencing and Sample Prep

Plasmids from both populations underwent extraction using the Qiagen QIAprep Spin Miniprep Kit and served as templates for subsequent rounds of PCR. The generation of sequencing libraries involved the amplification of 1000 ng of plasmid templates through 10 cycles of PCR, employing Q5 High-Fidelity 2X Master Mix and primers designed with partial Illumina adapters featuring overhands on the 5’ end. After confirming the expected amplicon size (6128 bp) on a 0.8% agarose gel, Qiagen QIAquick PCR Purification Kit was used to eliminate any residual impurities from the remaining plasmid. Following this purification, the plasmids were digested with DpnI enzyme, then phosphorylated. The plasmids were then purified again and transformed into the *E. coli* cells. After transformation, the cells were then miniprepped to isolate the plasmids and the plasmids were sent for a sequencing service offered by GeneWiz.

2.6: NGS Pre-Processing Workflow

Pooled libraries were deep sequenced using 2 x 250 bp paired-end Illumina Miseq sequencing (Genewiz). To analyze the resulting FASTQ data, the UseGalaxy.org platform was used (Afgan et al., 2018). A FASTQ file containing 329,816 raw sequence reads underwent quality control analysis with the FASTQC program (Andrew, 2010). Reads were demultiplexed with four 5 bp barcodes via Barcode Splitter (Gordon, 2010). Read trimming was performed to filter out low-quality sequences (quality score > 25) and trim the barcode using Trimmomatic. The sequence was then cropped to the specific region of interest (Bolger, et al., 2014). Trimmed reads were mapped to the reference sequence using Bowtie2 and unmapped reads were removed (Langmead and Salzberg, 2012). The aligned FASTQ files were exported for further analysis. Within each population of the experiment, we calculated the frequency of each base at each position. Then, the abundance and fitness scores were calculated.

2.7: Individual Scaffold Validations

Biosensor plasmids in *E. coli* NEB5α strains were cultured overnight in LB medium (250 rpm, 37°C; Thermo MaxQ) supplemented with 100 µg/ml ampicillin (Amp100). Cultures were diluted in LB with Amp100 medium, with or without lactose, to a final volume of 160 µl in sterile Corning 96-well flat clear bottom black microplates and sealed with breathable rayon film (VRW).

3: Results

The relative "fitness levels" of each of the mutated ML regions was first determined by using a ratio between the number of trials in which iLOV, a green fluorescent protein, expression was reduced because of SNPs within the ML region and the number of trials in which iLOV expression was unaffected, and then by taking the binary logarithmic function of this ratio. These fitness values are representative of the ability of the different ML regions to sustain SNPs without compromising the functionality of the Cas-12a scaffold and, by extension, CRISPR's ability to remove the iLOV gene.

Smaller and more negative average fitness values are directly correlated with a decrease in overall iLOV expression, indicating a loss of function in the scaffold region due to SNPs within a specific ML region. The larger positive fitness values are directly correlated with a decrease in overall iLOV expression, indicating improved function of the scaffold region. The fitness levels of each of the mutated ML regions indicate the effect of various SNPs within the respective ML regions on the overall ability of the Cas-12a enzyme to eliminate the target gene (iLOV). Lastly, induced mutations with less than 40 reads were not included in the figure below to ensure the results presented are statistically significant.

The results suggest that all mutations made to the ML-3 region resulted in decreased fitness of the CRISPR-Cas12a enzyme. To a lesser extent, mutations made to ML-5 also resulted in lowered fitness. In contrast, the ML-1 and ML-2 regions were better able to withstand mutations. Most mutations to these two regions resulted in zero to positive fitness levels, indicating potential neutral or beneficial effects of mutations on fitness. The data for ML-1 does not indicate a significant difference between the fitness levels of single nucleotide mutated strands compared to multiple nucleotide mutated strands. Only one specific mutation in ML-6 produced enough reads to analyze; this read had high fitness levels and could withstand mutations. The reads for ML-7 had relatively low counts for both high and low iLOV expression but overall, the reads had high, positive fitness values. The majority of ML-8 was able to retain function when mutated except for a small number of mutations with low read counts. Overall, the mutations made to ML-3, ML-4, and ML-5 illustrate that these are core regions that cannot be mutated since altering them results in a loss of function. On the other hand, mutations made to ML-8 show that these regions are still functional after being mutated. This would mean that this would be a good target for designing mutated hair-pin structures.

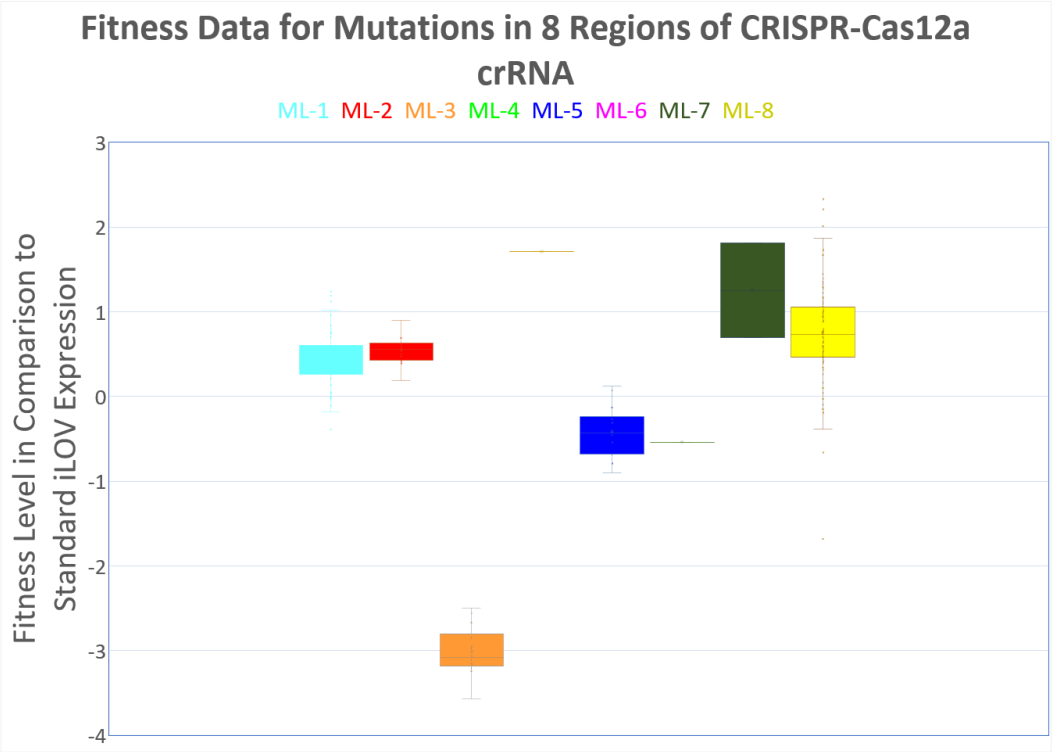

Figure 1. Box-and-whiskers graph depicting fitness levels for mutations in each region of CRISPR-Cas12a crRNA.

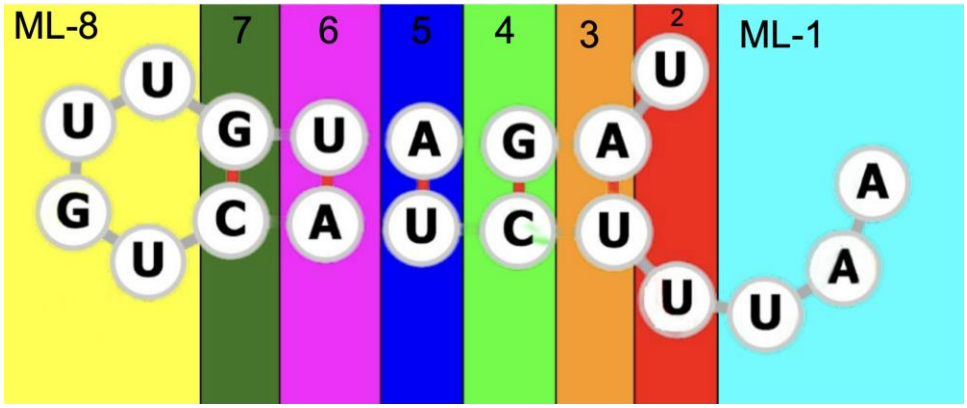

Figure 2. Diagram depicting the eight regions of the crRNA hairpin. Colors highlighting each region correspond to those used in Figure 1.

4: Conclusion

The CRISPR-Cas12 crRNA region is highly susceptible to homologous recombination due to the proximity of repeated elements. Homologous recombination slows or, in some cases, inhibits the enzyme's ability to multiplex. Homologous recombination occurs most frequently in regions with complementary, non-bounded sequences of nucleotides on opposing strands. These nucleotides bind, and leave unbound single strands, which are promptly cleaved. This experiment aims to

identify specific regions—more precisely, certain nucleotides—within the CRISPR-Cas12 crRNA hairpin that can be mutated to reduce the probability of homologous recombination and thereby increase the functionality of the enzyme. The collected data represents the ability of each region to be mutated without compromising the scaffold region and by extension the entire enzyme's ability to block expression of the iLOV gene. Unfortunately, a fully informed conclusion on the mutability of the scaffold regions cannot be made in this paper as the flow cytometry data is not yet available. The verification of the scaffold regions mutability, specifically the level of iLOV gene expression identified via flow cytometry, is still ongoing.

The higher mutability and stability of CRISPR-Cas12a compared to CRISPR-Cas9 will allow for further implementation of this system. The versatility of CRISPR-Cas12a makes it a great candidate for genetic engineering projects, especially for genetically modified organisms (GMOs) such as plants. While our project focused on prokaryotic cells, CRISPR in agriculture is expanding to create genetically modified plants. Using CRISPR-Cas12a in genetically modified schemes will help increase agricultural productivity as climate conditions worsen (Bandyopadhyay et al., 2020).

Acknowledgements

Tulane SynBio would not be possible without the support we have received from Tulane University. First, we want to thank Dr. Nicholas Sandoval for his endless mentorship, ideas, and funding. We thank the Tulane University Department of Chemical and Biomolecular Engineering for providing the laboratory facilities necessary to conduct this research, and the National Science Foundation (NSF) for funding the research done in Dr. Sandoval's lab, and concurrently, our research with EPSCoR (CAREER award 1847226). We also want to thank the following members of the Sandoval Lab for their mentorship: Nancy Kim, Mitch Wheeler, and Lily Rothschild. Additionally, we thank the following founding Tulane SynBio members for spearheading this project: Alina Tiemroth, David Sirkin, Josh Frydman, Joel Winderweedle, and Keefer Boone. Finally, we extend our gratitude to the Tulane University Registered Student Organizations for funding our lab work and offsetting the cost of supplies.

Human Centered Design

The stakeholders of this project could use its results and data to help generate a potential model in future CRISPR research. The mutations and their respective effects on the efficacy of the CRISPR-Cas12a system alter the potential use of this genetic technology. Future work in this field can use the mutated segments produced from this project to observe further downstream effects of these changes such as how these different versions interact with different organisms, the potential changes to the molecular mechanism, and how translatable these results are to different CRISPR-Cas systems. It is important to consider how future work that involves this project must consider the ethical and biosafety rules that are intertwined with this work. Future work could include CRISPR-Cas12a applications in agriculture to create genetically modified plants with increased productivity and greater survivability.

Biosafety

This project has met the biosafety rules that were defined by GOGEC. Our team and our research verified that the various aspects of the procedure did not interfere with the GOGEC Biosafety and Biosecurity Blacklist. Firstly, there was no experimentation that was conducted on human subjects or samples. Next, we used NEB5α *E. coli*, which is considered a Risk Group 1 microorganism. Additionally, although we used CRISPR technology, it was never used on any organism other than the NEB5α *E. coli* as previously stated. Lastly, our work did not help pathogens interfere with the immune system or interfere with the normal host cell's processes of replication, transcription, and translation. Although we are mutating coding sequences which interfere with the processes of replication, transcription, and translation, they are not being altered in a host organism and are instead solely altered in the pathogen itself.

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research, 46(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Andrew, S. FastQC A Quality Control tool for High Throughput Sequence Data. In.; 2010.

Bandyopadhyay, A., Kancharla, N., Javalkote, V. S., Dasgupta, S., & Brutnell, T. P. (2020). CRISPR-Cas12a (Cpf1): A Versatile Tool in the Plant Genome Editing Tool Box for Agricultural Advancement. Frontiers in plant science, 11, 584151. https://doi.org/10.3389/fpls.2020.584151

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Gordon, A. FASTQ/A short-reads pre-processing tools. In.; 2010. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (New York, N.Y.), 337(6096), 816–821. https://doi.org/10.1126/science.1225829.

Joseph, R. C., & Sandoval, N. R. (2022). Single and multiplexed gene repression in solventogenic *Clostridium* via Cas12a-based CRISPR interference. *Synthetic and systems biotechnology*, *8*(1), 148–156. https://doi.org/10.1016/j.synbio.2022.12.005.

Kim, N. M., Sinnott, R. W., Rothschild, L. N., & Sandoval, N. R. (2022). Elucidation of Sequence–Function Relationships for an Improved Biobutanol In Vivo Biosensor in E. coli. Frontiers in bioengineering and biotechnology, 10, 821152. https://doi.org/10.3389/fbioe.2022.821152.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923.

Leitão, A. L., Costa, M. C., & Enguita, F. J. (2017). Applications of genome editing by programmable nucleases to the metabolic engineering of secondary metabolites. Journal of biotechnology, 241, 50–60. https://doi.org/10.1016/j.jbiotec.2016.11.009.

Paul, B., & Montoya, G. (2020). CRISPR-Cas12a: Functional overview and applications. Biomedical journal, 43(1), 8–17. https://doi.org/10.1016/j.bj.2019.10.005.

Sanders, R. (2018, February 15). CRISPR scissors: Cas12a enables cutting-edge diagnostics. Berkeley News. Retrieved from https://news.berkeley.edu/2018/02/15/crispr-scissors-cas12a-enables-cutting-edge-diagnostics/#:~:text=Cas12a%2C%20discovered%20in%202015%20and%20originally%20called%20Cpf1%2C,turned%20into%20a%20powerful%20gene-editing%20tool%20in%202012.

Swarts DC, Jinek M. Cas9 versus Cas12a/Cpf1: Structure–function comparisons and implications for genome editing. WIREs RNA. 2018; 9:e1481. https://doi.org/10.1002/wrna.1481.

Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. Computational and structural biotechnology journal, 18, 2401–2415. https://doi.org/10.1016/j.csbj.2020.08.031.