Directed Evolution and Engineering of CRISPR-Associated Nucleases

Citation

Davis, Kevin Michael. 2016. Directed Evolution and Engineering of CRISPR-Associated Nucleases. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Directed Evolution and Engineering of CRISPR-Associated Nucleases

A dissertation presented
by
Kevin Michael Davis
to
The Committee on Higher Degrees in Chemical Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Chemical Biology

Harvard University
Cambridge, Massachusetts
September 2016

Dissertation Advisor: David R. Liu   Kevin Michael Davis

Directed Evolution and Engineering of CRISPR-Associated Nucleases

Abstract

CRISPR-Cas systems provide prokaryotes with a remarkable mechanism for adaptive immunity, and recent efforts to understand these systems have provided tremendous insight into the complex nature of these systems. Despite being highly diverse, all CRISPR-Cas systems are able to mediate genomic integration of DNA fragments, sequence-specific RNA processing, and programmable DNA cleavage. Excitingly, each of these processes has numerous uses in biotechnology, medicine, and biological research, and significant effort in the last five years has focused on repurposing CRISPR-Cas systems for applications in these areas. While many Cas-proteins have been successfully applied outside of their native context, protein engineering and directed evolution can be used to enhance their function or optimize their activity for a specific application. This thesis presents our work on engineering two CRISPR-associated nucleases: Csy4, a sequence-specific endoribonuclease, and Cas9, an RNA-guided endonuclease.

We report the development of a phage-assisted continuous evolution (PACE) system for Csy4 that enables the directed evolution of Csy4 variants with optimized activity or altered specificity. We demonstrate that this system can be used to evolve Csy4 variants with improved cleavage rates and to evolve Csy4 variants with orthogonal RNA-binding specificity that are more specific than wild-type Csy4. We also report the development of intein-Cas9 nucleases that are activated by the presence of a cell-permeable small molecule. We demonstrate that these intein-Cas9s enable small-molecule control of Cas9 activity and have improved genome-editing specificity compared to wild type Cas9.

To those who cared.

Acknowledgments

The work presented here was performed under the guidance and support of David Liu, and I am grateful to him for giving me the opportunity to undertake graduate research in his lab. David has created a resourceful, unique, and highly successful research environment that I was fortunate to be a part of. David was an incredibly supportive advisor, and I thank him for his time and mentorship.

Throughout my years in the Liu lab, I have received mentorship and support from numerous lab members. Ralph Kleiner served as my rotation mentor and introduced me to the lab, and I am grateful for his mentorship during this period. Aaron Leconte served as my PACE mentor, and I thank him for teaching me numerous molecular biology techniques. Jacob Carlson provided a tremendous amount of PACE assistance and helped me design and interpret many of my early experiments. I am truly grateful for his mentorship and support. Ahmed Badran provided endless support throughout my graduate degree, and I thank him for this and for his friendship over the years. Vikram Pattanayak provided invaluable assistance with the Cas9 project, and I thank him for his contributions to this project and for general support throughout my time in lab. I thank Mitchel Cole, a talented undergraduate researcher, for his assistance on the Csy4 project.

I have had the pleasure of sharing a lab bay with Grace Chen, Margie Li, and Dmitry Usanov, and I thank them for their friendship and support throughout my graduate degree. I also thank Chihui An, Brent Dorr, John Guilinger, Ryan Hili, Johnny Hu, Bill Kim, Juan Pablo Maianti, Tim Roth, David Thompson, and John Zuris for their assistance and friendship. Aleks Markovic works tirelessly to maintain the Liu Lab, and I thank him for all he has done for me.

Thanks are also due to the remaining Liu lab members who have undoubtedly provided assistance or helpful suggestions, or contributed to making the experience more enjoyable.

The Csy4 project is a collaboration with Jennifer Doudna's lab, and I thank her for her guidance and support. Samuel Sternberg and Mitchell O'Connell have been great collaborators, and I thank them for their assistance with the project.

I am indebted to my Preliminary Qualifying Examination and Dissertation Advisory Committee members who have provided valuable wisdom over the years. I thank Dan Kahne, George Whitesides, Nathanael Gray, and Jack Szostak for serving in one or both of these roles. I am truly grateful for their encouragement and insight. I additionally thank Dan and George for taking time from their busy schedules, on several occasions, to speak with me and provide a helpful perspective. I thank Suzanne Walker for serving on my Dissertation Defense Committee.

As a graduate student, I have been privileged to be part of the Harvard Chemical Biology Program, and I will miss the community of students and faculty that this program brings together. I thank Jason Millberg, Samantha Reed, and KeyAnna Schmiedl for their roles in managing this exceptional program and for their assistance over the years. Ahmed Badran, Sixun Chen, Erin May, and Stephanie Wang, have been awesome Chemical Biology classmates, and I thank them for their friendship and support.

My graduate degree has been supported by the Natural Sciences and Engineering Council of Canada (NSERC), through a PGS-M and PGS-D scholarship, and the Harvard Chemical Biology Program. I am most grateful for this financial support.

Lastly, I thank my family for the tremendous support that they have provided throughout my graduate degree, and my wife, Annie Krylova, for her unconditional love, support, and encouragement during this time.

Table of Contents

List of Figures

List of Tables

Abbreviations

4-HT	4-Hydroxytamoxifen
AP	Accessory plasmid
Cas	CRISPR associated
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
crRNP	CRISPR ribonucleoprotein
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
EMSA	Electrophoretic mobility shift assay
FBS	Fetal bovine serum
gRNA	Guide RNA
HDR	Homology-directed repair
HEK	Human embryonic kidney
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HTS	High-throughput sequencing
indel	Insertion or deletion mutation
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Infrared
MP	Mutagenesis plasmid
NHEJ	Nonhomologous end-joining
NSP	Negative selection plasmid
OD	Optical density
pIII	Protein III (M13 coat protein required for infectivity)
PACE	Phage-assisted continuous evolution
PAM	Protospacer-adjacent motif
PCR	Polymerase chain reaction
PDB	Protein data bank
PES	Polyethersulfone

PFU	Plaque forming units
PMSF	Phenylmethylsulfonyl fluoride
pre-crRNA	Precursor crRNA
RBS	Ribosome binding site
REP	RNA expression plasmid
RNA	Ribonucleic acid
sgRNA	Single guide RNA
SP	Selection phage
TALE	Transcription activator-like effector
TALEN	Transcription activator-like effector nuclease
TCEP	Tris(2-carboxyethyl)phosphine hydrochloride
tracrRNA	*trans*-activating CRISPR RNA
ZF	Zinc finger
ZFN	Zing finger nuclease

Chapter One:

Protein Engineering, Directed Evolution, and CRISPR-Cas Systems

1.1 Introduction Part I: Protein Engineering and Directed Evolution

Proteins are the workhorses of the cell, playing an integral role in nearly all functions required for cellular life. Despite being primarily composed of the 20 standard amino acids, proteins are highly diverse structurally and functionally, allowing them to perform functions ranging from catalysis to intercellular and intracellular signaling to structural support. Outside of their native context, proteins are applied industrially and therapeutically, and are critical to biological research. Owing to the central role that proteins play in biology and biotechnology, there is great interest in developing proteins with new functions that can be applied to solve challenging problems or used to further our understanding of biological processes.

Protein engineering is the generation of new proteins through the modification of existing proteins or *de novo* protein design. Protein engineering can be used to optimize or alter existing protein function, or to create proteins with entirely novel functionality. There are two general strategies for protein engineering: rational protein design and directed evolution. Importantly, these strategies are not mutually exclusive, and are routinely applied together.

In rational protein design, detailed structural information and mechanistic knowledge of protein function guide the design of engineered proteins. This approach can be applied at high resolution to introduce specific mutations into a protein through site-directed mutagenesis or at lower resolution to combine functional domains from several proteins. Rational protein design is increasingly aided by computational protein design algorithms that can predict amino acid sequences that will fold to a targeted protein structure. Despite numerous functional examples of rationally designed proteins, this method is still limited by our inability to accurately predict the structure and dynamics of proteins. It is likely that rational protein design will become an increasingly powerful technique as our understanding of protein folding and dynamics improves.

In directed evolution, proteins are subjected to iterative rounds of diversification and selection in a process mimicking Darwinian evolution. The directed evolution of proteins involves four general steps: diversification, translation, selection or screening, and replication (Figure 1.1). Diversification at the genetic level is introduced through mutations, which can be added in an unbiased manner throughout the entire gene or focused to specific regions in a semi-rational approach. Homologous recombination can also be used to increase genetic diversity. During translation, proteins must remain linked to their encoding gene through molecular linkage or spatial segregation. Improved protein variants are identified in selections by directly selecting for a specific function or activity, or in screens that analyze the phenotype of each protein variant. Following selection or screening, genes encoding desired variants are amplified and can be subjected to additional rounds of evolution. Unlike rational protein design, directed evolution does not require structural information and mechanistic knowledge of the protein; however, structural and mechanistic insights can aid the diversification and selection or screening steps.

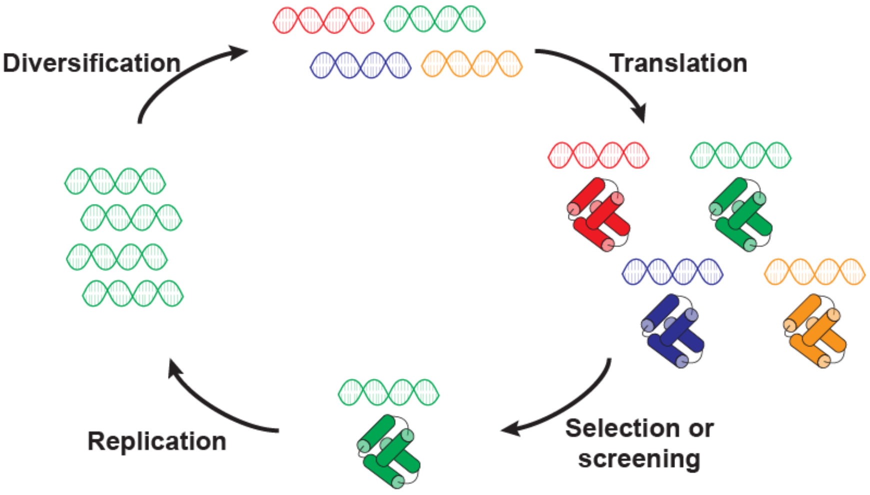


Figure 1.1 | Key steps in the directed evolution of proteins. Protein directed evolution begins with diversification at the genetic level. Following translation, proteins with desired properties are identified through selection or screening. Genes encoding the identified proteins are replicated and can be subjected to additional rounds of evolution.

1.2 Introduction Part II: CRISPR-Cas Adaptive Immunity in Prokaryotes

Bacteria and archaea are the most abundant organisms, inhabiting a diverse array of environments on the planet. However, these prokaryotic organisms are greatly outnumbered by prokaryotic viruses, necessitating the need for antiviral defense systems. Because they generally confer a selective advantage on the host organism, numerous and diverse defense mechanisms have evolved over time. The clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes encode a unique defense system capable of adapting to rapidly evolving viruses. Among prokaryotic defense mechanisms, only the CRISPR-Cas system is known to provide adaptive immunity against foreign genetic material. The defining feature of the CRISPR-Cas system is the CRISPR locus, which contains an array of direct repeats (approximately 20-50 base pairs and often partially palindromic) interspaced by unique spacer sequences (of similar length) of viral or plasmid origin[1]. These spacer sequences specify immunity against foreign genetic elements that contain a complementary sequence, known as the protospacer. While CRISPR-Cas systems can vary in the sequence and length of the CRISPR repeat[2], the main diversity in CRISPR-Cas systems stems from the variable cassette of *cas* genes located adjacent to the CRISPR array. Despite the extremely diverse set of *cas* genes that exist[3], CRISPR-Cas adaptive immunity involves three general stages: spacer acquisition, RNA biogenesis, and target interference (Figure 1.2).

In order to adapt to new or evolving viral threats, novel spacers must be added to the CRISPR array. Spacer acquisition can be divided into two stages: sampling and selection of new spacer sequences from foreign DNA and their integration into the CRISPR locus. Potential spacer sequences are likely generated from non-specific DNA breaks that occur during the replication of foreign DNA[4]. How chromosomally derived sequences (self) are avoided in this

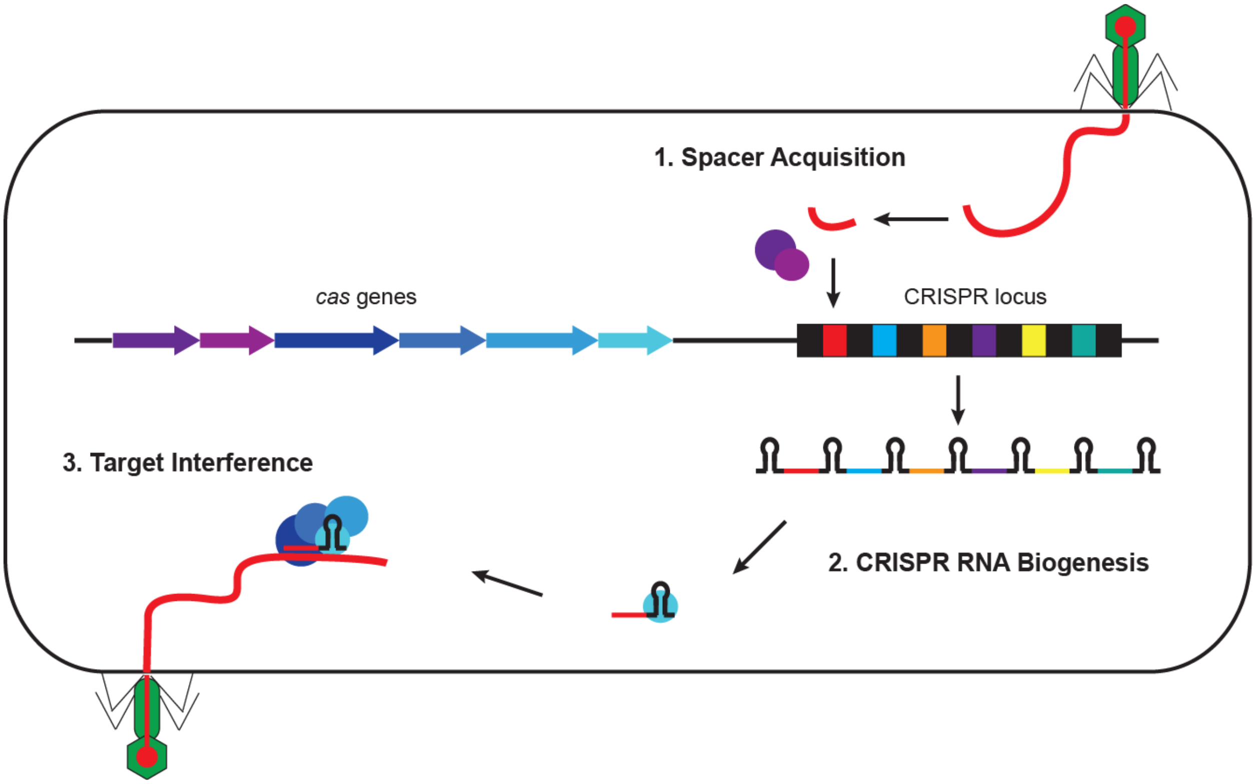


Figure 1.2 | CRISPR-Cas adaptive immunity in prokaryotes. CRISPR-Cas immunity begins with the acquisition of spacer sequences from foreign DNA, which are integrated into the CRISPR locus. During CRISPR RNA biogenesis, the CRISPR array is transcribed and processed into individual crRNAs. Upon subsequent exposure, crRNAs guide Cas nucleases to destroy the target.

process remains unclear, however, the higher copy number of viral and plasmid DNA relative to chromosomal DNA likely plays a role. In most CRISPR-Cas systems, potential protospacers must also be flanked by a protospacer-adjacent motif (PAM)[5], a short sequence (3-7 base pairs) required in the interference stage, and the mechanism of how this requirement is ensured is still not fully understood. The integration of suitable new spacers is catalyzed by Cas1 and Cas2[6-8], the only universally conserved Cas proteins. Integration occurs at the 5′ end of the CRISPR array and requires the simultaneous synthesis of an additional repeat[9,10].

While spacer acquisition enables the CRISPR-Cas system to adapt to invasive genetic elements, successful defense requires the generation of small CRISPR RNAs (crRNAs) that consist of a single spacer. In the first stage of crRNA biogenesis, the CRISPR array is transcribed as a single long precursor crRNA (pre-crRNA) from a promoter sequence upstream of the CRISPR locus. In the second stage, the pre-crRNA is processed, at each of the CRISPR repeat sequences, into mature individual crRNAs. In many CRISPR-Cas systems, pre-crRNA processing is carried out by a sequence-specific Cas endoribonuclease that recognizes and cleaves the CRISPR repeat. In the remaining CRISPR-Cas systems, an additional small *trans*-activating crRNA (tracrRNA) is required for pre-crRNA processing[11]. The tracrRNAs contain a region of complementarity to the CRISPR repeat allowing them to hybridize to the pre-crRNA. The resulting regions of dsRNA are cleaved by endogenous RNase III to yield mature individual crRNAs.

To target foreign DNA for destruction during the interference stage, crRNAs associate with Cas proteins to form large CRISPR ribonucleoprotein (crRNP) effector complexes that possess nuclease activity. The spacer portion of the crRNA guides the crRNP complex to complementary sequences present in the foreign DNA. The nuclease activity of the crRNP complex is activated upon hybridization of the crRNA to the target, leading to double-stranded cleavage of the foreign genetic material. In most CRISPR-Cas systems, the presence of the PAM on the foreign DNA is essential for full binding and activation of nuclease activity. The absence of PAMs in the CRISPR locus prevents self-targeting by the crRNP complex. While the majority of CRISPR-Cas systems target DNA, several examples of RNA-targeting systems have been discovered[12,13].

CRISPR-Cas systems are highly diverse and close to 50% of analyzed bacterial and archaea genomes encode CRISPR-Cas loci[3]. Often times, more than one CRISPR-Cas system can be found in a single organism[3]. The classification of CRISPR-Cas systems continues to evolve as new CRISPR-Cas systems are discovered and our understanding of CRISPR-Cas biology improves. CRISPR-Cas systems were originally classified into three types (I, II, and III)[3]; however, an updated classification system[14] has introduced a broader level of classification, classes (1 and 2), and added additional putative types (IV and V). Class 1 CRISPR-Cas systems are defined by a multi-subunit crRNP effector complex and the class includes type I, III, and IV CRISPR-Cas systems. Class 2 CRISPR-Cas systems are defined by the presence of a single Cas protein in the crRNP effector complex and the class includes type II and V CRISPR-Cas systems.

In their native context, CRISPR-Cas systems offer the ability to generate phage-resistant strains of bacteria, something of particular use for industrially applied microbes. More exciting, perhaps, is the vast and diverse array of Cas proteins that can be repurposed for alternative applications. CRISPR-Cas systems integrate DNA fragments into genetic loci, process RNA in a sequence-specific manner, and direct targeted double-stranded DNA cleavage, biochemical activities with numerous applications in biological research, biotechnology, and therapeutic development. Indeed, the field of genome engineering has progressed rapidly in recent years due to the programmable RNA-guided endonuclease Cas9. As CRISPR-Cas systems continue to be discovered and understood, it is likely that additional applications and uses for Cas proteins will be uncovered.

1.3 Thesis Overview

This thesis presents our work on engineering two CRISPR-associated nucleases: Csy4, a sequence-specific endoribonuclease, and Cas9, an RNA-guided endonuclease. In Chapter Two, we report the development of a phage-assisted continuous evolution (PACE) system for evolving Csy4, and we demonstrate that this system can be used to evolve Csy4 variants with improved cleavage rates. In Chapter Three, we report the use of the Csy4 PACE system to evolve Csy4 variants with orthogonal RNA-binding specificity. In Chapter Four, we report the development of a small molecule-triggered Cas9 protein that demonstrates improved genome-editing specificity.

1.4 References

1. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181-190 (2010).

2. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**, R61 (2007).

3. Makarova, K.S. et al. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477 (2011).

4. Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510 (2015).

5. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740 (2009).

6. Yosef, I., Goren, M.G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research* **40**, 5569-5576 (2012).

7. Nunez, J.K. et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528-534 (2014).

8. Nunez, J.K., Lee, A.S., Engelman, A. & Doudna, J.A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198 (2015).

9. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

10. Garneau, J.E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).

11. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011).

12. Hale, C.R. et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956 (2009).

13. Abudayyeh, O.O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

14. Makarova, K.S. et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736 (2015).

Chapter Two:

Phage-Assisted Continuous Evolution of Csy4

Kevin M. Davis, Mitchell R. O'Connell, Samuel H. Sternberg,
Jennifer A. Doudna, and David R. Liu

Kevin Davis generated all constructs, developed the Csy4 PACE system, performed the PACE experiments, and purified the Csy4 variants. Mitchell O'Connell and Samuel Sternberg performed the *in vitro* RNA cleavage assays.

2.1 Introduction Part I: Phage-Assisted Continuous Evolution (PACE)

Phage-assisted continuous evolution is a method that enables the continuous directed evolution of proteins by harnessing the life cycle of the M13 bacteriophage[1]. M13 is a non-lytic filamentous bacteriophage, and the M13 genome encodes 11 genes required for the replication and assembly of progeny. In PACE, selection phage (SP) are engineered by replacing a phage gene, gene III, with the gene to be evolved (Figure 2.1). Gene III encodes protein III (pIII), a phage coat protein that is required for infectivity, and, thus, these SPs are unable to propagate. If, however, pIII is supplied, in *trans*, by the host bacterial cell, these SPs are capable of propagation. By placing gene III on an accessory plasmid (AP) that conditionally expresses pIII in response to a desired activity, SPs encoding proteins with this activity are selected through their ability to propagate.

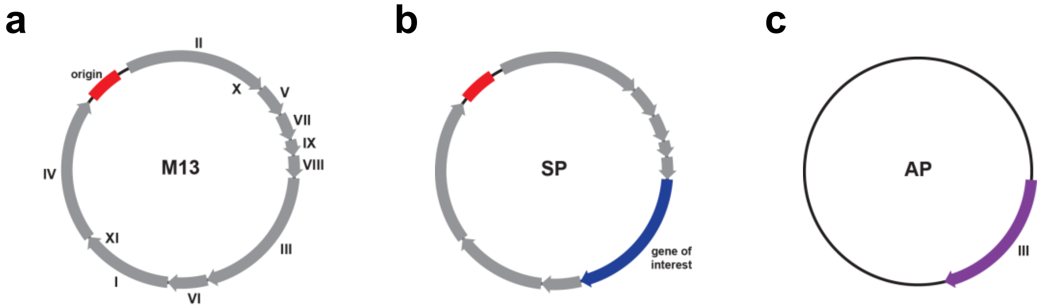


Figure 2.1 | M13 bacteriophage genome and engineered PACE vectors. (a) Schematic of the M13 genome, illustrating the genomic organization of the 11 phage genes required for replication and assembly of progeny. (b) Selection phage (SP) encode the gene of interest in place of gene III. (c) The accessory plasmid (AP) encodes gene III and any regulatory elements required for conditional pIII expression.

In PACE, an actively replicating population of SPs is maintained in a fixed-volume vessel (the “lagoon”) through which *E. coli* host cells continuously flow (Figure 2.2). These host cells contain the AP and any regulatory components required for the conditional expression of pIII. Importantly, the output of infectious progeny directly correlates with increasing levels of pIII, giving activities that enable higher levels of pIII production a selective advantage. To increase genetic diversity, host cells also contain an arabinose-inducible mutagenesis plasmid (MP) that elevates the error rate during SP DNA replication[1-3]. Since host cells are being continuously pumped into the lagoon, SPs must propagate at a sufficient rate to persist in the continuously diluting lagoon. Adjustments to this flow rate can be used to modulate selection stringency, and the use of multiple lagoons enables evolution experiments to be performed in parallel. In principle, any protein activity that can be linked to pIII production can be evolved in PACE.

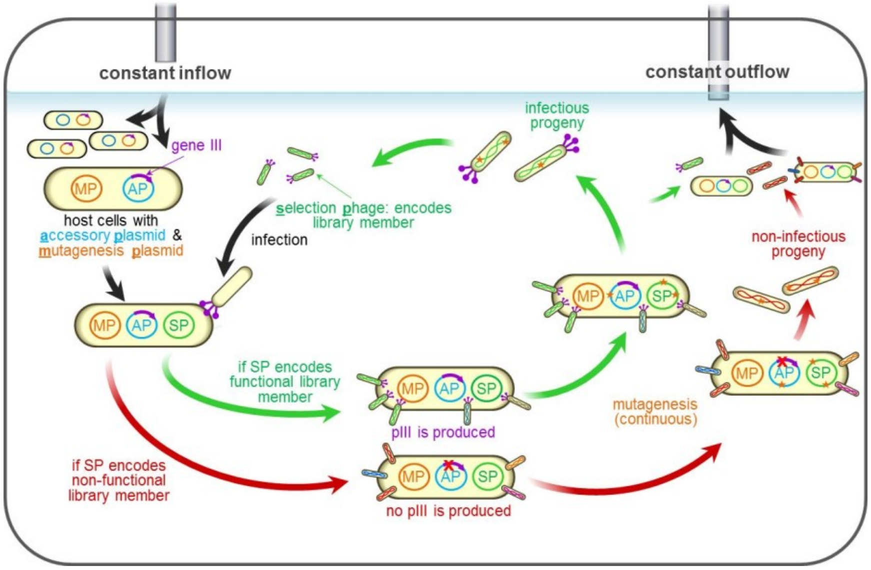


Figure 2.2 | Overview of the PACE system. Host cells continuously flow through the lagoon, where they are infected by selection phage (SP) encoding the protein of interest. Functional proteins induce the production of pIII from the accessory plasmid (AP) and allow for the production of progeny capable of infecting new cells. Increased mutation rates are triggered through the induction of a mutagenesis plasmid (MP).

2.2 Introduction Part II: The Sequence-Specific CRISPR Endoribonuclease Csy4

Csy4 is a sequence-specific endoribonuclease isolated from the type I CRISPR-Cas system of *P. aeruginosa* UCBPP-PA14 that is responsible for processing the CRISPR array transcript into mature crRNAs during crRNA biogenesis[4]. It recognizes a five-base stem-loop present in the CRISPR repeat and cleaves immediately 3′ of the stem-loop (Figure 2.3). Csy4 has a high affinity for both its substrate and product, binding both with a 50 pM equilibrium dissociation constant[5]. Due to product inhibition, Csy4 acts as a single-turnover enzyme. Csy4-mediated RNA cleavage does not require a divalent metal ion[4], and based on structural and biochemical characterization, two active site-proximal residues, His29 and Ser148, have been implicated in substrate positioning and cleavage[6]. Ser148 orients the stem-loop in a cleavage-compatible conformation, while His29 functions as a general base. Mutating His29 to alanine renders Csy4 catalytically inactive. Csy4 is highly selective for its cognate substrate and relies on base-specific hydrogen-bonds between two residues, Arg102 and Gln104, and the major groove faces of G20 and A19, respectively[4].

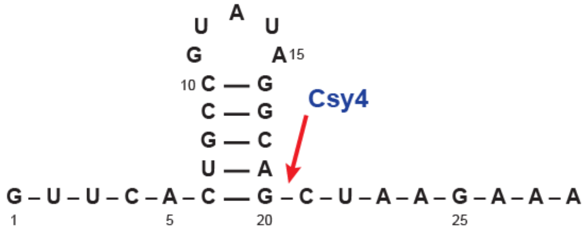


Figure 2.3 | Sequence-specific cleavage by Csy4. Csy4 recognizes and cleaves 3' of the stem-loop present in the CRISPR repeat. The full 28 base CRISPR repeat is shown and the Csy4 cleavage site is indicated.

Sequence-specific endoribonucleases like Csy4 enhance our ability to manipulate RNA structure and function *in vitro* and *in vivo*. Outside of its native context, Csy4 has been used to process RNA transcripts *in vivo*[7-9] and, in its catalytically inactive form, to pull down tagged RNAs[10]. Evolving Csy4 variants with orthogonal specificity would expand the utility of these techniques and enable multiplex applications.

2.3 Linking RNA Cleavage to Gene Expression

Csy4 PACE requires a selection scheme that links RNA cleavage to increased pIII expression. To accomplish this, we focused on developing a translational derepression selection scheme in which targeted RNA cleavage relieves a translationally repressed gene III transcript. This strategy relies on sequestering the ribosome binding site (RBS) in an RNA secondary structure, inhibiting ribosomal access (Figure 2.4). To enable Csy4-mediated derepression, the Csy4 recognition element must be engineered into the repressive structure. To test and validate this translational derepression strategy we used a luciferase assay, in which host cells carry a modified AP containing a luciferase gene in place of gene III, allowing us to quantify relative changes in gene expression by monitoring luminescence. Transcription of the repressed transcript is regulated by the phage-shock promoter, which can be induced by phage infection, and a strong SD8[11] RBS enables translation initiation. In this assay, host cells also carry a plasmid encoding an arabinose-inducible Csy4 gene.

We compared several repressive secondary structures that varied in the length of complementarity to the RBS region by manipulating the length of overlap upstream and downstream of the RBS (Figure 2.5). To increase the length of overlap upstream of the RBS, we inserted a string of cytosine bases directly upstream of the RBS and inserted a corresponding

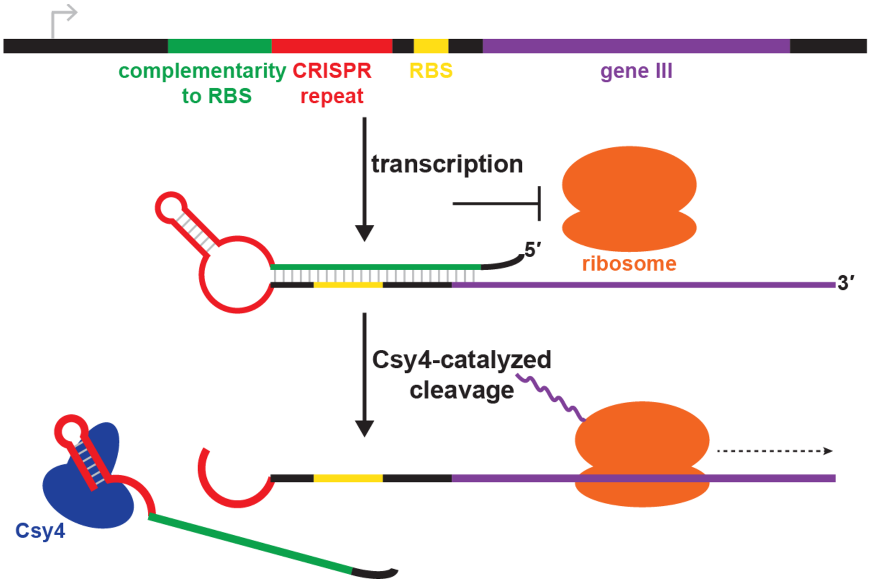


Figure 2.4 | Selection scheme for linking Csy4 activity to pIII production. The gene III transcript is translationally repressed by an RNA secondary structure that sequesters the ribosome binding site (RBS) from the ribosome. Csy4-mediated cleavage leads to derepression, allowing for translation initiation and pIII expression.

number of guanosine bases in the complementary segment. To increase the length of overlap downstream of the RBS, we simply extended the length of the complementary segment to include the start codon. Several positive hits were identified in the luciferase assay. We moved forward with the top candidate from the luciferase assay (Figure 2.6), which demonstrated a 41-fold increase in luminescence with Csy4 induction. Surprisingly, switching the cytosine bases upstream of the RBS to guanosines and making the corresponding guanosine to cytosine change to the complementary segment led to a non-functional AP. We also tested whether truncating the CRISPR repeat would improve the responsiveness of this AP to Csy4 induction, but found that this actually had the opposite effect. To allow for the RNA substrate to be present upon phage infection, we swapped the phage-shock responsive promoter with a proD[12] constitutive promoter. Importantly, inducing the catalytically inactive variant of Csy4 (H29A-Csy4) does not lead to an increase in luminescence (Figure 2.7).

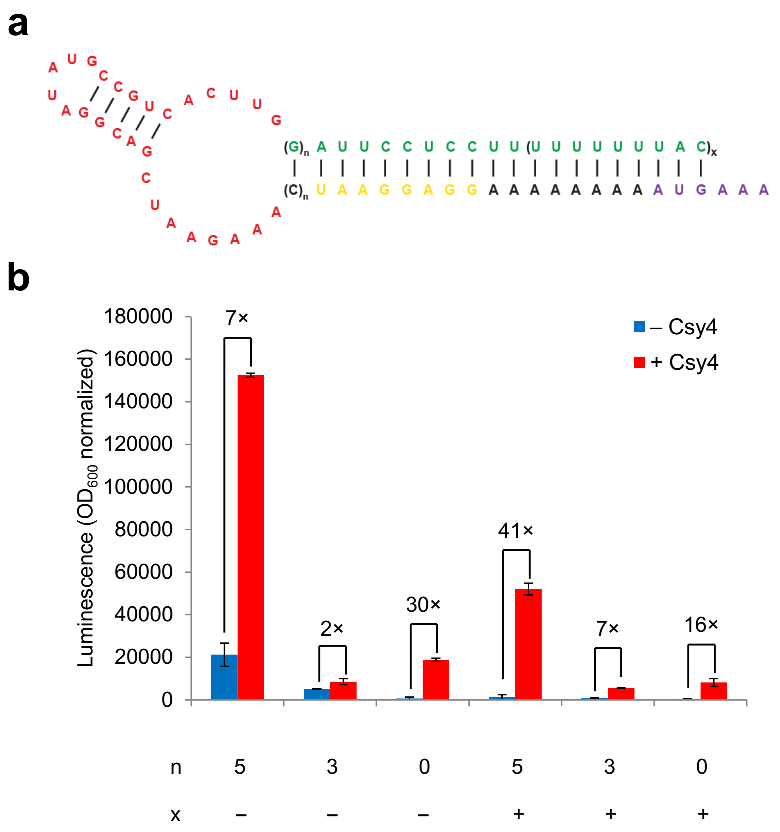


Figure 2.5 | Candidate translational repression secondary structures. (a) General structure of the repressive secondary structures investigated. The CRISPR repeat, the RBS sequence, the start codon and second codon of gene III, and the complementary region are colored red, yellow, purple, and green, respectively. A variable number of G–C base pairs were encoded upstream of the RBS, and two overlap lengths downstream of the RBS were tested. (b) Luciferase assay results for 6 candidate repressive secondary structures. n and x are defined in (a). Fold increases in luminescence with Csy4 induction are indicated.

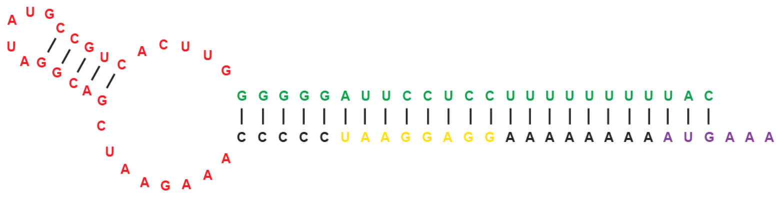

G G G G G A U U C C U C C U U U U U U U U U A C
C C C C C U A A G G A G G A A A A A A A A A U G A A A
A U G C C G U C A C U U G
U A G G C A G C U A A G A A A

Figure 2.6 | Top translational repression secondary structure. The repressive secondary structure demonstrating the largest fold-increase (~ 40×) in the luciferase assay was used in subsequent experiments.

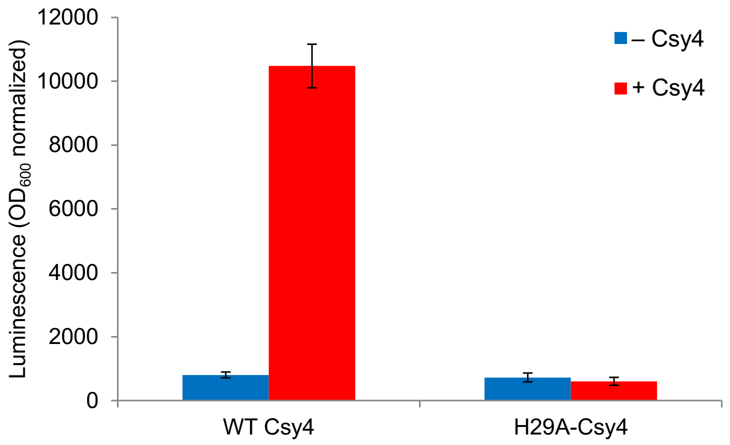


Figure 2.7 | Luciferase assay with H29A-Csy4. Catalytically inactive Csy4 (H29A) does not lead to an increase in luminescence upon induction.

2.4 Linking RNA Cleavage to SP Propagation

With the translational derepression system demonstrating Csy4-dependent gene expression in luciferase assays, we moved towards validating that this selection system could enable Csy4-SP propagation in an activity-dependent manner. We cloned the corresponding PACE AP by adding gene III downstream of the RBS. We kept the luciferase gene on the AP, translationally coupled to gene III, allowing relative gene expression levels to still be quantified by monitoring luminescence. Csy4 was cloned onto a modified SP vector (Figure 2.8), which had previously been used in T7 RNA polymerase PACE experiments[13]. Csy4 was placed downstream of the kanamycin resistance gene (KanR) with an SD8 RBS. The mRNA transcript containing Csy4 is initiated from an M13 promoter upstream of KanR.

To visualize actively replicating phage, we used plaque assays, which involve mixing a small number of infectious phage with a large excess of host cells, and plating the mixture in a semisolid medium. Infected host cells produce phage progeny that spread to neighboring cells in

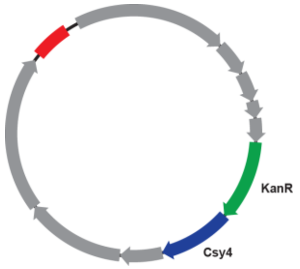


Figure 2.8 | Csy4-SP vector. Csy4 was cloned onto a modified SP vector downstream of the kanamycin resistance gene (KanR).

the medium. Because phage-infected host cells grow slower than uninfected cells, turbid plaques form on the plate, with each plaque representing a single initial infectious phage. Using plaque assays, we validated that Csy4-SPs could propagate on host cells containing the AP. Importantly, Csy4-SP propagation is activity-dependent since Csy4-SPs do not propagate on host cells lacking the AP, and SPs lacking Csy4 do not propagate on host cells containing the AP.

2.5 Phage-Assisted Continuous Evolution of Csy4

To validate Csy4 PACE, we tested the propagation of Csy4-SPs in the continuous liquid culture format required for PACE. For this initial PACE demonstration, host cells carried both the AP and MP, and we maintained 2 lagoons, one of which had mutagenesis induced. Phage propagation in both lagoons was robust over 96 hours of PACE even as the lagoon flow rate was increased stepwise from 1 to 4 lagoon volumes per hour (Figure 2.9). In plaque assays with phage isolated at the end of the experiment, the plaques appeared larger, suggesting that these phage were able to propagate more efficiently. Sequencing the Csy4 gene of individual Csy4-SPs at the 48 and 96 hour time points (Table 2.1) revealed that the phage populations in both lagoons converged on a Q30L mutation.

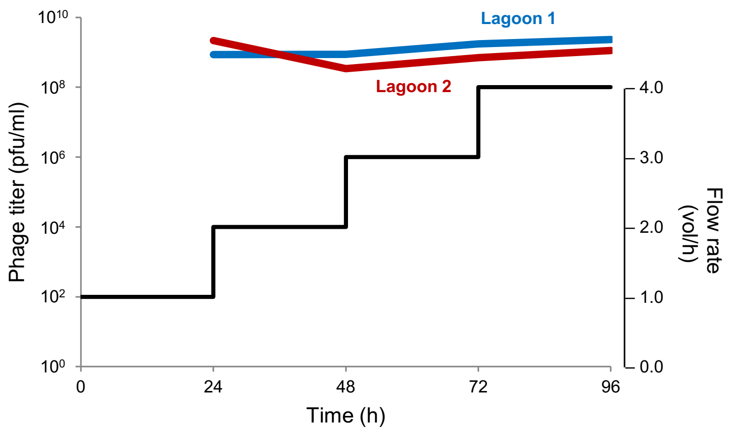


Figure 2.9 | Csy4 PACE demonstration. Each lagoon was inoculated with 10^6 Csy4-SPs, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP1 in lagoon 2.

Table 2.1 | Csy4 PACE sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue			
		17	30	34	82
WT		P	Q	A	H
L1	48 h				
	96 h		L		
			L		
			L		
			L		
			L		
			L		
			L		
			L		
L2	48 h		L		
			L		
			L	T	
			L		
		S	L		
		S	L		
			L		
	96 h		L		
			L		Y
			L		Y
			L		Y
			L		
			L		
			L		Y
			L		Y

2.6 Improving Csy4 Activity in PACE

We were interested to see if we could further optimize Csy4-SP propagation and, ideally, Csy4 activity by performing PACE experiments with increased selection pressure. To increase the selection pressure we undertook two general strategies: reducing Csy4 expression from the SP and reducing the amount of pIII expressed per Csy4 cleavage event. To reduce Csy4 expression from the SP, we removed the strong RBS upstream of the Csy4 gene and translationally coupled it to the preceding gene, KanR. To reduce the amount of pIII expressed per Csy4 cleavage event, we developed an RNA expression plasmid (REP) that produces an RNA transcript analogous to the one produced from the AP, except lacking gene III downstream of the repressive RNA secondary structure (Figure 2.10). By expressing an excess of this non pIII-producing transcript relative to the gene III transcript, multiple Csy4-mediated cleavage events are required to derepress a single gene III transcript.

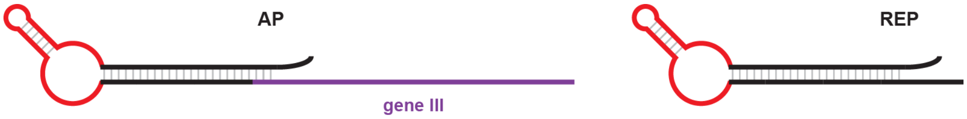


Figure 2.10 | Increasing selection pressure using a non pIII-producing substrate. The RNA expression plasmid (REP) expresses an RNA transcript analogous to the one produced from the AP, except lacking gene III. This transcript serves as a substrate for Csy4 but does not lead to pIII production.

Incorporating both strategies, we performed two successive PACE experiments with increased selection pressure. In the first experiment (medium stringency), we used an REP with a medium-copy number origin of replication, and in the second experiment (high stringency), we used an REP with a high-copy number origin of replication. Both PACE experiments were performed with mutagenesis induction in duplicate lagoons. Phage propagation in both experiments was robust over 96 hours (Figure 2.11 and Figure 2.12). The phage isolated at the end of the high stringency PACE experiment yielded even larger plaques in plaque assays, relative to the phage isolated after the initial PACE demonstration, suggesting the evolution of further improvements related to phage propagation. Sequencing the Csy4 gene of individual Csy4-SPs at the 48 and 96 hour time points of each PACE experiment (Table 2.2 and Table 2.3) revealed that that the phage populations in each lagoon converged on several consensus mutations.

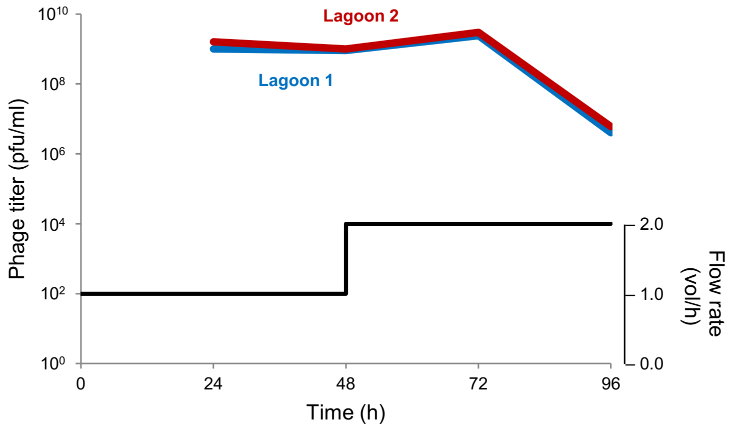


Figure 2.11 | Csy4 medium stringency PACE. Each lagoon was inoculated with 10^6 phage from the initial Csy4 PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

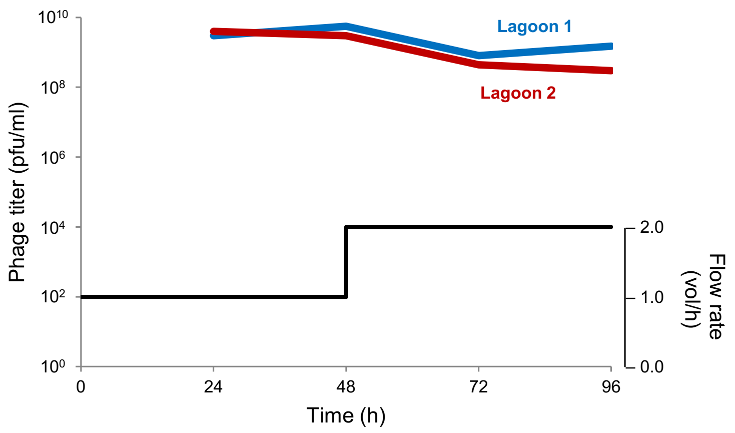


Figure 2.12 | Csy4 high stringency PACE. Each lagoon was inoculated with 10^6 phage from the medium stringency PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

Table 2.2 | Csy4 medium stringency sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

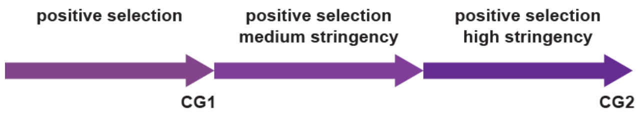
		Csy4 Residue											
		22	30	35	36	68	81	85	133	136	140	154	166
WT		S	Q	Q	G	R	D	F	D	A	D	H	V
L1	48 h		L							T			
			L										
		N	L										
			L							T			
			L										
			L							T			
			L										
			L										
	96 h		L	H									
			L										
		N	L										
			L									N	
			L						N				
		N	L										
		N	L										
		N	L		D			I					M
L2	48 h		L										
			L										
			L										
			L										
			L										
			L										
			L										
			L										
	96 h	N	L										
			L								N	N	
			L										
			L										
			L			H							E
			L				Y				A		
			L										
			L										

Table 2.3 | Csy4 high stringency sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																				
		12	16	18	22	27	30	50	73	85	96	117	118	120	121	130	133	135	140	147	154	160
WT		D	P	A	S	K	Q	S	R	F	P	M	R	H	D	R	D	V	D	R	H	R
L1	48 h					Q	L				L							G		L		
							L								G		Y				N	
						Q	L				L							G		L		
						R	L			I							Y		G			
				S		Q	L				L		H					G		L		
						Q	L										Y					
						R	L			I			C				Y					
						R	L										Y					
	96 h	Y	S		N	Q	L				L				G			G		L		
		Y	S		N	Q	L		Q		L				G			G		L		
		Y	S		N	Q	L				L	I			G			G		L		
		Y	S		N	Q	L				L				G			G		L		
		Y	S		N	Q	L				L				G			G		L		
		Y	S		N	Q	L				L				G			G		L		
		Y	S		N	R	L				L						A	G		L		
		Y	S		N	Q	L				L					H		G		L		
L2	48 h				N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
					N	R	L										Y					H
	96 h				N	R	L										Y				N	H
					N	R	L										Y				N	H
					N	R	L										Y				N	H
		Y	S		N	R	L							Y					G			
					N	R	L	R									Y					H
					N	R	L										Y					H
					N	R	L										Y				N	H
					N	R	L										Y				N	H

2.7 *In Vitro* Characterization of Optimized Csy4 Variants

Mutations that enhance the ability of Csy4-SPs to propagate faster or produce more progeny gain a selective advantage in PACE. In the ideal case, these propagation-enhancing mutations correspond to improvements in Csy4 activity. Due to the complex nature of PACE and the Csy4 PACE selection system, however, mutations in the phage genes on the SP or mutations that effect Csy4 expression and stability can also lead to improvements in phage propagation. To determine if any of the mutant Csy4 variants demonstrated improved activity, we isolated and purified representative Csy4 variants from the initial PACE demonstration and the high stringency PACE experiment (Figure 2.13) and assayed their activity *in vitro*. In RNA cleavage assays, both variant CG1, the Q30L mutant, and variant CG2, the consensus genotype from lagoon 2 of the high stringency PACE experiment, demonstrate an improvement in the rate of RNA cleavage (Figure 2.14). These results indicate that the Q30L mutation enhances the rate of substrate cleavage by approximately 2.5-fold relative to wild-type Csy4. It is interesting to note that this mutation is adjacent to the catalytic His29 residue. Since variant CG2 does not show any further improvements over variant CG1, it is likely that the additional mutations in this variant do not directly impact the rate of cleavage.


positive selection
positive selection medium stringency
positive selection high stringency
CG1
CG2

	Csy4 Residue					
	22	27	30	133	154	160
WT	S	K	Q	D	H	R
CG1			L			
CG2	N	R	L	Y	N	H

Figure 2.13 | Summary of the Csy4 PACE experiments. Representative variants (CG1, CG2) from the initial PACE experiment and the high stringency PACE experiment were purified for *in vitro* characterization.

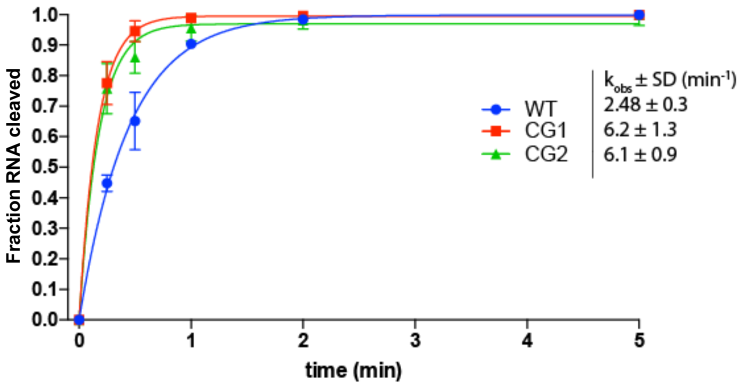


Figure 2.14 | RNA cleavage assays with the evolved Csy4 variants. Single-turnover RNA cleavage assays were performed with wild-type Csy4, variant CG1, and variant CG2.

2.8 Conclusion

The Csy4 PACE system developed here enables the continuous evolution of Csy4, and, in principle, it can be modified to facilitate the continuous evolution of other sequence-specific RNA cleaving enzymes. In this system, Csy4 activity mediates the production of pIII through derepression of the translationally repressed gene III transcript. Using this Csy4 PACE system, we evolved a Csy4 variant that cleaves the wild-type substrate 2.5-fold faster than wild-type Csy4.

2.9 Methods

Bacterial strains. DNA cloning was performed with NEB Turbo cells (New England Biolabs) or Mach1 cells (Thermo Fisher Scientific). Protein expression was performed with BL21 Star (DE3) cells (Thermo Fisher Scientific). All luciferase assays, plaque assays, and PACE experiments were performed with S1030 cells[13]. Electrocompetent S1030 cells were prepared as previously described[14]. Briefly, cells were grown to an OD_{600} value of 0.8, pelleted by centrifugation at 10,000g, and the supernatant was decanted. The cells were resuspended in chilled 10% glycerol and pelleted by centrifugation at 10,000g with subsequent removal of the supernatant. Washing with 10% glycerol was repeated an additional three times. Cells were flash frozen with liquid nitrogen and stored at –80 °C. Plasmids were transformed into S1030 cells by electroporation using an *E. coli* Pulser Transformation Apparatus (Bio-Rad). Liquid cultures were supplemented with the appropriate antibiotics (Gold Biotechnology) in the following final concentrations: streptomycin (S1030 cells – *rpsL* marker; 50 µg/ml), tetracycline (S1030 cells – F plasmid; 10 µg/ml), carbenicillin (AP, Csy4 expression vector; 50 µg/ml), chloramphenicol (MP, arabinose-inducible Csy4 plasmid; 40 µg/ml), and spectinomycin (REP; 100 µg/ml).

Liquid cultures were incubated in a 37 °C shaker unless otherwise noted. Agar plates were similarly supplemented with the appropriate antibiotics; however, streptomycin and tetracycline were not routinely included in agar plates. Agar plates were incubated 37 °C.

General cloning methods. PCR fragments were generated using PfuTurbo Cx Hotstart DNA polymerase (Agilent Technologies), VeraSeq ULtra DNA polymerase (Enzymatics), or Phusion U Hot Start DNA polymerase (Life Technologies), and appropriate DNA primers containing an internal deoxyuracil base (Integrated DNA Technologies). PCR products were purified using a MinElute PCR Purification Kit (Qiagen). Plasmids and selection phage were constructed by USER cloning as previously described[14]. Briefly, PCR products were mixed in an equimolar ratio, and incubated with DpnI (New England Biolabs) and USER enzyme (New England Biolabs) at 37 °C for 30 min. The assembly reaction was heated to 80 °C and slowly cooled to 20 °C at 0.1 °C/s at which point it was transformed into *E. coli*. Plasmids were transformed into chemically competent NEB Turbo cells or Mach1 cells according to the manufacturer's instructions, and plated on 2xYT-agar (United States Biological) plates. Plasmid DNA was amplified from colonies using an illustra TempliPhi Amplification Kit (GE Healthcare) and sequence verified by Sanger sequencing. Sequence verified colonies were grown overnight in 3 ml 2xYT media (United States Biological) and plasmid DNA was harvested using a QIAprep Miniprep Kit (Qiagen). Selection phages were transformed into S1030 cells containing the phage-responsive accessory plasmid pJC175e[13] by electroporation, and were recovered overnight. Phage were isolated by pelleting the cells at 10,000g and subsequently filtering the supernatant through a 0.2 µm PES syringe filter (Corning). Plaque assays were used to generate clonal phage plaques. Phage DNA was amplified from plaques using an illustra TempliPhi Amplification Kit and sequence verified by Sanger sequencing.

Luciferase assays. S1030 cells carrying the AP and arabinose-inducible Csy4 plasmid were grown overnight in 2xYT media. Following overnight growth, cultures were diluted 500-fold into 500 µl Davis rich media[13] (DRM) in a 96-well deep well plate (Axygen) and incubated. To induce Csy4 expression, 20 µM arabinose (Gold Biotechnology) was added to appropriate wells. For APs carrying the phage-shock response promoter, T7 RNA polymerase SPs were added after 3 h of incubation. Cultures were incubated for a total of 4-5 h, and then 150 µl was transferred to a clear-bottom 96-well plate (Costar). OD_{600} and luminescence for each sample was measured using an Infinite M1000 PRO microplate reader (Tecan). Raw luminescence values for each sample were normalized by the OD_{600} value to adjust for differences in cell density. Luciferase assays were performed in triplicate.

Plaque assays. S1030 cells carrying the AP were grown overnight in 2xYT media. Following overnight growth, cultures were diluted 50-fold into 3 ml 2xYT media and grown to an OD_{600} value of 0.7. Phage were serially diluted 10-fold and 10 µl of each dilution was mixed with 90 µl of cells. The phage/cell mixture was mixed with 1 ml of warm top agar (7g/l agar in 2xYT) and plated onto bottom agar (1.6g/l agar in 2xYT) plates. Plates were grown overnight for 20 h.

PACE. PACE experiments were setup as previously described[13]. S1030 cells carrying the AP, MP, and REP (when appropriate) were tested for arabinose sensitivity as previously described[14]. Cultures were grown in 3ml DRM to an OD_{600} value of 0.5 and transferred to the chemostat, which was subsequently grown until it was visibly turbid. The chemostat culture was maintained at 100 ml and diluted at a rate of approximate 1.5 volumes per hour as previously described. Lagoons were maintained at 30 ml and supplemented with 20mM arabinose to induce mutagenesis. Lagoon flow rates were adjusted as described for each experiment. Lagoon samples were taken every 24 h, and phage were isolated by pelleting the cells at 10,000g and subsequently filtering the supernatant through a 0.2 µm PES syringe filter. Phage titers were determined by plaque assay.

Protein expression and purification. Csy4 variants were cloned into a pHMGWA expression vector[15] with an N-terminal His_6 tag and a $(GGS)_2$ linker. The Csy4 expression vector was transformed into chemically competent BL21 Star (DE3) cells according to the manufacturer's instructions. Colonies were grown in 150 ml 2xYT media to an OD_{600} value of 0.6, at which point protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Gold Biotechnology). Cultures were transferred to an 18 °C shaker and incubated for 16 h. Csy4 variants were purified as previously described[4] except without size exclusion chromatography. Briefly, cells were resuspended in a lysis buffer (15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 10 mM imidazole, 1 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), 0.5mM phenylmethylsulfonyl fluoride (PMSF), 5% glycerol, 0.01% Triton X-100, 100 U/ml DNase I, pH 7.4, supplemented with protease inhibitors (Roche)) and lysed by sonication. The clarified lysate was incubated with HisPur Ni-NTA Resin (Thermo Fisher Scientific) in batch, and the resin was washed with lysis buffer (lacking PMSF, DNase I, and protease inhibitors). Bound protein was eluted with a high imidazole buffer (15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 300 mM imidazole, 1 mM TCEP, 5% glycerol, pH 7.4) and dialyzed overnight in dialysis buffer (100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 500 mM potassium chloride, 1 mM TCEP, 5% glycerol, pH 7.5). Protein was washed with additional dialysis buffer and concentrated in an Amicon Ultra-15 Centrifugal Filter Unit (EMD Millipore) with a 10 kDa molecular weight cut off. Proteins were flash frozen with liquid nitrogen and stored at –80 °C.

RNA cleavage assays. RNA cleavage assays were performed as previously described[5].

2.10 References

1. Esvelt, K.M., Carlson, J.C. & Liu, D.R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499-503 (2011).

2. Dickinson, B.C., Packer, M.S., Badran, A.H. & Liu, D.R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nat Commun* **5**, 5352 (2014).

3. Badran, A.H. & Liu, D.R. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nat Commun* **6**, 8425 (2015).

4. Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J.A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355-1358 (2010).

5. Sternberg, S.H., Haurwitz, R.E. & Doudna, J.A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18**, 661-672 (2012).

6. Haurwitz, R.E., Sternberg, S.H. & Doudna, J.A. Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *The EMBO journal* **31**, 2824-2832 (2012).

7. Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A. & Arkin, A.P. RNA processing enables predictable programming of gene expression. *Nature biotechnology* **30**, 1002-1006 (2012).

8. Tsai, S.Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology* **32**, 569-576 (2014).

9. Nissim, L., Perli, S.D., Fridkin, A., Perez-Pinera, P. & Lu, T.K. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Molecular cell* **54**, 698-710 (2014).

10. Lee, H.Y. et al. RNA-protein analysis using a conditional CRISPR nuclease. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5416-5421 (2013).

11. Ringquist, S. et al. Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Mol Microbiol* **6**, 1219-1229 (1992).

12. Davis, J.H., Rubin, A.J. & Sauer, R.T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic acids research* **39**, 1131-1141 (2011).

13. Carlson, J.C., Badran, A.H., Guggiana-Nilo, D.A. & Liu, D.R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat Chem Biol* **10**, 216-222 (2014).

14. Badran, A.H. et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **533**, 58-63 (2016).

15. Busso, D., Delagoutte-Busso, B. & Moras, D. Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. *Anal Biochem* **343**, 313-321 (2005).

Chapter Three:

Continuous Directed Evolution of Csy4 Variants with Orthogonal Binding Specificity

Kevin M. Davis, Mitchell R. O'Connell, Samuel H. Sternberg,
Jennifer A. Doudna, and David R. Liu

Kevin Davis generated all constructs, performed the PACE experiments, and purified the Csy4 variants. Mitchell O'Connell and Samuel Sternberg performed the *in vitro* RNA cleavage assays and electrophoretic mobility shift assays.

3.1 Introduction: Altering the Specificity of Csy4 with PACE

The main utility of having a Csy4 PACE system is the ability to evolve activity on RNA substrates currently inaccessible, ideally, in an orthogonal manner. To demonstrate that the Csy4 PACE system can evolve Csy4 variants with altered substrate activity, we targeted two modified stem-loop substrates in which the bottom base pair is mutated from C–G to U–A or G–C (Figure 3.1). Relative to the wild-type substrate, the U–A substrate is bound 20-fold weaker and cleaved 170-fold slower by wild-type Csy4, while the G–C substrate is bound 1,200-fold weaker and cleaved 7,500-fold slower[1]. Based on these relative binding and cleavage defects, we hypothesized that evolving activity on the U–A substrate would represent an intermediate evolutionary challenge and evolving activity on the G–C substrate would represent a difficult evolutionary challenge.

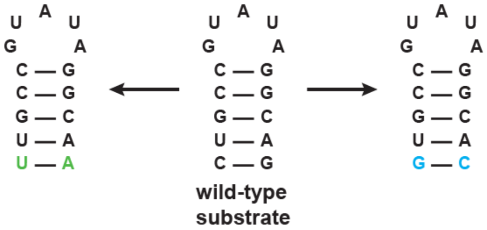


Figure 3.1 | Modified stem-loop substrates. The U–A and G–C substrates contain a mutated bottom base pair relative to the wild-type substrate.

3.2 Evolving Activity on Modified Stem-Loop Substrates

In order to evolve Csy4 activity on the U–A and G–C substrates, we modified the AP to contain the U–A or G–C stem-loop in place of the wild-type stem-loop. In plaque assays with host cells carrying either the U–A- or G–C-modified AP and phage from the end of the high stringency wild-type PACE experiment, no plaques were visible, suggesting extremely weak propagation or no propagation at all. We performed separate PACE experiments with the U–A-

and G–C-modified APs, both of which were seeded with phage from the high stringency wild-type PACE experiment. Each PACE experiment was performed with mutagenesis induction in duplicate lagoons. Phage propagation in the U–A PACE experiment was robust over 96 hours (Figure 3.2); however, the phage washed out of the G–C PACE experiment within 24 hours. Plaque assays with phage isolated at the end of the U–A PACE experiment yielded large plaques on the U–A substrate and, surprisingly, small plaques on the G–C substrate. By seeding a new G–C PACE experiment with phage isolated at the end of the U–A PACE experiment, we were able to maintain robust phage propagation over 96 hours (Figure 3.3). Plaque assays with phage isolated at the end of this successful G–C PACE experiment yielded moderate sized plaques on the G–C substrate. Sequencing the Csy4 gene of individual Csy4-SPs at the 48 and 96 hour time points of the U–A and G–C PACE experiments (Table 3.1 and Table 3.2) revealed that the phage populations had accumulated multiple consensus mutations in Csy4.

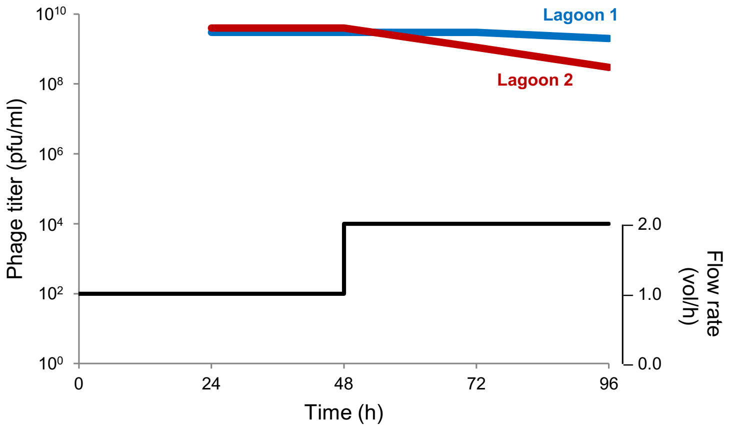


Figure 3.2 | U–A positive selection PACE. Each lagoon was inoculated with 10^6 phage from the high stringency wild-type PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

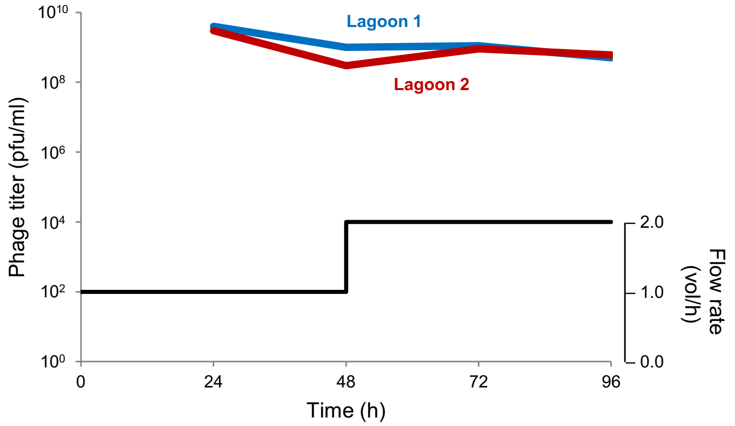


Figure 3.3 | G–C positive selection PACE. Each lagoon was inoculated with 10^6 phage from the U–A positive selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

Table 3.1 | U–A positive selection sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																	
		10	12	16	22	27	30	56	84	85	87	105	117	129	138	139	147	170	183
WT		R	D	P	S	K	Q	E	Q	F	E	A	M	K	A	L	R	E	F
L1	48 h		Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S		K	
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
	96 h	L	Y	S	N	R	L	K		I	K	V				S			
		L	Y	S	N	R	L	K		I	K	V				S			
		L	Y	S	N	R	L	K		I	K	V				S			
		L	Y	S	N	R	L	K		I	R	V				S			
		L	Y	S	N	R	L	K		I	K	V				S			
		L	Y	S	N	R	L	K		I	K	V			T	S			
		L	Y	S	N	R	L	K	R	I	K	V				S			
		L	Y	S	N	R	L	K	H	I	K	V				S			
L2	48 h		Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V			V	S			
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V	I			S			
			Y	S	N	R	L	K		I		V			S	S			
			Y	S	N	R	L	K		I		V				S			
	96 h	L	Y	S	N	R	L	K		I	K	V			T	S			
		L	Y	S	N	R	L	K	K	I		V			T	S			S
		L	Y	S	N	R	L	K	R	I		V				S			
			Y	S	N	R	L	K		I		V				S			
		L	Y	S	N	R	L	K	R	I		V				S	H		
			Y	S	N	R	L	K		I		V				S			
			Y	S	N	R	L	K		I		V		N		S			
			Y	S	N	R	L	K		I		V				S			

Table 3.2 | G–C positive selection sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																						
		10	12	16	21	22	26	27	30	37	38	56	69	84	85	87	105	138	139	144	148	155	157	172
WT		R	D	P	M	S	G	K	Q	G	D	E	A	Q	F	E	A	A	L	V	S	F	L	G
L1	48 h	L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S				F	
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
	96 h	L	Y	S		N		R	L	R	G	K		R	I	K	V	T	S		C			
		L	Y	S		N		R	L			K	G	R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					R
		L	Y	S		N		R	L	R	G	K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S				I	
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N	D	R	L		G	K		R	I	K	V	T	S					
L2	48 h	L	Y	S		N		R	L			K		R	I		V	T	S					
		L	Y	S		N		R	L			K		R	I		V	T	S					
		L	Y	S		N		R	L			K		R	I		V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K			I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I		V	T	S					
		L	Y	S		N		R	L			K		R	I		V	T	S					
	96 h	L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S	R	N		R	L			K		Y	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S	I				
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S					
		L	Y	S		N		R	L			K		R	I	K	V	T	S			Y		
		L	Y	S		N		R	L			K		R	I	K	V	T	S				I	
		L	Y	S		N		R	L			K		R	I	K	V	T	S					

3.3 Developing Negative Selection for Csy4 PACE

Based on the plaque assay results, we had successfully evolved Csy4-SPs with activity on the U–A and G–C substrates. However, plaque assays also indicated that the evolved phage populations from both PACE experiments maintained high activity on the wild-type substrate. We were interested in evolving variants with orthogonal activity, and we focused on removing activity on the wild-type substrate as well as cross-activity on the U–A or G–C substrate using negative selection. To develop negative selection for Csy4 PACE, we investigated two general strategies: engineering the gene III transcript such that off-target cleavage results in knockdown of the transcript and developing a negative selection plasmid (NSP) that expresses pIII-neg[2], a dominant negative variant of pIII, in response to off-target cleavage. Both strategies maintain use of the AP, allowing for simultaneous positive and negative selection.

To engineer the gene III transcript such that off-target cleavage results in knockdown of the transcript, we modified the AP to encoded off-target RNA substrates within gene III (Figure 3.4). Because the RNA substrates are encoded within gene III, they are placed immediately downstream of the gene III signal peptide, in frame, such that only a small N-terminal fusion peptide is present in the mature pIII protein. For the U–A-modified AP, we encoded the C–G and G–C stem-loops inside the gene III transcript, and for the G–C-modified AP, we encoded the C–G and U–A stem-loops inside the gene III transcript. Based on plaque assays, this strategy only enables low stringency negative selection. To develop a NSP that expresses pIII-neg in response to off target cleavage, we created a plasmid identical to the AP, except with gene III-neg instead of gene III and a higher copy number origin of replication (Figure 3.5). In this format, we were able to manipulate the negative selection pressure by using origins of replication with varying copy number, and plaque assays indicated that medium and high stringency negative selection could be applied with medium- and high-copy origins of replication, respectively. The one limitation of the NSP is that only one off-target substrate can be negatively selected against at a time.

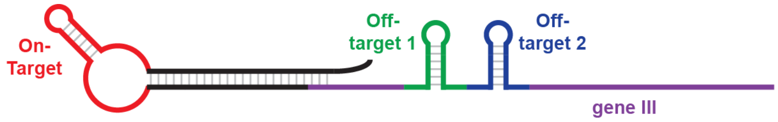


Figure 3.4 | Low stringency negative selection. The AP expresses a translationally repressed gene III transcript that also contains the off-target stem-loop substrates within the coding region of gene III. On-target RNA cleavage leads to derepression and pIII expression while off-target RNA cleavage leads to destruction of the transcript.

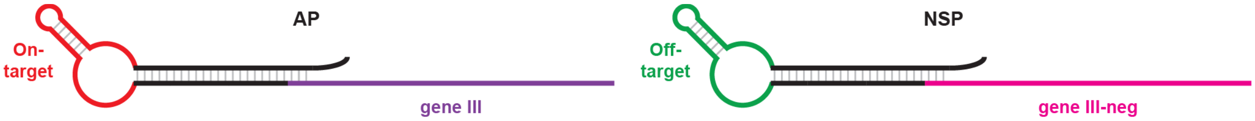


Figure 3.5 | Medium and high stringency negative selection. Host cells carry both an AP and a negative selection plasmid (NSP). The NSP expresses a translationally repressed gene III-neg transcript. Off-target RNA cleavage results in the production of pIII-neg, a dominant negative variant of pIII, which results in the production of non-infectious progeny. Negative selection stringency is a function of the copy number of the NSP and can be tuned by changing the origin of replication.

3.4 Removing Wild-Type Activity Using Negative Selection

To generate Csy4-SPs with orthogonal activity on the U–A substrate we performed successive PACE experiments with low, medium, and high stringency negative selection. During low stringency negative selection, the C–G and G–C stem-loops were encoded within the gene III transcript of the AP, while during the medium and high stringency negative selection, a NSP encoding the C–G stem-loop was used. Each PACE experiment was performed with mutagenesis induction in duplicate lagoons, and the low stringency negative selection PACE experiment was seeded with phage from the initial U–A positive selection PACE experiment. Phage propagation in each PACE experiment was robust over 96 hours (Figure 3.6, Figure 3.7, and Figure 3.8). Plaque assays with phage isolated at the end of the high stringency negative selection yielded large plaques on the U–A substrate and no visible plaques on the C–G and G–C substrates. Sequencing the Csy4 gene of individual Csy4-SPs at the 48 and 96 hour time points of each PACE experiment (Table 3.3, Table 3.4, and Table 3.5) revealed the accumulation of numerous consensus mutations.

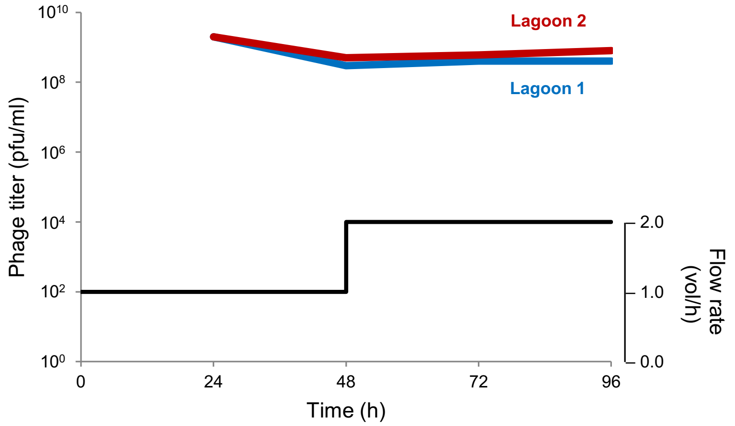


Figure 3.6 | U–A positive selection with low stringency negative selection PACE. Each lagoon was inoculated with 10^6 phage from the U–A positive selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

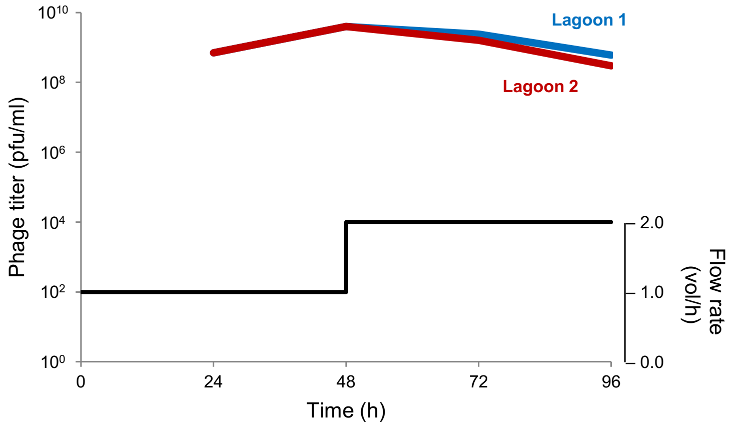


Figure 3.7 | U–A positive selection with medium stringency negative selection (against C–G) PACE. Each lagoon was inoculated with 10^6 phage from the U–A low stringency negative selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP6 in both lagoons.

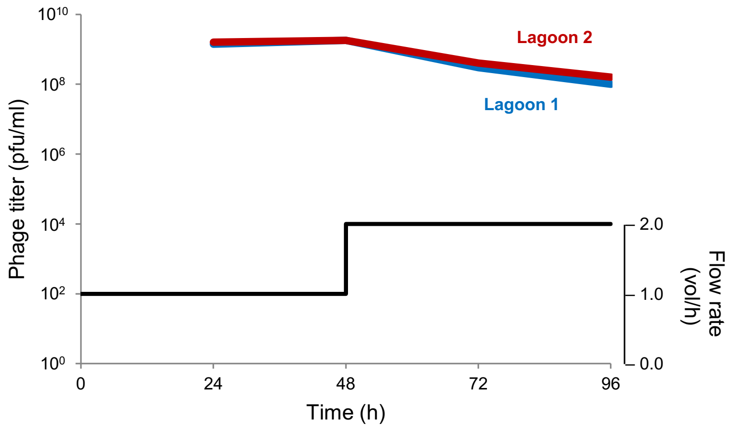


Figure 3.8 | U–A positive selection with high stringency negative selection (against C–G) PACE. Each lagoon was inoculated with 10^6 phage from the U–A medium stringency negative selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP6 in both lagoons.

Table 3.3 | U–A positive selection with low stringency negative selection sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																									
		10	12	16	22	23	26	27	29	30	33	56	65	84	85	87	96	105	107	111	125	129	138	139	144	153	167
WT		R	D	P	S	V	G	K	H	Q	V	E	D	Q	F	E	P	A	S	R	E	K	A	L	V	Q	T
L1	48 h	L	Y	S	N		D	R		L		K		R	I	K	S	V	N				S	S			
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S			
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S			
		L	Y	S	N			R		L		K		R	I	K		V	N					S			
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S			
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S			A
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S			
		L	Y	S	N			R		L		K		R	I	K		V	N					S			
	96 h	L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S		H	
		L	Y	S	N		D	R		L	M	K		R	I	K		V	N				S	S			
		L	Y	S	N		D	R		L		K		R	I	K		V	N				S	S		H	
		L	Y	S	N			R		L		K		R	I	K		V		C	A			S		H	
		L	Y	S	N			R		L		K		R	I	K		V		C				S		H	
		L	Y	S	N			R		L		K		R	I	K		V		C				S		H	
		L	Y	S	N			R		L		K		R	I	K		V		C				S		H	
		L	Y	S	N		D	R		L		K		R	I	K		V		C				S		H	
L2	48 h	L	Y	S	N			R		L		K		R	I	K		V	N					S	I		
		L	Y	S	N			R		L		K	A	R	I	K		V	N			N		S			
		L	Y	S	N			R		L		K	A	R	I	K		V	N			N		S			
		L	Y	S	N			R		L		K	A	R	I	K		V	N			N		S			
		L	Y	S	N			R		L		K		R	I	K		V	N					S	I		
		L	Y	S	N			R		L		K	A	R	I	K		V	N			N		S			
		L	Y	S	N			R		L		K		R	I	K		V	N					S	I		
		L	Y	S	N			R		L		K		R	I	K		V	N					S	I		
	96 h	L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I	H	
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I	H	
		L	Y	S	N		D	R		L		K	A	R	I	K		V	N			N		S		H	
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I	H	
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I		
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I		
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I	H	
		L	Y	S	N	E		R	Y	L		K		R	I	K		V	N					S	I		

Table 3.4 | U–A positive selection with medium stringency negative selection (against C–G) sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																												
		10	12	16	22	23	27	30	31	33	34	38	50	56	84	85	87	102	103	104	105	106	107	121	139	141	144	146	150	170
WT		R	D	P	S	V	K	Q	A	V	A	D	S	E	Q	F	E	R	V	Q	A	K	S	D	L	L	V	L	S	E
L1	**48 h**	L	Y	S	N	E	R	L	V	E	T	G		K	R	I	K				V		D		S	P	I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K				V		D		S	P	I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E	T	G		K	R	I	K				V		D		S	P	I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K				V		D		S	P	I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
	96 h	L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D	G	S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			K
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
		L	Y	S	N	E	R	L		E	V	G		K	R	I	K	G	G		V		D		S		I			
L2	**48 h**	L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L	D	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K				V	I	D		S		I		G	
		L	Y	S	N	E	R	L		E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L		E		G		K	R	I	K			L	V		D		S		I			
	96 h	L	Y	S	N	E	R	L	V	E		G		K	R	I	K			K	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G	N	K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I			
		L	Y	S	N	E	R	L	V	E		G		K	R	I	K			L	V		D		S		I	I		

Table 3.5 | U–A positive selection with high stringency negative selection (against C–G) sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																																							
		10	12	16	22	23	25	27	30	31	33	34	38	47	49	51	56	64	84	85	87	88	100	102	103	105	107	114	117	123	124	125	138	139	144	146	150	154	169	172	183
WT		R	D	P	S	V	F	K	Q	A	V	A	D	L	E	R	E	A	Q	F	E	P	V	R	V	A	S	R	M	S	E	E	A	L	V	L	S	H	E	G	F
L1	48 h	L	Y	S	N	E		R	L	V	E		G				K		R	I	K		I	G		V	D	Q						S	I		G			V	
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K			G	G	V	D		I					S	I						
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K			G	G	V	D				G			S	I						
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K	T		G	G	V	D		I					S	I						
		L	Y	S	N	E		R	L		E	V	G			C	K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G			C	K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K			G	G	V	D			G			S	S							
	96 h	L	Y	S	N	E		R	L	T	E		G				K		R	I	K			G	G	V	D				K			S	I						
		L	Y	S	N	E		R	L	T	E		G	F			K		R	I	K			G	G	V	D					K		S	I		G				
		L	Y	S	N	E		R	L	T	E		G				K		R	I	K			G	G	V	D					K		S	I						
		L	Y	S	N	E		R	L	T	E		G				K		R	I	K			G	G	V	D				K			S	I		G				
		L	Y	S	N	E		R	L		E	V	G				K		R	I	K			G	G	V	D				G			S	I						
		L	Y	S	N	E		R	L	T	E		G				K		R	I	K			G	G	V	D				K			S	I						
		L	Y	S	N	E		R	L	T	E		G				K		R	I	K		I	G	G	V	D							S	I		G				L
		L	Y	S	N	E		R	L	T	E		G				K		R	I	K			G	G	V	D				K			S	I						
L2	48 h	L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E	Y	R	L		E	V	G				K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L	V	E		G				K		R	I	K			G		V	D	Q						S	I		G			V	
		L	Y	S	N	E	Y	R	L		E	V	G				K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G				K	E	R	I	K			G	G	V	D					K		S	I						
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E	Y	R	L		E	V	G				K		R	I	K			G	G	V	D							S	I						
	96 h	L	Y	S	N	E		R	L		E		G				K		R	I	K			G		V	D					G		S	I	I	G			V	
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I				A		
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I						
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I			L			
		L	Y	S	N	E		R	L		E		G				K		R	I	K			G		V	D					G		S	I	I	G	L		V	
		L	Y	S	N	E		R	L		E	V	G		G		K		R	I	K			G	G	V	D							S	I			L			
		L	Y	S	N	E		R	L		E		G				K		R	I	K			G	A	V	D			R				S	I	I	G			V	

To generate Csy4-SPs with orthogonal activity on the G–C substrate we similarly performed successive PACE experiments with low, medium, and high stringency negative selection. During low stringency negative selection, the C–G and U–A stem-loops were encoded within the gene III transcript of the AP, while during the medium and high stringency negative selection, a NSP encoding the C–G stem-loop was used. Each PACE experiment was performed with mutagenesis induction in duplicate lagoons, and the low stringency negative selection PACE experiment was seeded with phage from the initial G–C positive selection PACE experiment. Phage propagation in each PACE experiment was robust over 96 hours (Figure 3.9, Figure 3.10, and Figure 3.11). Plaque assays with phage isolated at the end of the high stringency negative selection yielded large plaques on the G–C substrate, no visible plaques on the C–G substrate, and small plaques on the U–A substrate. To remove residual activity on the U–A substrate, we performed an additional PACE experiment using a high-copy number NSP encoding the U–A stem-loop (Figure 3.12). Sequencing the Csy4 gene of individual Csy4-SPs at the 48 and 96 hour time points of the each PACE experiment (Table 3.6, Table 3.7, Table 3.8, and Table 3.9) revealed the accumulation of numerous consensus mutations. Serendipitously, we identified two phage from the final negative selection PACE experiment that yielded large plaques on the C–G substrate. Sequencing the Csy4 gene of these two phage revealed that they contained a single W104R mutation relative to the consensus mutant (Table 3.9), suggesting that this single mutation abolishes substrate specificity.

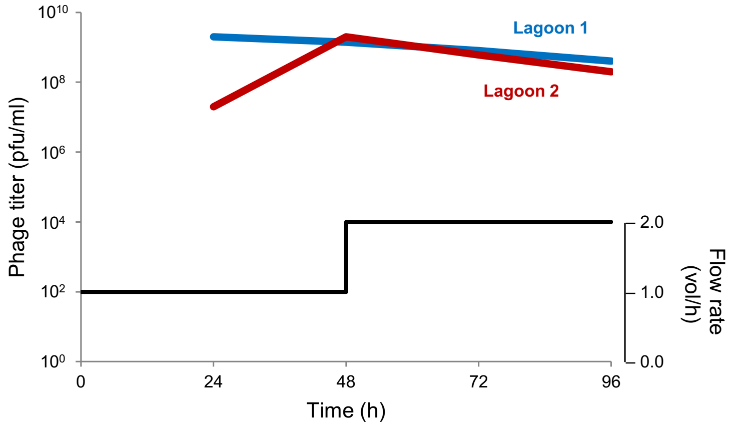


Figure 3.9 | G–C positive selection with low stringency negative selection PACE. Each lagoon was inoculated with 10^6 phage from the G–C positive selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP4 in both lagoons.

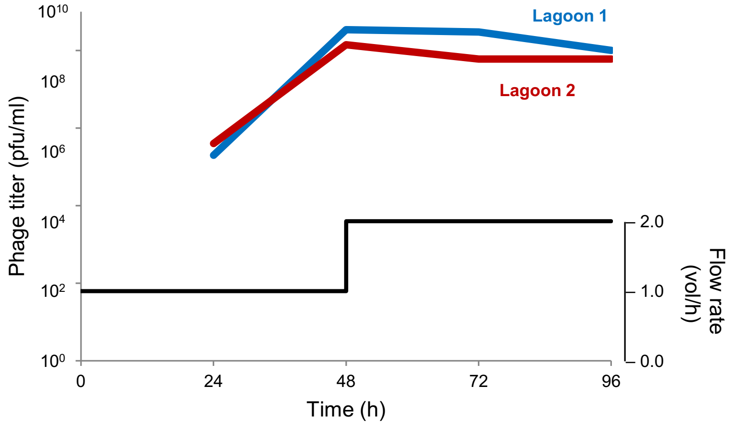


Figure 3.10 | G–C positive selection with medium stringency negative selection (against C–G) PACE. Each lagoon was inoculated with 10^6 phage from the G–C low stringency negative selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP6 in both lagoons.

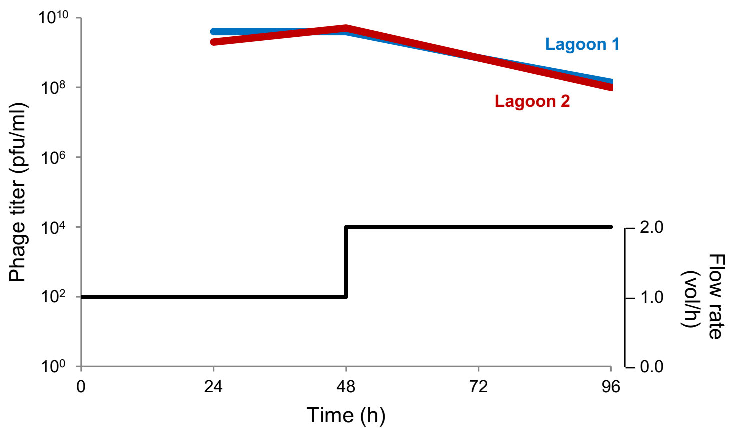


Figure 3.11 | G–C positive selection with high stringency negative selection (against C–G) PACE. Each lagoon was inoculated with 10^6 phage from the C–G medium stringency negative selection PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP6 in both lagoons.

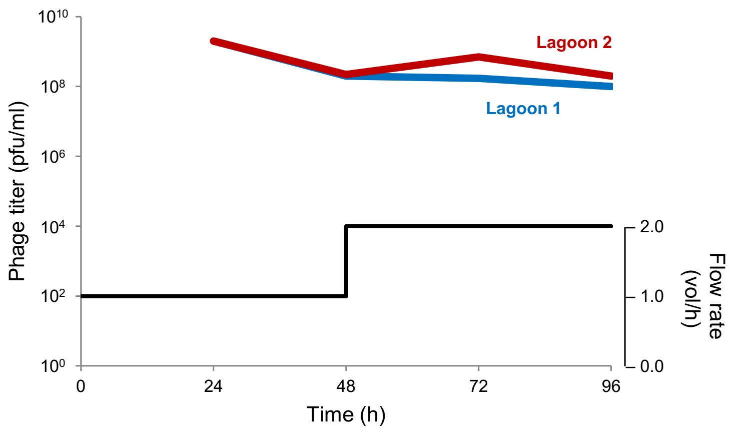


Figure 3.12 | G–C positive selection with high stringency negative selection (against U–A) PACE. Each lagoon was inoculated with 10^6 phage from the C–G high stringency negative selection (against C–G) PACE experiment, and PACE was run for a total of 96 hours. Lagoon samples were taken every 24 hours and the phage titer was determined by plaque assay. Mutagenesis was induced with arabinose from MP6 in both lagoons.

Table 3.6 | G–C positive selection with low stringency negative selection sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																		
		10	12	16	21	22	27	30	36	56	84	85	87	102	105	138	139	147	148	174
WT		R	D	P	M	S	K	Q	G	E	Q	F	E	R	A	A	L	R	S	T
L1	**48 h**	L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K		V	T	S		N	
	96 h	L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S	C	N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L		K	Y	I	K	G	V	T	S		N	
		L	Y	S	R	N	R	L	S	K	Y	I	K	G	V	T	S		N	
L2	**48 h**	L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			
	96 h	L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N
		L	Y	S	R	N	R	L		K	Y	I	K	Q	V	T	S			N

Table 3.7 | G–C positive selection with medium stringency negative selection (against C–G) sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																														
		10	12	16	21	22	27	30	38	39	47	56	84	85	87	102	105	107	108	109	111	116	118	127	129	135	136	138	139	147	148	169
WT		R	D	P	M	S	K	Q	D	R	L	E	Q	F	E	R	A	S	N	P	R	L	R	A	K	V	A	A	L	R	S	E
L1	**48 h**	L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C							T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C							T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
	96 h	L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D	L	C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C				R		V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
		L	Y	S	R	N	R	L	G			K	Y	I	K	G	V		D		C						V	T	S	C	N	
L2	**48 h**	L	Y	S	R	N	R	L				K	Y	I	K	G	V				C				E	A		T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V	N								A	V	T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C					A		T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V						H			A		T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C					A		T	S		T	
		L	Y	S	R	N	R	L		K		K	Y	I	K	G	V	N								A	V	T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V					F				A		T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V							T		A		T	S		T	
	96 h	L	Y	S	R	N	R	L			F	K	Y	I	K	G	G									A		T	P		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C					A		T	S		T	K
		L	Y	S	R	N	R	L			F	K	Y	I	K	G	G									A		T	P		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V									A	V	T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	G									A		T	P		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C					A		T	S		T	K
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C					A		T	S		T	
		L	Y	S	R	N	R	L				K	Y	I	K	G	V				C				E	A		T	S		T	

Table 3.8 | G–C positive selection with high stringency negative selection (against C–G) sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

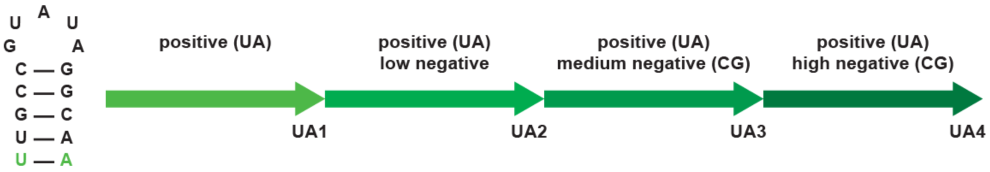
		Csy4 Residue																															
		7	10	12	16	17	21	22	23	27	30	31	34	36	38	42	46	51	56	57	73	84	85	87	102	104	105	108	111	136	138	139	147	148	170
WT		I	R	D	P	P	M	S	V	K	Q	A	A	G	D	V	D	R	E	R	R	Q	F	E	R	Q	A	N	R	A	A	L	R	S	E
L1	48 h		L	Y	S		R	N		R	L				G							Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S	S	R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G			C	K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G			C	K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G	M			K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G	M			K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
	96 h		L	Y	S		R	N	E	R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L	D			G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L	D	T		G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N	E	R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N	E	R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N	E	R	L			D	G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	K
			L	Y	S		R	N	E	R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
L2	48 h		L	Y	S		R	N		R	L				G			C	K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
		V	L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G		G		K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G			C	K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G			H	K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K	H		Y	I	K	G	W	V	D	C	V	T	S	C	N	
	96 h		L	Y	S		R	N		R	L	D			G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K		Q	Y	I	K	A	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L				G				K			Y	I	K	G	W	V	D	C	V	N	S	C	N	
			L	Y	S		R	N		R	L	D			G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		R	L	D			G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		S	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	
			L	Y	S		R	N		S	L				G				K			Y	I	K	G	W	V	D	C	V	T	S	C	N	

Table 3.9 | G–C positive selection with high stringency negative selection (against U–A) sequencing. Eight Csy4-SPs from the 48 and 96 hour time points of each lagoon (L1 and L2) were sequenced. Mutated residues are specified by the new amino acid at the indicated position; wild-type residues at these positions are specified at the top.

		Csy4 Residue																														
		10	12	16	17	21	22	23	27	28	30	38	50	51	56	69	84	85	87	102	104	105	108	111	135	136	138	139	147	148	154	183
WT		R	D	P	P	M	S	V	K	L	Q	D	S	R	E	A	Q	F	E	R	Q	A	N	R	V	A	A	L	R	S	H	F
L1	48 h	L	Y	S		R	N	G	R		L				K		Y	I	K	G	W	V	D	C	A		T	S		T	Y	S
		L	Y	S		R	N	E	R		L			C	K		Y	I	K	G	W	V	D	C	A		T	S		T		S
		L	Y	S		R	N	G	R		L				K		Y	I	K	G	W	V	D	C	A		T	S		T	Y	S
		L	Y	S		R	N		R		L	G		C	K	T	Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N	E	R		L			C	K		Y	I	K	G	W	V	D	C	A		T	S		T		S
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N	G	R		L	G			K		Y	I	K	G	W	V	D	C	A		T	S		T		
		L	Y	S	S	R	N	E	R		L	G			K		Y	I	K	G	W	V	D	C		V	T	S	C	N		
	96 h	L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K	D	Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N	E	R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
L2	48 h	L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N	E	R		L			C	K		Y	I	K	G	W	V	D	C	A		T	S		T		S
		L	Y	S		R	N		R	F	L	G	R		Q		Y	I	K	G	W	V	D	C		V	T	S	C	N		
		L	Y	S		R	N	G	R		L				K		Y	I	K	G	W	V	D	C	A		T	S		T	Y	S
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
	96 h	L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C	A	V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	W	V	D	C		V	T	S		T		
Activity on C–G		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	R	V	D	C		V	T	S		T		
		L	Y	S		R	N		R		L	G		C	K		Y	I	K	G	R	V	D	C		V	T	S		T		

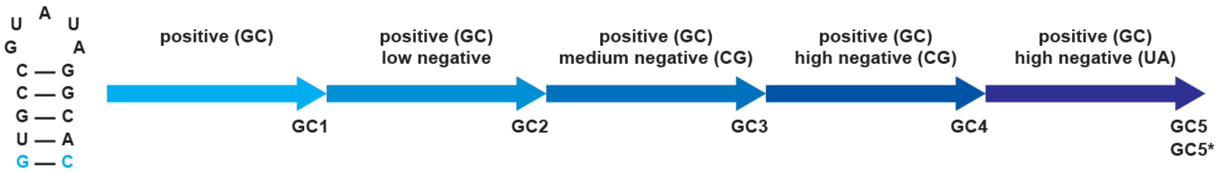
3.5 *In Vitro* Characterization of the Orthogonal Csy4 Variants

With Csy4-SPs demonstrating orthogonal activity on the U–A or G–C substrates in plaque assays, we moved forward to assess the *in vitro* activity of the evolved Csy4 variants. We isolated and purified representative variants from each of the four U–A PACE experiments (Figure 3.13) and each of the five G–C PACE experiments (Figure 3.14). Unfortunately, *in vitro* RNA cleavage experiments indicated that cleavage activity was lost for both the U–A and G–C evolutionary paths following the first low stringency negative selection. This result suggests that RNA binding alone had become sufficient to derepress the gene III transcript of the AP, something that was not the case for catalytically inactive wild-type Csy4 (see Figure 2.7).



	Csy4 Residue																						
	10	12	16	22	23	27	29	30	33	34	38	49	56	84	85	87	102	103	105	107	139	144	153
WT	R	D	P	S	V	K	H	Q	V	A	D	E	E	Q	F	E	R	V	A	S	L	V	Q
UA1	L	Y	S	N		R		L					K		I	K			V		S		
UA2	L	Y	S	N	E	R	Y	L					K	R	I	K			V	N	S	I	H
UA3	L	Y	S	N	E	R		L	E	V	G		K	R	I	K	G	G	V	D	S	I	
UA4	L	Y	S	N	E	R		L	E	V	G	G	K	R	I	K	G	G	V	D	S	I	

Figure 3.13 | Summary of the U–A PACE experiments. Csy4 variants with orthogonal activity on the U–A substrate were isolated after four successive PACE experiments. Representative variants (UA1, UA2, UA3, UA4) from each experiment were purified for *in vitro* characterization.



	Csy4 Residue																						
	10	12	16	21	22	27	30	38	51	56	84	85	87	102	104	105	108	111	136	138	139	147	148
WT	R	D	P	M	S	K	Q	D	R	E	Q	F	E	R	Q	A	N	R	A	A	L	R	S
GC1	L	Y	S		N	R	L			K	R	I	K			V				T	S		
GC2	L	Y	S	R	N	R	L			K	Y	I	K	G		V				T	S		N
GC3	L	Y	S	R	N	R	L	G		K	Y	I	K	G		V	D	C	V	T	S	C	N
GC4	L	Y	S	R	N	R	L	G		K	Y	I	K	G	W	V	D	C	V	T	S	C	N
GC5	L	Y	S	R	N	R	L	G	C	K	Y	I	K	G	W	V	D	C	V	T	S		T
GC5*	L	Y	S	R	N	R	L	G	C	K	Y	I	K	G	R	V	D	C	V	T	S		T

Figure 3.14 | Summary of the G–C PACE experiments. Csy4 variants with orthogonal activity on the G–C substrate were isolated after five successive PACE experiments. Representative variants (GC1, GC2, GC3, GC4, GC5, GC5*) from each experiment were purified for *in vitro* characterization.

Despite the lack of cleavage activity, we tested the RNA-binding specificity of the Csy4 variants from the U–A and G–C evolutionary paths in electrophoretic mobility shift assays (EMSAs). Comparing equilibrium dissociation constants (K_d) for each UA variant and the wild-type and U–A stem-loops reveals a change in substrate preference over the course of the four U–A PACE experiments (Figure 3.15). Overall, this represents an approximate 320-fold change in specificity relative to catalytically inactive wild-type Csy4. RNA binding curves for wild-type Csy4 and variant UA4 (Figure 3.16) highlight the change in substrate specificity. Unfortunately, despite demonstrating altered substrate specificity, UA4 has a much weaker binding affinity for its preferred substrate than wild-type Csy4 has for its cognate substrate.

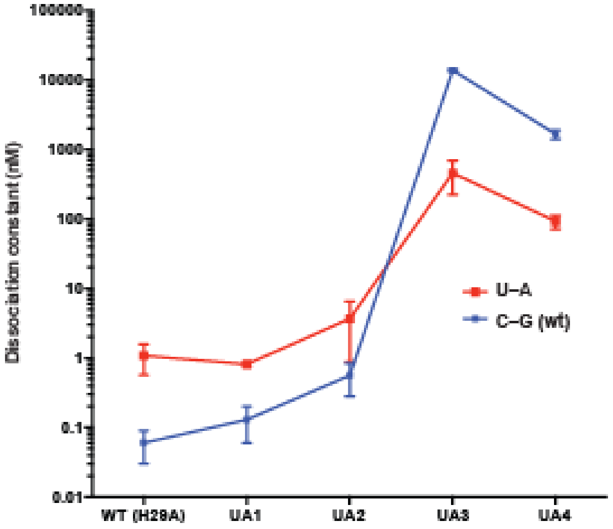


Figure 3.15 | Substrate specificity of the UA variants. Equilibrium dissociation constants for each UA variant and the wild-type and U–A stem-loops were determined by electrophoretic mobility shift assays.

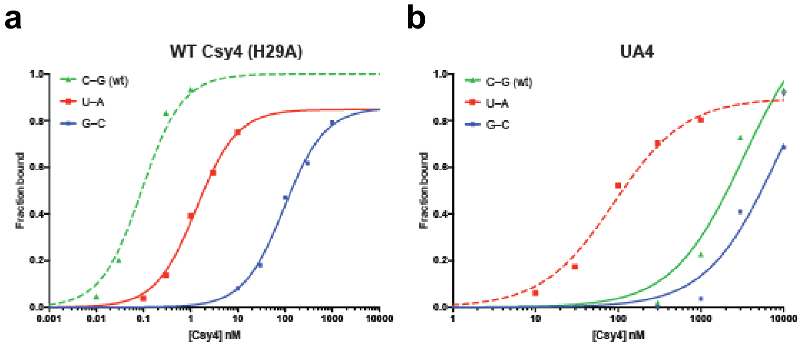


Figure 3.16 | UA4 RNA binding curves. RNA binding curves with the wild-type, U–A, and G–C stem-loops are shown for (a) wild-type Csy4 and (b) UA4.

Comparing K_d values for each GC variant and the wild-type and G–C stem-loops reveals a more striking swap in substrate preference over the course of the five G–C PACE experiments (Figure 3.17). Overall, this represents an approximate 12,000,000-fold change in specificity relative to catalytically inactive wild-type Csy4. RNA binding curves for wild-type Csy4 and variant GC5 reveal that GC5 is actually more specific than wild-type Csy4 (Figure 3.18). Similar to UA4, GC5 shows a weaker binding affinity for its preferred substrate than wild-type Csy4 has for its cognate substrate; however, GC5 still maintains low nanomolar affinity for the G–C stem-loop. GC5*, the W104R mutant with high activity on the wild-type substrate in plaque assays, demonstrates almost no preference among the wild-type, U–A, and G–C substrates, which is pretty remarkable given that it only has a single mutation relative to GC5. It is interesting to note that both of Csy4's base-specific residues, Arg102 and Gln104, are mutated in GC5.

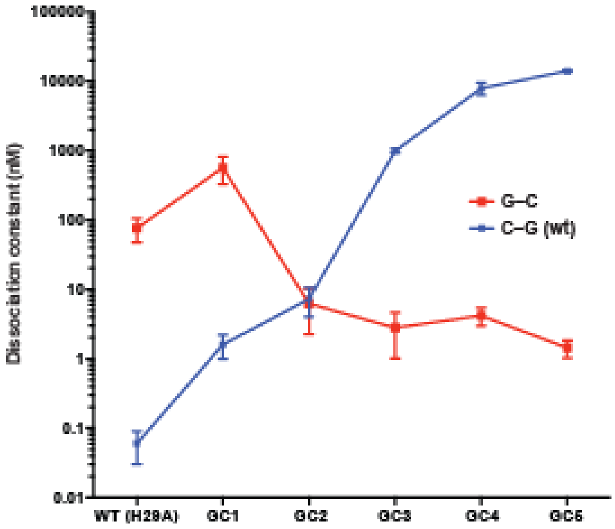


Figure 3.17 | Substrate specificity of the GC variants. Equilibrium dissociation constants for each GC variant and the wild-type and G–C stem-loops were determined by electrophoretic mobility shift assays.

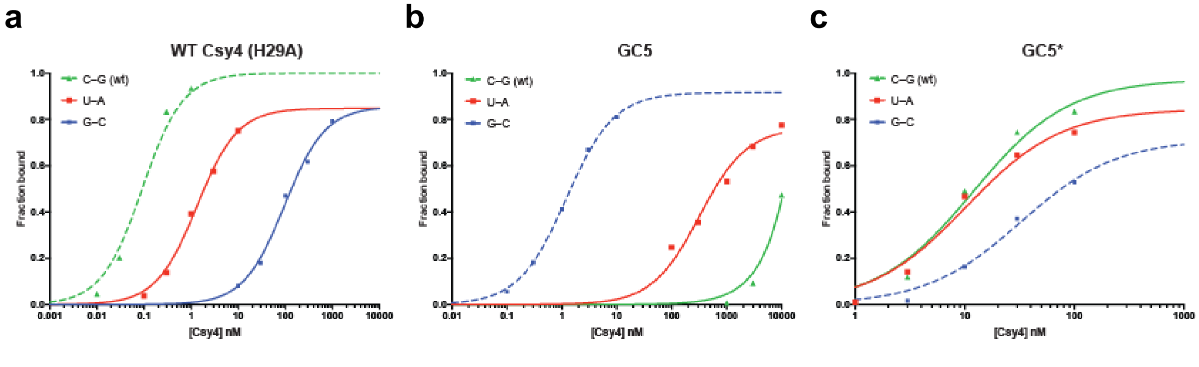


Figure 3.18 | GC5 and GC5* RNA binding curves. RNA binding curves with the wild-type, U–A, and G–C stem-loops are shown for (a) wild-type Csy4, (b) GC5, and (c) GC5*.

3.6 Conclusion

The Csy4 PACE system can be used to evolve Csy4 variants with activity on modified stem-loop substrates, and the negative selection capabilities developed here enable the removal of off-target activity. Using negative selection we evolved Csy4 variants with orthogonal RNA-binding specificity for two modified stem-loops. While the variants ultimately lost cleavage activity, several variants demonstrate a remarkable change in RNA-binding specificity relative to wild-type Csy4.

3.7 Methods

Bacterial strains. DNA cloning was performed with NEB Turbo cells (New England Biolabs) or Mach1 cells (Thermo Fisher Scientific). Protein expression was performed with BL21 Star (DE3) cells (Thermo Fisher Scientific). All luciferase assays, plaque assays, and PACE experiments were performed with S1030 cells[2]. Electrocompetent S1030 cells were prepared as previously described[3]. Briefly, cells were grown to an OD_{600} value of 0.8, pelleted by centrifugation at 10,000g, and the supernatant was decanted. The cells were resuspended in chilled 10% glycerol and pelleted by centrifugation at 10,000g with subsequent removal of the supernatant. Washing with 10% glycerol was repeated an additional three times. Cells were flash frozen with liquid nitrogen and stored at –80 °C. Plasmids were transformed into S1030 cells by electroporation using an *E. coli* Pulser Transformation Apparatus (Bio-Rad). Liquid cultures were supplemented with the appropriate antibiotics (Gold Biotechnology) in the following final concentrations: streptomycin (S1030 cells – *rpsL* marker; 50 µg/ml), tetracycline (S1030 cells – F plasmid; 10 µg/ml), carbenicillin (AP, Csy4 expression vector; 50 µg/ml), chloramphenicol (MP; 40 µg/ml), and spectinomycin (NSP; 100 µg/ml). Liquid cultures were incubated in a 37 °C shaker unless otherwise noted. Agar plates were similarly supplemented with the appropriate antibiotics; however, streptomycin and tetracycline were not routinely included in agar plates. Agar plates were incubated 37 °C.

General cloning methods. PCR fragments were generated using PfuTurbo Cx Hotstart DNA polymerase (Agilent Technologies), VeraSeq ULtra DNA polymerase (Enzymatics), or Phusion U Hot Start DNA polymerase (Life Technologies), and appropriate DNA primers containing an internal deoxyuracil base (Integrated DNA Technologies). PCR products were purified using a MinElute PCR Purification Kit (Qiagen). Plasmids were constructed by USER cloning as previously described[3]. Briefly, PCR products were mixed in an equimolar ratio, and incubated with DpnI (New England Biolabs) and USER enzyme (New England Biolabs) at 37 °C for 30 min. The assembly reaction was heated to 80 °C and slowly cooled to 20 °C at 0.1 °C/s at which point it was transformed into *E. coli*. Plasmids were transformed into chemically competent NEB Turbo cells or Mach1 cells according to the manufacturer's instructions, and plated on 2xYT-agar (United States Biological) plates. Plasmid DNA was amplified from colonies using an illustra TempliPhi Amplification Kit (GE Healthcare) and sequence verified by Sanger sequencing. Sequence verified colonies were grown overnight in 3 ml 2xYT media (United States Biological) and plasmid DNA was harvested using a QIAprep Miniprep Kit (Qiagen).

Plaque assays. S1030 cells carrying the AP were grown overnight in 2xYT media. Following overnight growth, cultures were diluted 50-fold into 3 ml 2xYT media and grown to an OD_{600} value of 0.7. Phage were serially diluted 10-fold and 10 µl of each dilution was mixed with 90 µl of cells. The phage/cell mixture was mixed with 1 ml of warm top agar (7g/l agar in 2xYT) and plated onto bottom agar (1.6g/l agar in 2xYT) plates. Plates were grown overnight for 20 h.

PACE. PACE experiments were setup as previously described[2]. S1030 cells carrying the AP, MP, and NSP (when appropriate) were tested for arabinose sensitivity as previously described[3]. Cultures were grown in 3ml DRM to an OD_{600} value of 0.5 and transferred to the chemostat, which was subsequently grown until it was visibly turbid. The chemostat culture was maintained at 100 ml and diluted at a rate of approximate 1.5 volumes per hour as previously described. Lagoons were maintained at 30 ml and supplemented with 20mM arabinose to induce mutagenesis. Lagoon flow rates were adjusted as described for each experiment. Lagoon samples were taken every 24 h, and phage were isolated by pelleting the cells at 10,000g and subsequently filtering the supernatant through a 0.2 μm PES syringe filter (Corning). Phage titers were determined by plaque assay.

Protein expression and purification. Csy4 variants were cloned into a pHMGWA expression vector[4] with an N-terminal His_6 tag and a $(GGS)_2$ linker. The Csy4 expression vector was transformed into chemically competent BL21 Star (DE3) cells according to the manufacturer's instructions. Colonies were grown in 150 ml 2xYT media to an OD_{600} value of 0.6, at which point protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Gold Biotechnology). Cultures were transferred to an 18 °C shaker and incubated for 16 h. Csy4 variants were purified as previously described[5] except without size exclusion chromatography. Briefly, cells were resuspended in a lysis buffer (15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 10 mM imidazole, 1 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), 0.5mM phenylmethylsulfonyl fluoride (PMSF), 5% glycerol, 0.01% Triton X-100, 100 U/ml DNase I, pH 7.4, supplemented with protease inhibitors (Roche)) and lysed by sonication. The clarified lysate was incubated with HisPur Ni-NTA Resin (Thermo Fisher Scientific) in batch, and the resin was washed with lysis buffer (lacking PMSF, DNase I, and protease inhibitors). Bound protein was eluted with a high imidazole buffer (15.5 mM disodium hydrogen phosphate, 4.5 mM sodium dihydrogen phosphate, 500 mM sodium chloride, 300 mM imidazole, 1 mM TCEP, 5% glycerol, pH 7.4) and dialyzed overnight in dialysis buffer (100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 500 mM potassium chloride, 1 mM TCEP, 5% glycerol, pH 7.5). Protein was washed with additional dialysis buffer and concentrated in an Amicon Ultra-15 Centrifugal Filter Unit (EMD Millipore) with a 10 kDa molecular weight cut off. Proteins were flash frozen with liquid nitrogen and stored at –80 °C.

RNA cleavage assays. RNA cleavage assays were performed as previously described[1].

Electrophoretic mobility shift assays. Electrophoretic mobility shift assays were performed as previously described[1].

3.8 References

1. Sternberg, S.H., Haurwitz, R.E. & Doudna, J.A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18**, 661-672 (2012).

2. Carlson, J.C., Badran, A.H., Guggiana-Nilo, D.A. & Liu, D.R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat Chem Biol* **10**, 216-222 (2014).

3. Badran, A.H. et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **533**, 58-63 (2016).

4. Busso, D., Delagoutte-Busso, B. & Moras, D. Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. *Anal Biochem* **343**, 313-321 (2005).

5. Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J.A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355-1358 (2010).

Chapter Four:

Small-Molecule Control of Cas9 Activity

Kevin M. Davis, Vikram Pattanayak, David B. Thompson,
John A. Zuris, and David R. Liu

Kevin Davis generated all constructs and performed all experiments. Vikram Pattanayak analyzed all high-throughput DNA sequencing data and performed all statistical analyses. David Thompson assisted with flow cytometry experiments. John Zuris assisted with high-throughput DNA sequencing experiments.

Adapted from *Nat. Chem. Biol.* **11**, 316-318 (2015) and *Cell Chem. Biol.* **23**, 57-73 (2016).

4.1 Introduction: Genome Editing with Programmable Nucleases

The ability to precisely manipulate a DNA sequence of interest in living cells can reveal the function of genes and regulatory sequences and has the potential to enable treatments for genetic diseases. Targeting a single locus within a vast genome, however, is a substantial challenge. The discovery and development of programmable nucleases, which can introduce double-stranded breaks at specified locations in the genome, has greatly increased our ability to make precise and efficient genetic modifications. Double-stranded breaks at a DNA locus of interest can trigger endogenous cellular DNA repair processes that lead to the disruption or replacement of genes. Gene disruption, as a result of genetic insertion or deletion mutations (indels), can occur during error-prone nonhomologous end-joining (NHEJ)[1,2], while replacement of the original DNA sequence can occur, with the aid of an exogenous DNA template, during homology-directed repair (HDR)[2,3] (Figure 4.1).

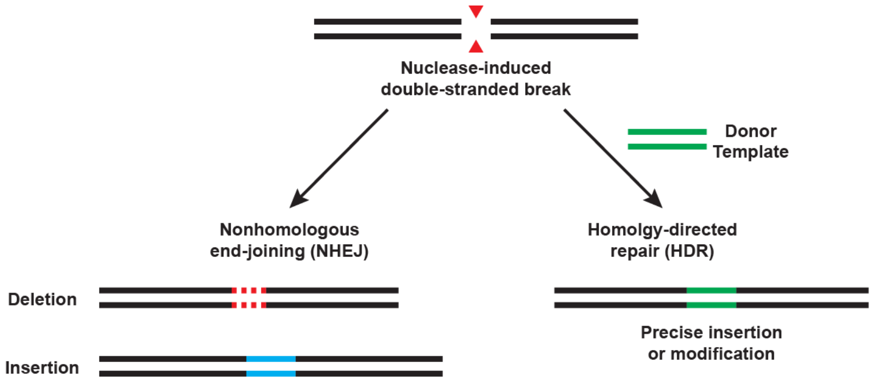

Nuclease-induced double-stranded break
Donor Template
Nonhomologous end-joining (NHEJ)
Homolgy-directed repair (HDR)
Deletion
Insertion
Precise insertion or modification

Figure 4.1 | Repair pathways following a double-stranded DNA break. Nuclease induced double-stranded breaks are typically repaired by nonhomologous end-joining (NHEJ) or homology-directed repair (HDR). The imprecise nature of NHEJ can lead to random insertion or deletion mutations (indels), which may lead to gene disruption. In the presence of a single- or double-stranded DNA donor template, HDR can mediate precise DNA replacement.

Zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) were the first two widely used programmable DNA-cleaving proteins. Both ZFNs and TALENs rely on protein domains with intrinsic DNA specificity. ZFNs are composed of ZF domains, each of which specifies binding to three base pairs of DNA[4-7], while TALENs are composed of TALE domains, each of which recognizes a single DNA base pair[8-11]. Linking together a string of ZF or TALE domains in a modular fashion generates programmable DNA-binding proteins capable of recognizing long contiguous DNA sequences. Fusing these ZF or TALE DNA-binding proteins to the FokI endonuclease leads to ZFNs[12] and TALENs[13-15], respectively, capable of sequence-specific DNA cleavage.

The recent discovery and development of the CRISPR-Cas9 system has further increased the accessibility of programmable nucleases. Cas9 is an RNA-programmable endonuclease from the type II CRISPR-Cas system that is targeted using a guide RNA (gRNA)[16,17]. In the native type II CRISPR-Cas system, the gRNA is composed of two short pieces of RNA, the crRNA and the tracrRNA[18]. The crRNA contains a 20 nucleotide spacer sequence that guides the Cas9 protein to the target DNA locus through complementary base pairing. By changing the spacer sequence, Cas9 can be directed to DNA sequences of interest, with the main requirement being a conserved PAM sequence downstream of the target site. Early studies demonstrated that the two-RNA crRNA and tracrRNA system can be simplified by combining them into a single RNA molecule, the single guide RNA (sgRNA), that maintains Cas9's DNA cleavage activity[16]. While many other CRISPR-Cas systems also mediate RNA-programmable DNA cleavage, class 2 CRISPR-Cas systems, requiring only a single Cas protein, are simpler than the multi-subunit class 1 CRISPR-Cas systems, and have been favored for genome-editing applications. The simplicity and programmable nature of the CRISPR-Cas9 system has led to its widespread adoption in biological research and therapeutic development. To date, Cas9-mediated genome editing has been demonstrated in a wide variety of cell types and organisms[19].

Beyond genome editing, programmable DNA-binding proteins can also be used to localize heterologous effector domains to specific genomic loci. Effector domains can be directly fused to ZFs and TALES, which don't possess nuclease activity in the absence of the FokI fusion. Catalytically dead variants of Cas9 (dCas9)[16,17] can similarly be fused to effector domains, and modifications to the gRNA can also be used to recruit specific effector domains through the use of sequence-specific RNA-binding proteins. DNA-programmable transcriptional activators and repressors[20-27], epigenetic modulators[28,29], and fluorophores for sequence-specific DNA labeling[30] have been developed using this strategy.

The ability of programmable DNA-binding proteins and nucleases to distinguish on-target from off-target DNA sequences is crucial for their application as research tools and potential human therapeutics. Unfortunately, all three types of programmable nucleases demonstrate some level of off-target genome modification[31-35]. Compared with other techniques that perturb biological systems, genome editing is somewhat unique in the fact that only one or two copies of the target are present in the cell. As a result, once the on-target site has been successfully modified, the continued presence of active genome-editing proteins is undesirable since sustained activity can only mediate off-target DNA modification. In light of the imperfect specificity of programmable nucleases, researchers have sought to improve specificity using a number of strategies including protein engineering and directed evolution.

4.2 Post-Translational Small-Molecule Control of Cas9 Activity

One general strategy to improve Cas9 specificity is to reduce its activity once it has had sufficient opportunity to modify the target DNA locus. Unfortunately, wild-type Cas9 nucleases are not known to be regulated by other molecules and, therefore, are used in constitutively active form. While Cas9 can be regulated at the transcriptional level through the use of inducible promoters[36-39], transcriptional control cannot limit activity to the short temporal windows that may be necessary to maximize genome-editing specificity[40,41], in contrast with the high temporal resolution of post-translational strategies that directly control protein activity.

To enable post-translational control over Cas9, we sought to engineer variants of Cas9 that can be directly controlled with a cell-permeable small molecule. Previous work in the Liu lab resulted in the development of ligand-dependent inteins that undergo protein splicing only in the presence of 4-hydroxytamoxifen (4-HT)[42]. These inteins were developed by inserting the human estrogen receptor ligand-binding domain into the *M. tuberculosis* RecA intein and evolving the resulting inactive fusion protein into a conditionally active intein that requires the presence of 4-HT. Subsequent evolution at 37 °C yielded a second-generation intein, 37R3-2, with improved splicing properties in mammalian cells[43]. We envisioned that inserting the 37R3-2 intein into Cas9 at a location that disrupts Cas9 activity until protein splicing has taken place could result in conditionally active Cas9 nucleases that are active only in the presence of 4-HT (Figure 4.2).

We genetically inserted the 4-HT-dependent intein at each of fifteen positions in Cas9 (Cys80, Ala127, Thr146, Ser219, Thr333, Thr519, Cys574, Thr622, Ser701, Ala728, Thr995, Ser1006, Ser1154, Ser1159, and Ser1274), chosen to distribute the location of the intein across the structural domains of Cas9[44-46]. Because intein splicing leaves behind a single cysteine

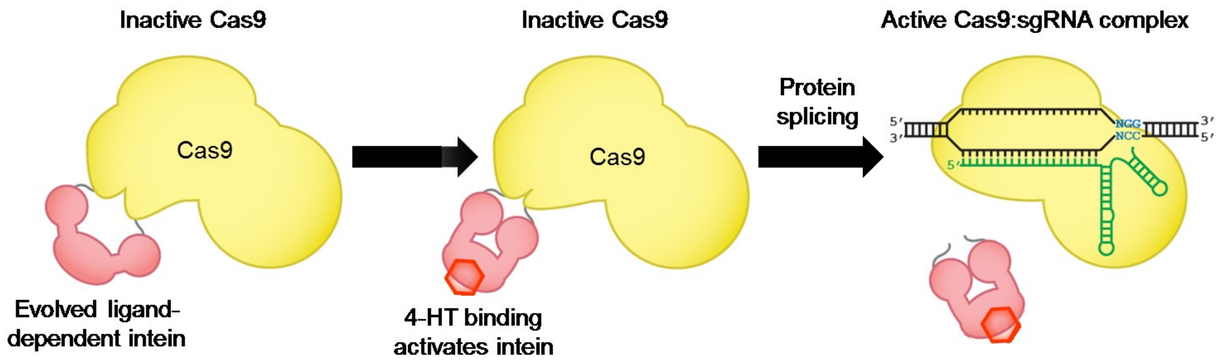

Inactive Cas9
Cas9
Evolved ligand-dependent intein
Inactive Cas9
Cas9
4-HT binding activates intein
Protein splicing
Active Cas9:sgRNA complex
5′
3′
NGG
NCC
3′
5′
5′

Figure 4.2 | Small-molecule control of Cas9 using a ligand-dependent intein. Insertion of the evolved ligand-dependent intein renders Cas9 inactive. Upon 4-HT binding, the intein undergoes conformational changes that trigger protein splicing and restore Cas9 activity.

residue, the intein was inserted in place of one Cas9 amino acid in each of the 15 candidate constructs. In addition to replacing natural cysteine amino acids, we also favored replacing alanine, serine, or threonine residues to minimize the likelihood that the resulting cysteine point mutation resulting from protein splicing would disrupt Cas9 activity. The 15 intein-Cas9 candidates were expressed in HEK293-GFP cells together with a sgRNA that targets the genomic *EGFP* locus in these cells. Twelve hours post-transfection, cells were treated with or without 1 μM 4-HT. Five days post-transfection, cells were analyzed on a flow cytometer for loss of GFP expression from Cas9-mediated *EGFP* cleavage and subsequent NHEJ-mediated indel formation.

Eight of the candidates, corresponding to intein insertion at Ala127, Thr146, Ser219, Thr333, Thr519, Cys574, Ser1006, and Ser1159, demonstrated 4-HT-dependent loss of GFP expression consistent with 4-HT-triggered Cas9 activity (Figure 4.3). Interestingly, three intein-Cas9 proteins (insertion at Ala728, Thr995, and Ser1154) showed high DNA modification rates both in the presence and absence of 4-HT, suggesting that large protein insertions at these

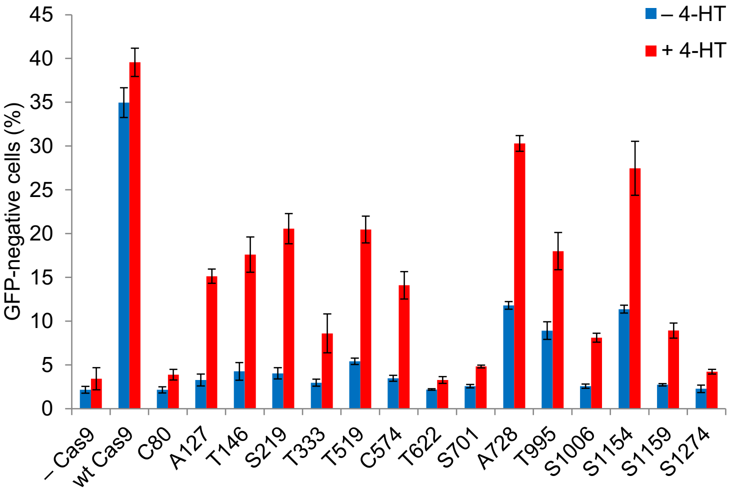


Figure 4.3 | Genomic *EGFP* disruption activity of intein-Cas9 variants. Intein-Cas9 variants were tested for their ability to mediate *EGFP* disruption in the absence or presence of 4-HT. Cas9-mediated cleavage of the *EGFP* locus followed by NHEJ-mediated indel formation leads to gene disruption and subsequent loss of GFP expression. Error bars reflect the standard deviation of three biological replicates.

positions do not significantly inhibit nuclease activity, or that the intein lost its 4-HT dependence due to context-dependent conformational perturbations. We speculate that it may be possible to engineer split Cas9 variants by dividing the protein at these locations, given their tolerance of a 413-residue insertion. The lack of nuclease activity of the remaining four Cas9-inteins (insertion at Cys80, Thr622, Ser701, and Ser1274) in the presence or absence of 4-HT could result from the inability of the intein to splice in those contexts, the inability of Cas9 to refold properly following splicing, or intolerance of replacement of native threonine or serine residues with cysteine. We pursued two intein-Cas9 variants corresponding to insertion at Ser219 (S219) and Cys574 (C574) (Figure 4.4). These two variants combined high activity in the presence of 4-HT and low activity in the absence of 4-HT.

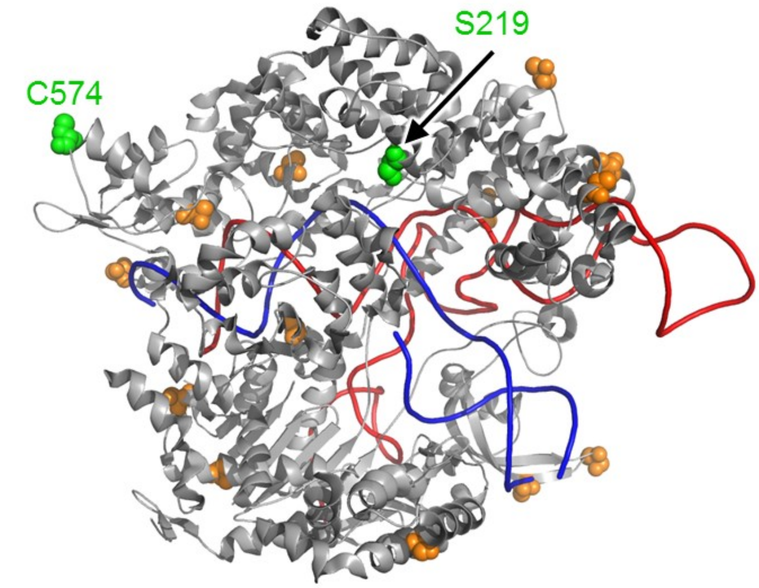


Figure 4.4 | Cas9 sites of intein insertion. Intein insertion sites (orange and green) are highlighted on the structure of Cas9 (PDB 4UN3). Intein-Cas9(S219) and intein-Cas9(C574) (green) were used in subsequent experiments.

4.3 Genome Modification Specificities of Intein-Cas9s

To evaluate the genome modification specificity of conditionally active Cas9 variants, we expressed intein-Cas9(S219), intein-Cas9(C574), and wild-type Cas9 in HEK293-GFP cells together with each of three previously described[47] sgRNAs that target the well-studied *EMX*, *VEGF*, and *CLTA* genomic loci. We assayed these Cas9:sgRNA combinations in human cells for their ability to modify the three on-target loci as well as 11 known off-target genomic sites (Table 4.1)[33,35,48,49]. Cells were treated with or without 1 μM 4-HT during transfection, and after 12 hours the media was replaced with fresh media lacking 4-HT. We observed no cellular toxicity arising from 12 or 60 hours of treatment with 1 μM 4-HT in untransfected or transfected HEK293 cells (Figure 4.5). Genomic DNA was isolated 60 hours post-transfection and analyzed by high-throughput DNA sequencing

Table 4.1 | Genomic on-target and off-targets sites of the *EMX*, *VEGF*, and *CLTA* sgRNAs. On-target and 11 known off-target substrates of Cas9:sgRNAs that target sites in *EMX*, *VEGF,* and *CLTA* are listed. Mutations from the on-target sequence are shown in red lower case. Protospacer-adjacent motifs (PAMs) are shown in blue.

EMX On	GAGTCCGAGCAGAAGAAGAAGGG
EMX Off 1	GAGgCCGAGCAGAAGAAagACGG
EMX Off 2	GAGTCCtAGCAGgAGAAGAAGaG
EMX Off 3	GAGTCtaAGCAGAAGAAGAAGaG
EMX Off 4	GAGTtaGAGCAGAAGAAGAAAGG
VEGF On	GGGTGGGGGGAGTTTGCTCCTGG
VEGF Off 1	GGaTGGaGGGAGTTTGCTCCTGG
VEGF Off 2	GGGaGGGtGGAGTTTGCTCCTGG
VEGF Off 3	cGGgGGaGGGAGTTTGCTCCTGG
VEGF Off 4	GGGgaGGGGaAGTTTGCTCCTGG
CLTA On	GCAGATGTAGTGTTTCCACAGGG
CLTA Off 1	aCAtATGTAGTaTTTCCACAGGG
CLTA Off 2	cCAGATGTAGTaTTcCCACAGGG
CLTA Off 3	ctAGATGaAGTGcTTCCACATGG

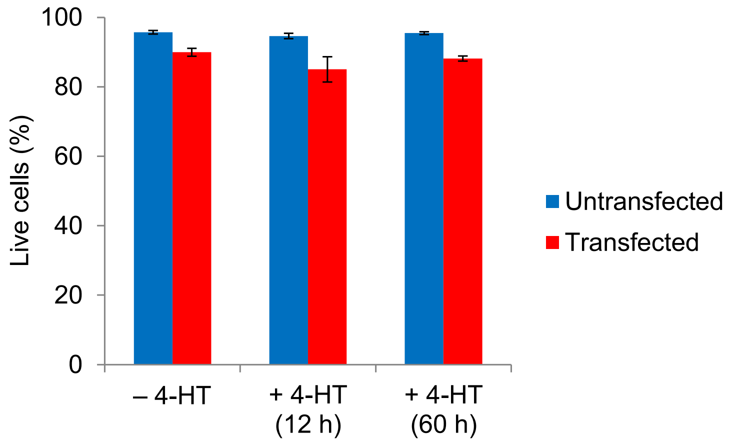


Figure 4.5 | Effect of 4-HT on cellular toxicity. Untransfected HEK293-GFP stable cells, and cells transfected with intein-Cas9(S219) and sgRNA expression plasmids, were treated with or without 4-HT. The live cell population was determined by flow cytometry. Error bars reflect the standard deviation of six technical replicates.

Overall on-target genome modification frequency of intein-Cas9(S219) and intein-Cas9 (C574) in the presence of 1 μM 4-HT was similar to that of wild-type Cas9 (Figure 4.6). On-target modification frequency in the presence of 4-HT was 3.4- to 7.3-fold higher for intein-Cas9(S219), and 3.6- to 9.6-fold higher for intein-Cas9(C574), than in the absence of 4-HT, whereas modification efficiency for wild-type Cas9 was 1.2- to 1.8-fold lower in the presence of 4-HT. Both intein-Cas9 variants exhibited a low level of background activity in the absence of 4-HT, consistent with previous reports[42,43,50]. Western blot analysis of intein-Cas9(S219) from transfected HEK293 cells confirmed the presence of spliced product at the earliest assayed time point (4 hours) following 4-HT treatment; no spliced product was detected in the absence of 4-HT (Figure 4.7). Together, these results indicate that intein-Cas9(S219) and intein-Cas9(C574) are slightly less active than wild-type Cas9 in the presence of 4-HT, likely due to incomplete splicing but much less active in the absence of 4-HT.

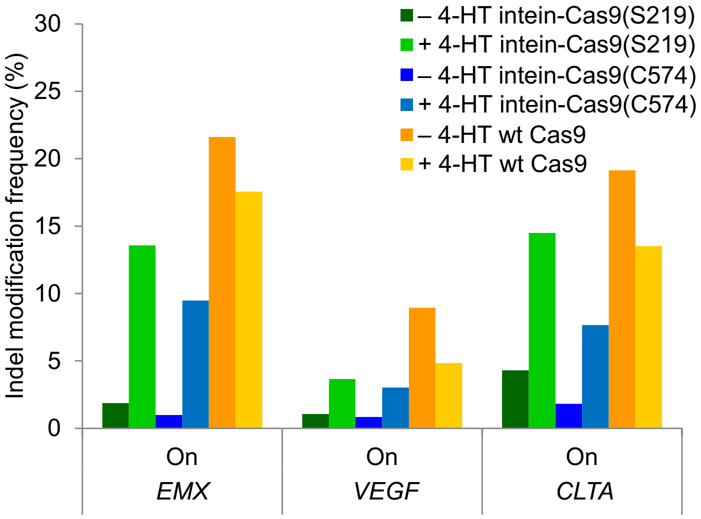


Figure 4.6 | Genomic on-target DNA modification by intein-Cas9(S219) and intein-Cas9(C574). Indel modification frequency from high-throughput DNA sequencing of amplified genomic on-target sites by intein-Cas9(S219), intein-Cas9(C574), and wild-type Cas9 in the absence or presence of 4-HT. *P*-values are $< 10^{-15}$ for the Fisher exact test (one-sided up) on comparisons of indel modification frequency in the presence versus the absence of 4-HT for intein-Cas9(S219) and intein-Cas9(C574).

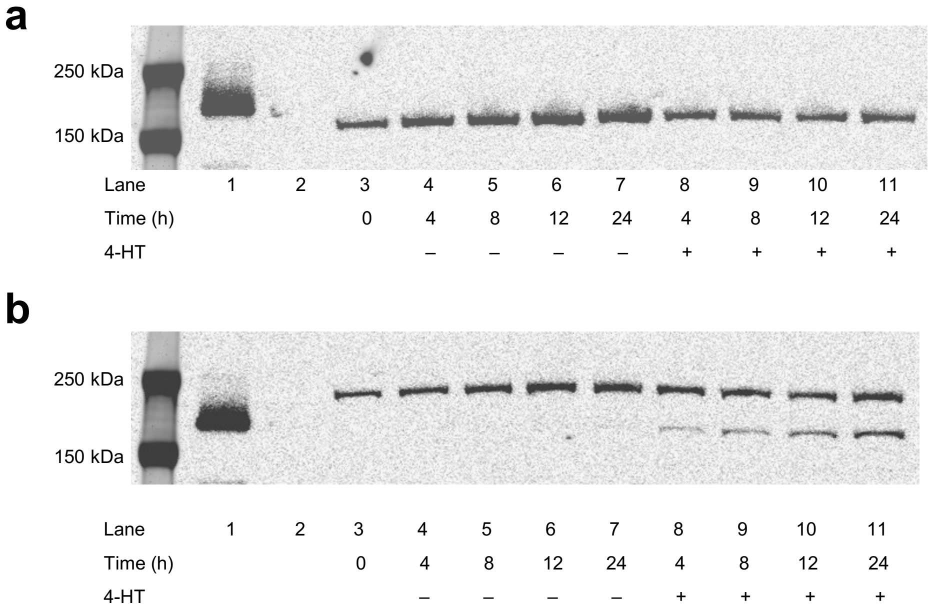

a
250 kDa
150 kDa
Lane 1 2 3 4 5 6 7 8 9 10 11
Time (h) 0 4 8 12 24 4 8 12 24
4-HT – – – – + + + +
b
250 kDa
150 kDa
Lane 1 2 3 4 5 6 7 8 9 10 11
Time (h) 0 4 8 12 24 4 8 12 24
4-HT – – – – + + + +

Figure 4.7 | Western blot analysis of intein splicing. HEK293-GFP stable cells were transfected with (**a**) wild-type Cas9 or (**b**) intein-Cas9(S219) expression plasmid and treated with or without 4-HT. Cells were lysed and pooled from three technical replicates at 4, 8, 12, or 24 hours after 4-HT treatment. Lanes 1 and 2 contain purified dCas9-VP64-3×FLAG protein and lysate from untransfected HEK293-GFP cells, respectively.

High-throughput sequencing of 11 previously described off-target sites that are modified by wild-type Cas9:sgRNA complexes targeting the *EMX*, *VEGF*, and *CLTA* loci revealed that both intein-Cas9 variants when treated with 4-HT for 12 hours exhibit substantially improved specificity compared to that of wild-type Cas9 (Figure 4.8). On-target:off-target indel modification ratios for both intein-Cas9 variants were on average 6-fold higher, and as much as 25-fold higher, than that of wild-type Cas9 (Figure 4.9). In the absence of 4-HT, the genome modification specificity of both intein-Cas9 variants was on average 14-fold higher than that of wild-type Cas9 in the absence of 4-HT, presumably resulting from the much lower activity of the intein-Cas9 variants in the absence of 4-HT[33-35].

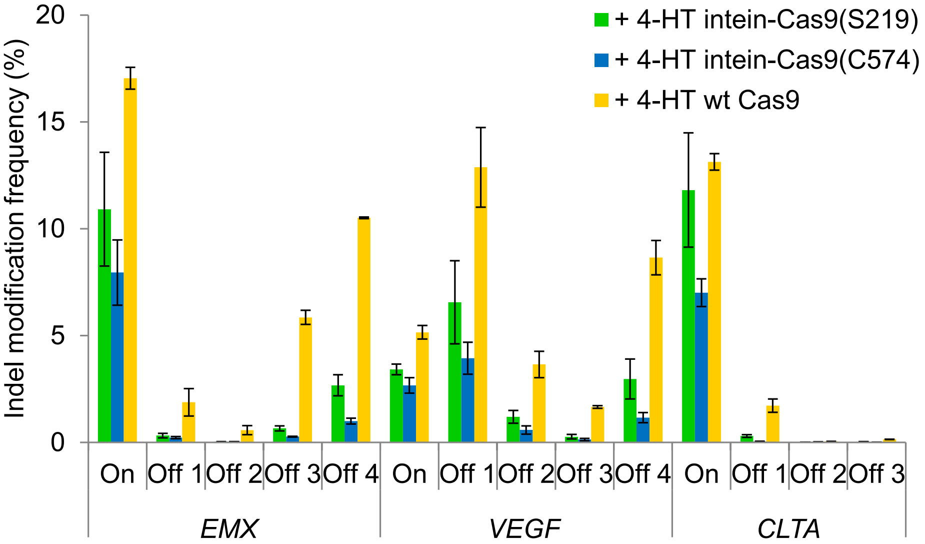


Figure 4.8 | Genomic off-target DNA modification by intein-Cas9(S219) and intein-Cas9(C574). Indel modification frequency from high-throughput DNA sequencing of amplified genomic on-target sites and off-target sites ("Off 1-Off 4") by intein-Cas9(S219), intein-Cas9(C574), and wild-type Cas9 in the presence of 4-HT. *P*-values are < 0.005 for the Fisher exact test (one-sided down) on all pairwise comparisons within each independent experiment of off-target modification frequency between either intein-Cas9 variant in the presence of 4-HT versus that of wild-type Cas9 in the presence of 4-HT. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Error bars reflect the range of two independent experiments conducted on different days.

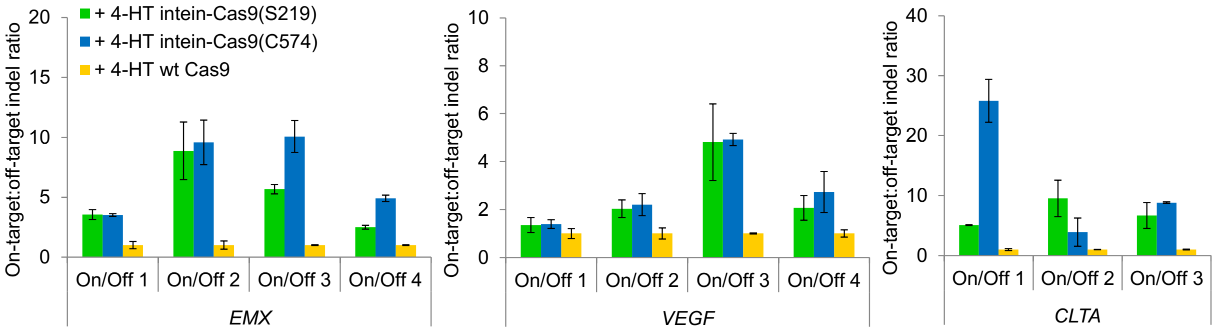


Figure 4.9 | DNA modification specificity of intein-Cas9(S219) and intein-Cas9(C574). DNA modification specificities, defined as on-target/off-target indel modification frequency ratio, in the presence of 4-HT were calculated for intein-Cas9(S219) and intein-Cas9(C574), and normalized to wild-type Cas9. Error bars reflect the range of two independent experiments conducted on different days.

Since intein-Cas9s can result in slightly lower on-target modification rates compared to wild-type Cas9, we sought to verify that the improvements in specificity among the intein-Cas9s were not simply a result of reduced activity. Both on- and off-target activity of Cas9 has been shown to be dependent on the amount of Cas9 expression plasmid transfected[33-35]. By transfecting lower amounts of the wild-type Cas9 expression plasmid, we compared intein-Cas9s with wild-type Cas9 under conditions that result in very similar levels of on-target modification. To minimize potential differences in transfection efficiency, we supplemented with a plasmid that does not express Cas9 so that the same total amount of plasmid DNA was transfected into each sample. High-throughput sequencing revealed that wild-type Cas9 shows slightly improved specificity, as expected, as the on-target cleavage rate is reduced. The intein-Cas9 variants, however, remain substantially more specific than wild-type Cas9 at similar on-target DNA cleavage rates (Figures 4.10 and Figure 4.11). For example, intein-Cas9(C574) and wild-type Cas9 (80 ng) have virtually identical on-target DNA cleavage rates (both 6.4%) at the *EMX* locus but all four off-target sites are modified at an average of 4-fold lower frequencies ($P < 1 \times 10^{-13}$) by intein-Cas9(C574) than by wild-type Cas9. These findings indicate that specificity improvements of intein-Cas9 variants do not simply arise from differences in overall genome editing activity.

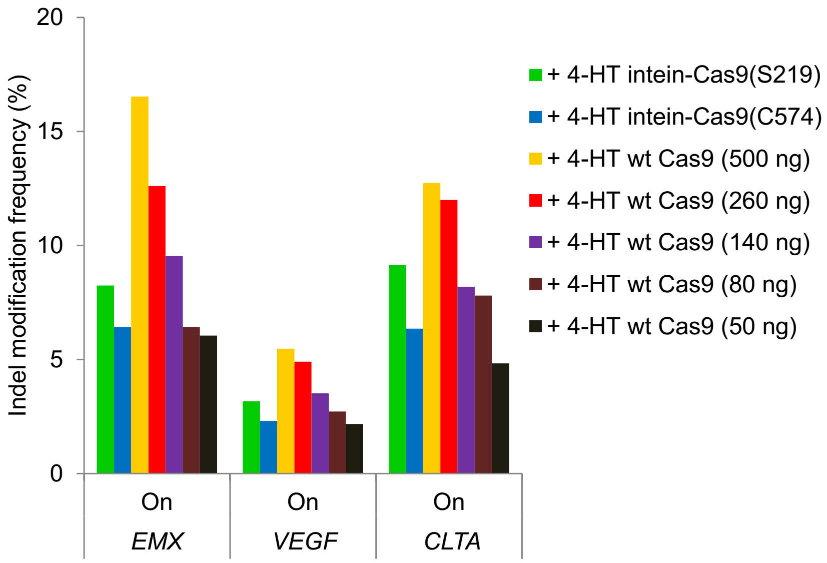


Figure 4.10 | Genomic on-target DNA modification comparisons with titrated amounts of wild-type Cas9. Indel modification frequency from high-throughput DNA sequencing of amplified genomic on-target sites by intein-Cas9(S219), intein-Cas9(C574), and different amounts of transfected wild-type Cas9 expression plasmid (specified in parentheses) in the presence of 4-HT.

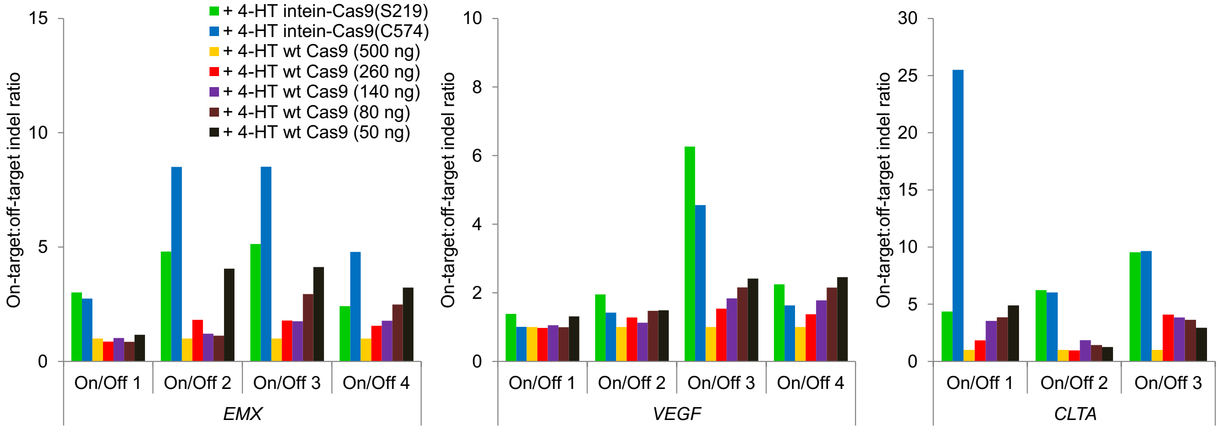


Figure 4.11 | DNA modification specificity comparisons with titrated amounts of wild-type Cas9. DNA modification specificities, defined as on-target/off-target indel modification frequency ratio, in the presence of 4-HT were calculated for intein-Cas9(S219), intein-Cas9(C574), and different amounts of transfected wild-type Cas9 expression plasmid (specified in parentheses). DNA modification specificities were normalized to wild-type Cas9 (500 ng).

4.4 Eliminating Intein Activation by β-estradiol

Intein 37R3-2 can be activated by other estrogen receptor modulators. To enable intein-Cas9 applications in which endogenous β-estradiol is present, we inserted into the estrogen receptor ligand-binding domain a point mutation (G521R) that renders the domain more specific for 4-HT[51]. This mutation slightly reduces affinity for 4-HT but almost abolishes affinity for β-estradiol. The addition of this mutation to intein-Cas9(S219) eliminates the ability of β-estradiol to trigger Cas9 activity (Figure 4.12).

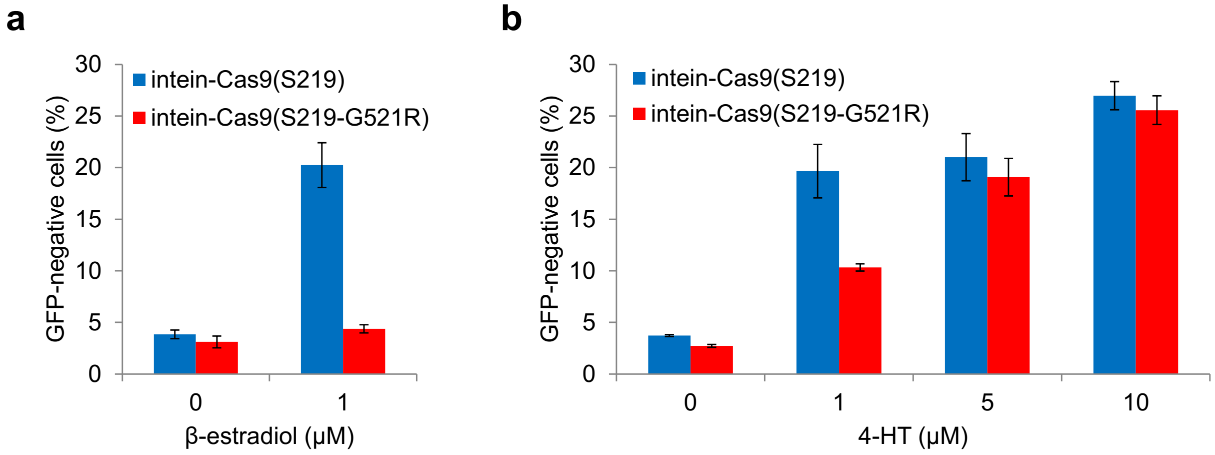


Figure 4.12 | Eliminating intein activation by β-estradiol. Genomic *EGFP* disruption activity of intein-Cas9(S219) and intein-Cas9(S219-G521R) in the presence of (**a**) β-estradiol or (**b**) 4-HT. Error bars reflect the standard deviation of three technical replicates.

4.5 Conclusion

The intein-Cas9 variants developed here enable small-molecule control of Cas9 function, thereby enhancing genome-modification specificity. These findings strongly suggest that limiting the opportunity of genome-editing agents to modify off-target loci following a period of activity sufficient to induce desired levels of on-target modification results in substantial specificity improvements. We anticipate that the use of ligand-dependent Cas9 variants will provide greater control over genomic modification efficiencies and specificities than is currently achievable with constitutively active or transcriptionally regulated genome editing. In principle, this approach could synergize with other specificity-augmenting strategies such as using truncated guide RNAs[49], paired Cas9 nickases[20,48], or FokI-dCas9 fusions[47,52]. This approach could also be applied to other genome engineering proteins to enable, for example, small-molecule control of TALE-based or Cas9-mediated transcriptional regulators.

4.6 Methods

Cas9, intein-Cas9, and sgRNA expression plasmids. A plasmid encoding the human codon-optimized *S. pyogenes* Cas9 nuclease with an NLS and 3×FLAG tag (Addgene plasmid 43861)[33] was used as the wild-type Cas9 expression plasmid. Intein 37R3-2 was subcloned at the described positions into the wild-type Cas9 expression plasmid using USER cloning (New England Biolabs) as previously described[53]. sgRNA expression plasmids used in this study have been previously described[47].

Modification of genomic GFP. HEK293-GFP stable cells (GenTarget), which constitutively express Emerald GFP, served as the reporter cell line. Cells were maintained in “full serum media": Dulbecco's Modified Eagle's Media plus GlutaMax (Life Technologies) with 10% (vol/vol) FBS and penicillin/streptomycin (1×, Amresco). 5 × 10^4 cells were plated on 48-well collagen-coated Biocoat plates (Becton Dickinson). 16-18 h after plating, cells were transfected with Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Briefly, 1.5 μL of Lipofectamine 2000 was used to transfect 650 ng of total plasmid: 500 ng Cas9 expression plasmid, 125 ng sgRNA expression plasmid, and 25 ng near-infrared iRFP670 expressing plasmid (Addgene plasmid 45457)[54]. 12 h after transfection, the media was replaced with full serum media, with or without 4-HT (1 μM, Sigma-Aldrich). The media was replaced again 3-4 days after transfection. Five days after transfection, cells were trypsinized and resuspended in full serum media and analyzed on a C6 flow cytometer (Accuri) with a 488-nm laser excitation and 520-nm filter with a 20-nm band pass. Transfections and flow cytometry measurements were performed in triplicate.

High-throughput DNA sequencing of genome modifications. HEK293-GFP stable cells were transfected with plasmids expressing Cas9 (500 ng) and sgRNA (125 ng) as described above. For treatments in which a reduced amount of wild-type Cas9 expression plasmid was transfected, pUC19 plasmid was used to bring the total amount of plasmid to 500 ng. 4-HT (1 μM final), where appropriate, was added during transfection. 12 h after transfection, the media was replaced with full serum media without 4-HT. Genomic DNA was isolated and pooled from three biological replicates 60 h after transfection using a previously described[47] protocol with a DNAdvance Kit (Agencourt). 150 ng or 200 ng of genomic DNA was used as a template to amplify by PCR the on-target and off-target genomic sites with flanking HTS primer pairs previously described[47]. PCR products were purified using RapidTips (Diffinity Genomics) and quantified using the PicoGreen dsDNA Assay Kit (Invitrogen). Purified DNA was PCR amplified with primers containing sequencing adaptors, purified with a MinElute PCR Purification Kit (Qiagen) and AMPure XP PCR Purification (Agencourt). Samples were sequenced on a MiSeq high-throughput DNA sequencer (Illumina), and sequencing data was analyzed as described previously[35].

Western blot analysis of intein splicing. HEK293-GFP stable cells were transfected with 500 ng Cas9 expression plasmid and 125 ng sgRNA expression plasmid. 12 h after transfection, the media was replaced with full serum media, with or without 4-HT (1 μM). Cells were lysed and pooled from three technical replicates 4, 8, 12, or 24 h after 4-HT treatment. Samples were run on a Bolt 4-12% Bis-Tris gel (Life Technologies). An anti-FLAG antibody (Sigma-Aldrich F1804) and an anti-mouse 800CW IRDye (LI-COR) were used to visualize the gel on an IR imager (Odyssey).

Statistcal analysis. Statistical tests were performed as described in the figure captions. All *p*-values were calculated with the R software package. *p*-values for the Fisher exact test were calculated using the fisher.test function, with a one-sided alternative hypothesis (alternative = "greater" or alternative = "less", as appropriate). Upper bounds on *p*-values that are close to zero were determined manually. The Benjamini-Hochberg adjustment was performed using the R function p.adjust (method = "fdr").

4.7 References

1. Lukacsovich, T., Yang, D. & Waldman, A.S. Repair of a specific double-strand break generated within a mammalian chromosome by yeast endonuclease I-SceI. *Nucleic acids research* **22**, 5649-5657 (1994).

2. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* **14**, 8096-8106 (1994).

3. Choulika, A., Perrin, A., Dujon, B. & Nicolas, J.F. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol Cell Biol* **15**, 1968-1973 (1995).

4. Miller, J., McLachlan, A.D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal* **4**, 1609-1614 (1985).

5. Parraga, G. et al. Zinc-dependent structure of a single-finger domain of yeast ADR1. *Science* **241**, 1489-1492 (1988).

6. Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A. & Wright, P.E. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**, 635-637 (1989).

7. Pavletich, N.P. & Pabo, C.O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809-817 (1991).

8. Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512 (2009).

9. Moscou, M.J. & Bogdanove, A.J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).

10. Deng, D. et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720-723 (2012).

11. Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J. & Stoddard, B.L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716-719 (2012).

12. Kim, Y.G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1156-1160 (1996).

13. Christian, M. et al. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757-761 (2010).

14. Li, T. et al. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **39**, 359-372 (2011).

15. Miller, J.C. et al. A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143-148 (2011).

16. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

17. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579-2586 (2012).

18. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011).

19. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology* **32**, 347-355 (2014).

20. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833-838 (2013).

21. Maeder, M.L. et al. CRISPR RNA-guided activation of endogenous human genes. *Nature methods* **10**, 977-979 (2013).

22. Konermann, S. et al. Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472-476 (2013).

23. Qi, L.S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).

24. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic acids research* **41**, 7429-7437 (2013).

25. Gilbert, L.A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

26. Cheng, A.W. et al. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res* **23**, 1163-1171 (2013).

27. Perez-Pinera, P. et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature methods* **10**, 973-976 (2013).

28. Hilton, I.B. et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology* **33**, 510-517 (2015).

29. Kearns, N.A. et al. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nature methods* **12**, 401-403 (2015).

30. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479-1491 (2013).

31. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**, 765-770 (2011).

32. Guilinger, J.P. et al. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nature methods* **11**, 429-435 (2014).

33. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology* **31**, 822-826 (2013).

34. Hsu, P.D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827-832 (2013).

35. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31**, 839-843 (2013).

36. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).

37. Gonzalez, F. et al. An iCRISPR platform for rapid, multiplexable, and inducible genome editing in human pluripotent stem cells. *Cell stem cell* **15**, 215-226 (2014).

38. Dow, L.E. et al. Inducible in vivo genome editing with CRISPR-Cas9. *Nature biotechnology* **33**, 390-394 (2015).

39. Aubrey, B.J. et al. An inducible lentiviral guide RNA platform enables the identification of tumor-essential genes and tumor-promoting mutations in vivo. *Cell Rep* **10**, 1422-1432 (2015).

40. Zuris, J.A. et al. Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature biotechnology* (2014).

41. Pruett-Miller, S.M., Reading, D.W., Porter, S.N. & Porteus, M.H. Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels. *PLoS genetics* **5**, e1000376 (2009).

42. Buskirk, A.R., Ong, Y.C., Gartner, Z.J. & Liu, D.R. Directed evolution of ligand dependence: small-molecule-activated protein splicing. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10505-10510 (2004).

43. Peck, S.H., Chen, I. & Liu, D.R. Directed evolution of a small-molecule-triggered intein with improved splicing properties in mammalian cells. *Chemistry & biology* **18**, 619-630 (2011).

44. Jinek, M. et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).

45. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014).

46. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* (2014).

47. Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature biotechnology* **32**, 577-582 (2014).

48. Ran, F.A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-1389 (2013).

49. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology* **32**, 279-284 (2014).

50. Yuen, C.M., Rodda, S.J., Vokes, S.A., McMahon, A.P. & Liu, D.R. Control of transcription factor activity and osteoblast differentiation in mammalian cells using an evolved small-molecule-dependent intein. *Journal of the American Chemical Society* **128**, 8939-8946 (2006).

51. Danielian, P.S., White, R., Hoare, S.A., Fawell, S.E. & Parker, M.G. Identification of residues in the estrogen receptor that confer differential sensitivity to estrogen and hydroxytamoxifen. *Molecular endocrinology* **7**, 232-240 (1993).

52. Tsai, S.Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology* **32**, 569-576 (2014).

53. Badran, A.H. et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **533**, 58-63 (2016).

54. Shcherbakova, D.M. & Verkhusha, V.V. Near-infrared fluorescent proteins for multicolor in vivo imaging. *Nature methods* **10**, 751-754 (2013).