Altering Parameters of CRISPR/Cas9 Mediated Genome Editing in *Escherichia coli* Helps Identify Sources of Variation in Editing Efficiency

By:

Reilly Fankhauser

B.A. Biochemistry, University of Colorado Boulder, May 2018

A thesis submitted to the Faculty of the Department of Chemistry and Biochemistry of the University of Colorado in partial fulfillment of the requirement for graduation with honors for the degree of Biochemistry

Defense date: April 6th, 2018

SIGNATURE PAGE

This thesis entitled:

Altering Parameters of CRISPR/Cas9 Mediated Genome Editing in *Escherichia coli* Helps Identify Sources of Variation in Editing Efficiency

written by Reilly Fankhauser

has been approved for the Department of Chemistry and Biochemistry

____
(Type committee chair name here)

____
(Type committee member name here)

Date_______________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.

ABSTRACT

Fankhauser, Reilly Garrett (B.A., Biochemistry, Department of Chemistry and Biochemistry)

Altering Parameters of CRISPR/Cas9 Mediated Genome Editing in *Escherichia coli* Helps Identify Sources of Variation in Editing Efficiency

Thesis directed by Dr. Deborah Wuttke

CRISPR/Cas9, originating from a bacterial immune system, has been identified as a powerful tool for genome modification (Jinek, et al., 2012). Historically, analyzing genotype-phenotype relationships have been limited by DNA synthesis and sequencing technology. Recent developments in genome editing (CRISPR/Cas9) coupled to improvements in synthesis and sequencing allows mapping these relationships at an entirely new scale. A technique novel to the Gill lab for genome modification that harnesses the CRISPR-Cas system is “CRISPR Enabled Trackable Genomic Engineering” or “CREATE” technology. CREATE technology is one such strategy for simultaneously making hundreds of thousands of single genomic edits, then tracking the resultant phenotypes. In this study, I alter parameters of the CREATE mechanism to elucidate the cause of variability in editing efficiency between different editing constructs. I demonstrate that editing efficiency is impacted by guide RNA toxicity, genomic position, expression levels of Cas9, availability of template for homologous recombination, and expression levels of phage-based recombination proteins. I find that high toxicity is related to high editing efficiency. The position of the gene being edited in the bacterial genome impacts editing efficiency in a pattern dependent on guide toxicity. Expression levels of Cas9 also affect editing efficiency in a guide toxicity-dependent fashion. Additional results studying recombination suggest increasing intracellular abundance of donor DNA can lead to increased editing efficiency, in a genomic position-specific manner. A final finding was that expressing recombination proteins using a constitutive promoter replicated the behavior of the traditional lambda Red expression system, opening up possibilities for future modifications.

ACKNOWLEDGEMENTS

Thesis advisors:
Alaksh Chourdhury, Ph.D candidate, Chemical and Biological Engineering
Dr. Deborah Wuttke, Department of Chemistry and Biochemistry
Dr. Joel Kaar, Chemical and Biological Engineering
Dr. Ryan Gill, Chemical and Biological Engineering

Honors Council Co-chairs:
Dr. Xuedong Liu, Department of Chemistry and Biochemistry
Dr. Jeffrey Cameron, Department of Chemistry and Biochemistry

Work Performed in Gill Laboratory: Genome Engineering

With funding provided by:
Colorado Center for Biorefining and Bioproducts (C2B2)
Undergraduate Research Opportunities Program (UROP)

CONTENTS

TABLES

FIGURES

LIST OF ABBREVIATIONS, SYMBOLS, AND NOMENCLATURE

Table 1. *Abbreviations, Symbols and Nomenclature.*

Term	**Meaning**
3HP	3-hydroxy propionic acid – a molecule used in synthesis of biodegradable polymers.
Cas9	CRISPR-associated protein 9
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
DSB	Double-strand break
dsDNA	Double stranded DNA
E.E.	Editing Efficiency $Editing\ Efficiency = \frac{number\ of\ cells\ successfully\ edited}{total\ number\ of\ cells}$
H.R.	Homologous recombination
HDR	Homology-directed Repair
Indel	Insertion or deletion
λ Red recombination	Phage-derived machinery for facilitating homologous recombination (HR)
NHEJ	Non-homologous end joining
O_R1 - O_R3, O_L1 - O_L3	Operator sites regulating expression of lambda Red recombination proteins isolated from bacteriophage.
ORF	Open Reading Frame
OriC	Chromosomal origin of replication.
PAM	Protospacer adjacent motif
PBS	Phosphate Buffered Saline
RCF	Relative centrifugal force
sgRNA	Single guide RNA
ssDNA	Single Stranded DNA
T.E.	Transformation Efficiency $Transformation\ Efficiency = \frac{CFU\ of\ E.coli}{(\mu g\ DNA)(mL\ Transformation)}$
tracrRNA	Trans-activating CRISPR RNA
TSS	Transcription Start Site

CHAPTER 1

INTRODUCTION

MOTIVATION

As the global energy crisis and concerns about carbon emissions and global warming progress, the demand for cleaner energy sources will increase. Ethanol and higher alcohols are promising alternatives to fossil fuels as a source of clean, carbon neutral energy. Currently, the production of ethanol is centered on the digestion of corn sugars by yeast and bacteria: a controversial approach because food is being used as a source for fuel. Cellulose from corn stalks and other sources is a great alternative to corn because it is the most abundant organic polymer on earth and is indigestible by humans (Klemm, Heublein, Fink, & Bohn, 2005). One major obstacle to this process is that cellulose is challenging to break down into simpler sugars that can then be converted into ethanol by yeast or bacteria (Amin, et al., 2017). Therefore, cellulose must be pretreated by physical, chemical, or biological methods: a thermodynamically unfavorable process (Amin, et al., 2017). Dr. Ryan Gill proposes that advancements in genetic engineering of yeast and bacteria will eventually allow the cells to digest cellulose from corn stalks and convert it into ethanol all in one reaction mixture, thus eliminating pretreatment. However, methods of genome modification must be optimized before this approach can work. The most widely used technique for genome modification utilizes Cas9 endonucleases.

A 2012 paper described how CRISPR-Cas systems could be adapted to make guided edits to bacterial genomes (Jinek, et al., 2012). This initiated a race to find applications for the technology in bacteria, plants, and animals. A technique novel to the Gill lab for genome modification that harnesses the older CRISPR-Cas system is “CRISPR Enabled Trackable Genomic Engineering” or “CREATE” technology. CREATE technology allows us to make multiple single edits across the entire genome simultaneously to make large mutant libraries (the population of all organisms modified at the genomic level) and provides the ability to track the genotypes on a population-wide scale. This trackable editing allows for the mapping of thousands of sequence-activity relationships at a time, under a variety of selective pressures. Essentially, once edits are made to specific regions of the genome (at promoter sites, on individual proteins, or on entire protein systems, etc.), selective pressures can be applied to the cells to determine what effect these regions of the genome had on the cell and its overall fitness. The precision of these edits to the genome range from alteration of single proteins to genome wide targeting of protein systems.

Some recent applications of CREATE Technology include:

1) Site saturation mutagenesis for protein engineering with single amino acid resolution, and genome wide fitness mapping for identification of tolerance to stress and antibiotic resistance (Garst, et al., 2017).
2) Construction and testing of approximately 1000 designer strains for increased yield and productivity during isopropanol production in *E. coli*. A maximum productivity of 0.62 g/L/h and a maximum yield of 0.75 mol/mol were observed. (Liang, et al., 2017).
3) Investigating a library of approximately 40,000 mutations spanning 30 different genes to increase 3-hydroxy propionic acid (3HP) production by 7-8 fold. 3HP is used in production of biodegradable polymers (Liu, et al., 2018).

Despite these successful demonstrations of CREATE technology, there are some issues that must be addressed in order to reach its full potential. Addressing these issues is the motivation for this study. It is not well understood what factors contribute to variation in editing efficiency (the number of edits that were made, out of the number of edits that were intended to be made) across individual editing constructs. This variation is problematic for two reasons. The first reason is that despite designing a large and diverse library of mutations, very few (~1%) of them are actually introduced into the genome. This decreases the heterogeneity of the starting population and therefore decreases the breadth of mutations that were intended to be studied. The second reason why this is problematic is that it convolutes the process of data analysis.

Addressing the source of these two issues gives rise to the aim for this study: altering parameters of the CREATE mechanism to probe what factors contribute to the variability in editing efficiency. The implications of this study are that the results can be used to optimize future CREATE experiments, which could lead to improved techniques for the sustainable production of fuels, commodity chemicals, and pharmaceuticals. This work is a subset of the Gill lab's research emphasis studying the creation, development, and demonstration of novel genome-engineering tools.

SCOPE OF STUDY AND LIMITATIONS

The goal of this study was to alter parameters of the CREATE mechanism to probe what factors contribute to the variability in editing efficiency. Using the primary literature, I selected five different aspects of the CREATE mechanism that were promising targets for investigation.

1) How does guide RNA toxicity impact editing efficiency?
2) The cut: how does the genomic position of the gene being targeted impact editing efficiency?
3) The cut: how does the expression level of Cas9 impact editing efficiency, and can increasing expression of Cas9 increase editing efficiency?
4) Recombination, the paste: how does altering the quantity of available homology arm impact editing efficiency?

5) Recombination, the paste: What do Gam, Exo, and RecA proteins contribute to toxicity and editing efficiency, and can altering the expression levels of recombination proteins lead to increased editing efficiency?

Additional research opportunities that could have been included in this study but were not, due to time and financial constraints, include:

1) Changing the growth medium used during recovery to slow down the rate of replication after transformation. Altering the growth rate could impact the chromosome copy number, which could help to explain the observed variation in editing efficiency with position.
2) Identifying the cause of the decrease in editing efficiency with increased expression of Cas9 in group one guides. The hypothesis of off target activity deserves further investigation.
3) Validations of protein expression for both Cas9 and the pSIM5 variants using western blots, or other means for quantifying expression levels. Evidence for increased expression was largely empirical in this study.
4) The future studies section gives a comprehensive list of potential research directions that could be pursued based on the findings of this study.

ARRANGEMENT OF THE THESIS

The motivation for this study is given in the introduction. Chapter two begins with an introduction to CRISPR-Cas9 mediated genome modification, followed by a summary of the CREATE mechanism. The background section is designed to highlight the logic behind choosing what aspects of the CREATE mechanism would make good targets for modification to determine sources of variability in editing efficiency. Chapter three is the methods section. This section explains how each of the major objectives were carried out. Illustrations are used to help elucidate the workflow in various experiments. Chapter four is the results and conclusions section. The chronology of the conclusion follows the order of the thesis objectives listed above, starting with the relationship between editing efficiency and toxicity, followed by the two sections studying the cut, and ending with the two sections studying the paste. The conclusions and future studies section provide answers to the questions introduced above. The supplements section includes verifications of some hypotheses that diverged from the main storyline, including a study contrasting the phenotype of gene repression due to Cas9 binding (but not creating a DSB and facilitating recombination) against the phenotype exhibited by successful gene editing following homologous recombination. Also included in the supplements is a data table giving information about all the editing constructs used in this study, and the primer sequences used in PCR experiments.

CHAPTER 2

BACKGROUND

CRISPR/CAS9 AS A TOOL FOR GENOMIC MODIFICATION

CRISPR/Cas-9 technology is a system for genome engineering that arose from a natural defense mechanism in the bacterial immune system (Barrangou, et al., 2007). In this immune system, invading DNA, from bacteriophages for instance, is cut into fragments and incorporated into the bacterial genome at CRISPR loci (Barrangou, et al., 2007). These loci are transcribed into crRNA which is used to guide endonucleases, like the cas9 protein, using sequence complementarity (Barrangou, et al., 2007). Cas9 scavenges for protospacer adjacent motif (PAM, a canonical NGG sequence) then attempts to anneal the 20 nucleotide long crRNA to the sequence downstream of the PAM (Jiang, Bikard, Cox, Zhang, & Marrafini, 2013). If the crRNA exhibits sufficient complementarity to the target site, the endonuclease/crRNA complex makes a double-strand break (DSB), rendering the invading DNA ineffective (Barrangou, et al., 2007). In the absence of a PAM sequence even full complementarity between crRNA and the DNA of interest will not cause the Cas9 protein to make a DSB (Barrangou, et al., 2007).

This natural bacterial immune response has proven an invaluable tool for targeting specific regions of the genome and making edits. The two general components of a cas9 endonuclease are the effector domain and the DNA-binding domain (Wang, La Russa, & Qi, 2016). The effector domain carries out DNA cleavage and the DNA-binding domain controls the endonuclease's ability to recognize a specific sequence and bind to it (Wang, La Russa, & Qi, 2016). The introduction of a double-strand break in the host genome would ordinarily lead to cell death, but the SOS response in cells activates machinery to repair damaged DNA, ultimately causing the cell to escape death (Voet & Voet, 2011). There are two mechanisms through which double-strand breaks can be repaired (Voet & Voet, 2011). The first is non-homologous end joining (NHEJ) and the other is homology directed repair or HDR. NHEJ can cause mutations at target sites like random insertions, deletions, or reading frame shifts. HDR incorporates donor DNA, in the case of CREATE coming from the engineered editing cassette, with sequence homology at the ends into the host genome; this leads to sequence replacement, gene correction, or gene knock-ins (Wang, La Russa, & Qi, 2016). The beauty of this system is that the sgRNAs can be designed to exhibit sequence complementarity to any sequence in the genome and induce a cut in the DNA so long as it is downstream of a PAM sequence.

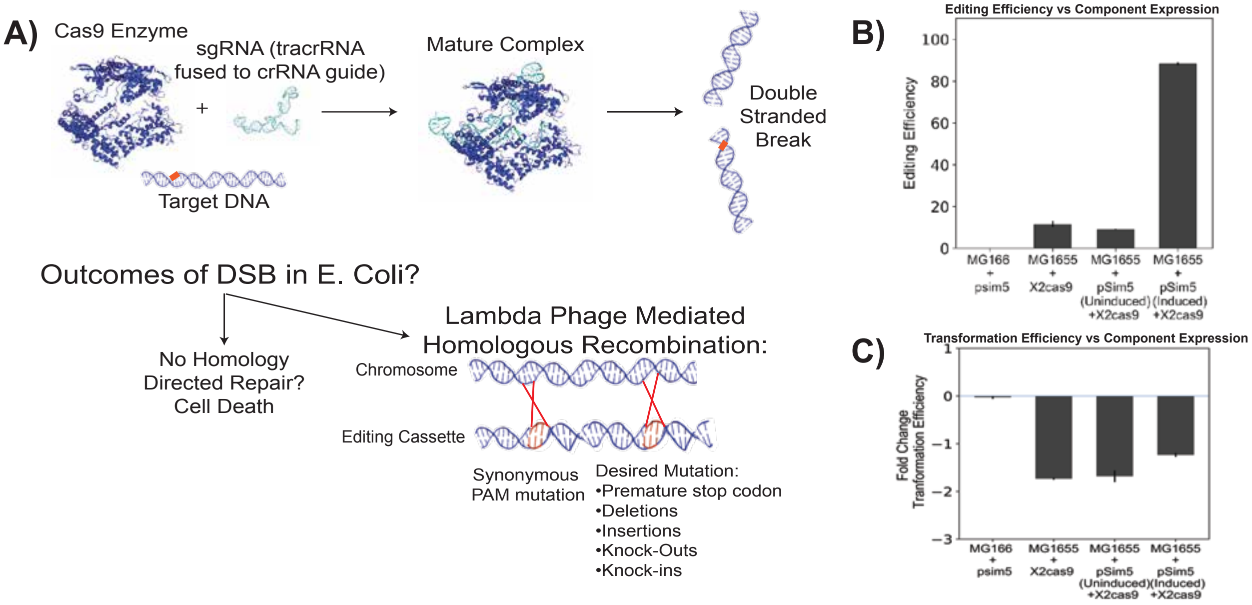


Figure 1. *The essentials of CRISPR-Cas9 mediated genome modification.*

(A) *Cas9 enzyme forms a mature complex when bound to a sgRNA. This complex then scavenges for a PAM sequence (NGG). Sequence homology between the guide RNA and the target genome will lead to a conformation change that induces a double-strand break in the genome. E. coli perform homologous recombination with the help of the lambda Red system to salvage a broken genome. If the DSB is left unrepaired, the cells will die due to a lack of a NHEJ pathway (Wilson, Topper, & Palmbos, 2003). (Crystal structures adapted from (Nishimasu, et al., 2014)).* ***(B)*** *High editing efficiency only resulted from expression of both Cas9 (to make the cut) and the lambda Red system for homologous recombination (transcribed from pSIM5 plasmid).* ***(C)*** *Cas9 and guide RNA expression led to high toxicity (negative fold change in transformation efficiency). Providing machinery for repair (lambda Red system) decreased the toxic effects by allowing the salvage of otherwise lethal double-strand breaks.*

E. coli lack a non-homologous end-joining pathway, thus the only means for repair of a broken genome is homology directed repair (Wilson, Topper, & Palmbos, 2003). This is significant because it provides a selection for successfully edited cells; any cells that fail to incorporate the homology arm with the desired mutation will die.

MODELS OF CRISPR CAS9 ACTIVITY

In 2016, Farasat and Salis constructed a model for site-specific Cas9 cleavage activity, taking into account many factors, including gRNA sequences, superhelical density of the genome, and Cas9 and crRNA expression levels (Farasat & Salis, 2016). While introducing a term for superhelical density into a model for predicting guide activity was novel at the time, they did so under the assumption that the genome exhibits uniform superhelical density, and did not take chromosome copy number into account.

In 2016, Lal et al. demonstrated that supercoiling is variable across the genome, thus indicating a fallacy in the Farasat and Salis assumption of uniform superhelical density (Lal, et al., 2016). This lack of understanding about positional dependence of not just cleavage activity, but also of actual genomic editing efficiency, prompted the first cut objective from this study: to investigate the positional dependence of genome modification.

While the Farasat and Salis study makes mention of the significance of the expression levels of Cas9 and guide RNA, they did so within the context of studying the kinetics of forming a mature complex, and thus, did not investigate neither how altering the number of active complexes impacts the ability to create a DSB, nor what impact that has on overall editing efficiency.

The Farasat and Salis study focused exclusively on the factors involved in creating a double-strand break. In so doing, their research did not demonstrate how any of their identified factors like superhelical density influence recombination, the second half of the process leading to a genomic modification. Their lack of characterization of the paste provided additional motivation for the final two studies of recombination in this project.

LAMBDA RED RECOMBINATION

Figure 1B illustrates how *E. coli* are inherently poor at facilitating homologous recombination. Thus, to achieve efficient recombination, CREATE relies upon exogenous sources of recombination machinery; the lambda Red system.

Homologous recombination has been previously characterized as a powerful tool to make genomic mutations, including insertions, deletions, and point mutations on chromosomes, plasmids, or BACs (Mosberg, 2010). Thus, homologous recombination is utilized to introduce genomic edits during a CREATE experiment. The proteins required to carry out homologous recombination arise from bacteriophage; the lambda Red system. These proteins are known as Gam, Exo, and Beta. The function of each of these proteins is as follows (Mosberg, 2010; Sharan, Thomason, Kuznetsov, & Court, 2009):

- **Gam:** Binds to the 3' end of ssDNA and prevents RecBCD and SbcCD (native *E. coli* nucleases) from destroying linear (double stranded) donor DNA. Blocking RecBCD is less significant with CREATE since circular, double stranded DNA is used as a recombination template. In systems that utilize linear double or single stranded DNA, blocking RecBCD activity becomes more significant for the availability of recombination template.
- **Exo:** Provides 5' to 3' exonuclease activity, which generates the single stranded overhangs necessary to anneal the donor DNA to homologous regions of the genome.

- **Beta:** Binds to the single stranded 3’ overhangs generated by Exo to promote annealing of the donor DNA to the homologous region of the genome.

The mechanism of making an edit using the lambda Red system is illustrated in Figure 2 (Pines, Freed, Winkler, & Gill, 2015).

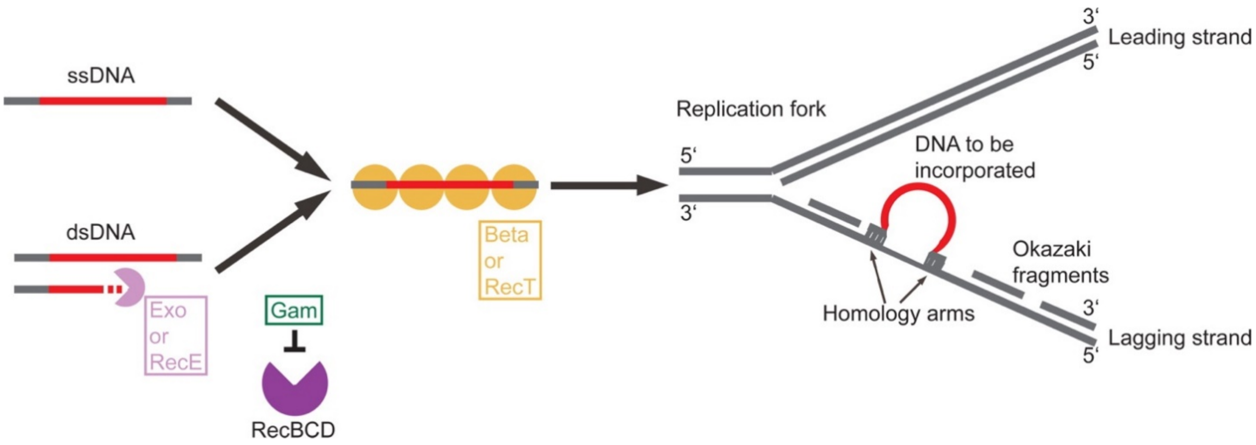


Figure 2 *Previously Characterized Mechanisms of Lambda Red Homologous Recombination.*

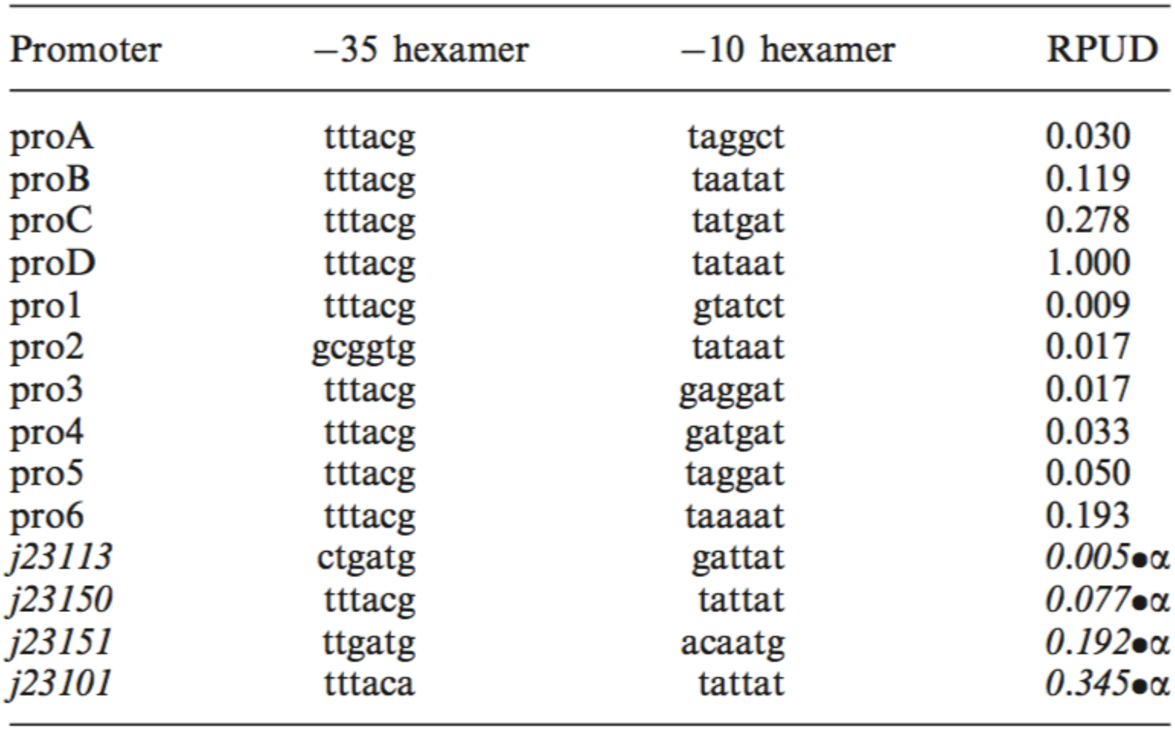
From (Pines, Freed, Winkler, & Gill, 2015). Illustrates how Exo degrades one strand of double stranded DNA to form a single stranded intermediate. Gam prevents host machinery from degrading the intermediate. Beta coats the single stranded intermediate to protect it from degradation and facilitate annealing to the homologous region of the genome. The donor DNA is incorporated in place of an Okazaki fragment during chromosomal replication.

Two parameters stand out when studying the mechanism of lambda Red recombination when aiming to make alterations. The first is the DNA donor; if this system is limited by the availability of the substrate for homologous recombination, could altering the availability of homology arm lead to increased editing efficiency? The second parameter that stands out in this system is the availability of recombination proteins: Gam, Exo, and Beta. If recombination is limited by the availability of recombination machinery, could increasing the expression of these genes could lead to increased editing efficiency?

MODIFICATION OF EXPRESSION THROUGH PROMOTER DESIGN:

In 2011, Davis, Rubin, and Sauer demonstrated that it was possible to design and built a set of insulated bacterial promoters of varying strength (Davis, Rubin, & Sauer, 2011).

Table 2. *Promoter strengths from table one of Davis, Rubin, and Sauer, Nucleic Acids Research 2010.*



Promoter	−35 hexamer	−10 hexamer	RPUD
proA	tttacg	taggct	0.030
proB	tttacg	taatat	0.119
proC	tttacg	tatgat	0.278
proD	tttacg	tataat	1.000
pro1	tttacg	gtatct	0.009
pro2	gcggtg	tataat	0.017
pro3	tttacg	gaggat	0.017
pro4	tttacg	gatgat	0.033
pro5	tttacg	taggat	0.050
pro6	tttacg	taaaat	0.193
j23113	ctgatg	gattat	*0.005*●α
j23150	tttacg	tattat	*0.077*●α
j23151	ttgatg	acaatg	*0.192*●α
j23101	tttaca	tattat	*0.345*●α

The sequences in shown in Table 2 were used to vary expression of Cas9 in the new system (see methods, section “building a set of varying strength promoters”).

CREATE TECHNOLOGY

The motivation for this study was improving strategies for high-throughput genome modification. Figure 3 illustrates the logic and workflow behind CREATE.

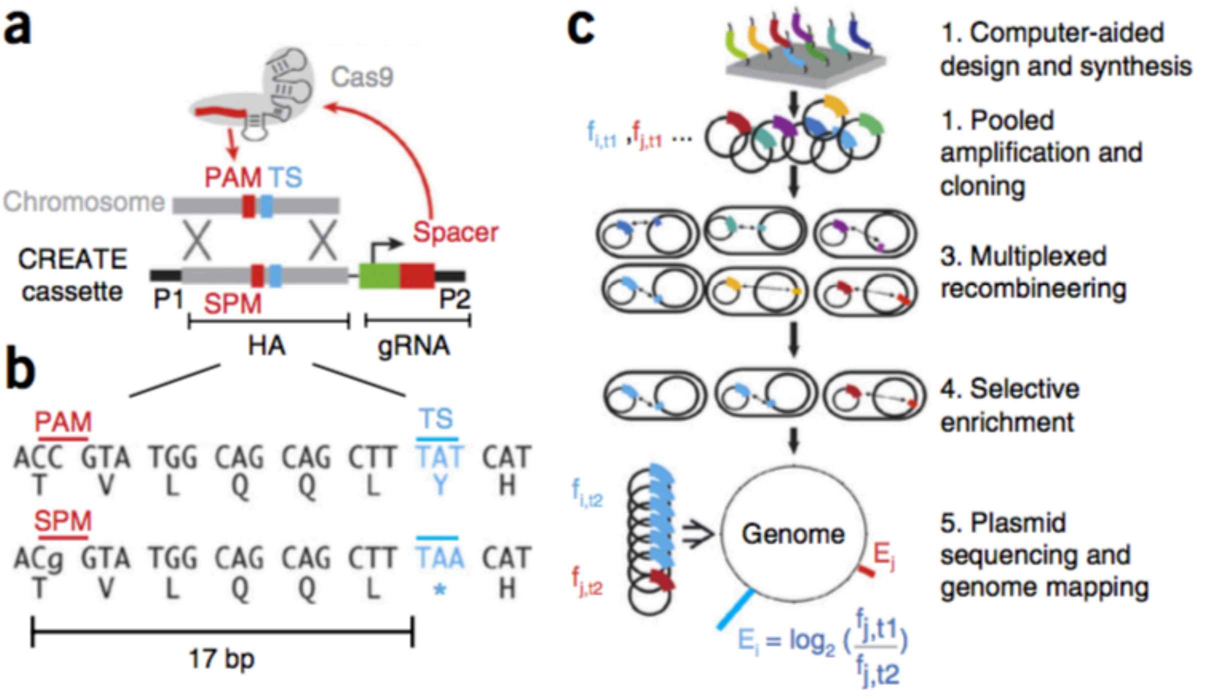


Figure 3. *CREATE technology.*

Figure 3 *from (Garst, et al., 2017)***. *(a)*** *Design of CREATE cassettes. A CREATE cassette contains a homology arm, with an embedded synonymous PAM mutation (SPM, red) and a target mutation (blue). The synonymous PAM mutation prevents mutated genes from being re-targeted by Cas9. Priming sequences in black allow for multiplexed amplification and cloning. The guide RNA is expressed from the same cassette, under constitutive expression.* ***(b)*** *Sample SPM (red) and target mutation (blue), introduction of a stop codon in this case. The SPM does not alter the coding sequence.* ***(c)*** *CREATE workflow. 1) The library of editing cassettes (designed using computational tools) is amplified off of a microarray chip. 2) the library is amplified in multiplex 3) multiplexed transformation of constructs into competent cells. 4) applying selective pressures. 5) sequencing the plasmids to determine enrichment of sequences, compared to the starting population, determined in stage 2. The change in frequency of cassettes between the initial and final time points allows for identification of beneficial variations.*

Three plasmids are introduced in a typical CREATE experiment. The first plasmid expresses the Cas9 endonuclease and contains a kanamycin antibiotic resistance marker. The second plasmid, pSIM5, contains proteins necessary for homologous recombination: Gam, Exo, and Beta. pSIM5 contains a chloramphenicol antibiotic resistance marker. The third plasmid, often called the CREATE plasmid, contains a CREATE cassette shown in Figure 3 and an antibiotic resistance marker.

CHALLENGES WITH CREATE

In an ideal CREATE experiment, all of the editing constructs exhibit uniform editing efficiency. The reality is that editing efficiency is not uniform, as illustrated in Figure 4. As mentioned in the introduction, wide variation in editing efficiency is problematic because it leads to noise during data analysis and decreases the heterogeneity of the starting population (not all of the desired mutations are actually introduced).

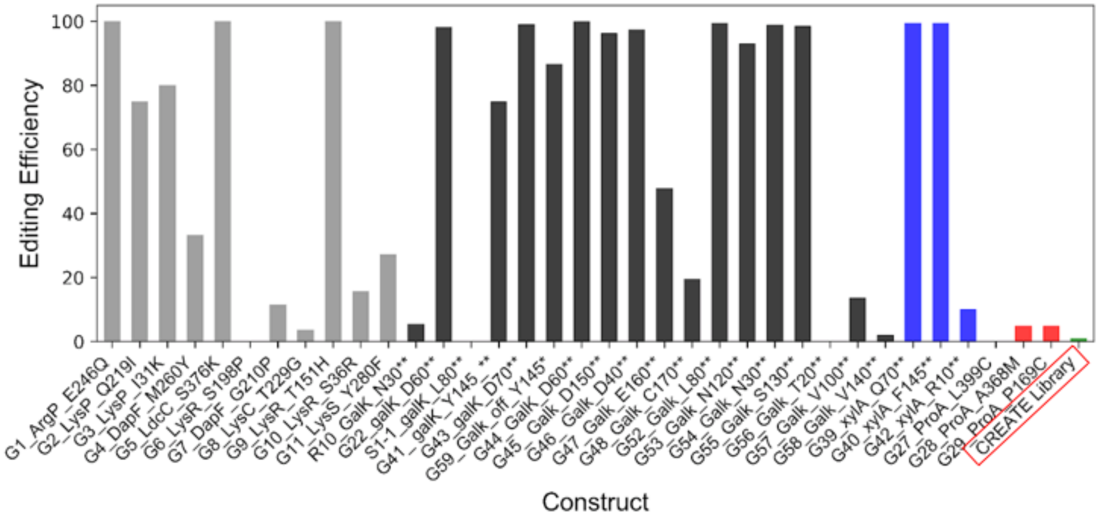


Figure 4. *Experimentally observed editing efficiencies across individual cassettes and at the full library scale.*

Different editing constructs exhibit large variability in editing efficiency, even within the same gene. Bars in black are galactokinase targeting constructs. Blue bars are xylose isomerase targeting constructs. Contsructs which target the ProA gene (gamma gluatmyl phosphate reductase) are in shown in red. Other genes are in gray. Editing efficiency at the full CREATE library scale is highlighted in the red box at the right of the graph.

CHAPTER THREE

METHODS

CULTURE MEDIA AND STRAINS

Strains were grown in Lysogeny Broth (LB) media (10g/L peptone, 5g/L NaCl, 5g/L yeast extract) at 37 °C (or 30 °C prior to recombineering)
LB agar plating media: (10g/L peptone, 5g/L yeast extract, 5g/L NaCl, 12g/L Agar)
MacConkey agar plating media: (17.0g/L peptone, 5.0g/L NaCl, 3.0g/L proteose peptone, 1.5g/L bile salts, 30.0mG/L neutral red, 1.0mg/L crystal violet, 13.5g/L agar, 10g/L galactose)
Escherichia coli strain MG1655 was used consistently in this work, to avoid strain to strain variance in editing efficiency.
E. cloni DH10B was used as a standard strain for cloning.
Strain SW105 was used to generate cassettes for the MG1655 *GalK* deletion strain, preceding safe site integration.

PLASMIDS

The Court lab provided the pSIM5 plasmid (Prior, Lynch, & Gill, 2010).

The Cas9 ORF was amplified off of the genome of *Streptococcus pyogenes* strain SF370 and cloned into the pBTBX-2 plasmid to generate the pX2-Cas9 plasmid (Datta, Costantino, & Court, 2006).

The pX2-Cas9 plasmid and the pSIM5 plasmid were further modified as described elsewhere in the methods section.

COMPETENT CELL CREATION PROTOCOL

To generate electrically competent cells, a starter culture of 5 mL LB broth (10 g/L peptone, 5 g/L NaCl, 5 g/L yeast extract) with the appropriate antibiotics was grown overnight in a shaking incubator at 30 °C if they contained the pSIM5 plasmid (lambda Red recombination proteins), and 37 °C if not. The next morning, 1 mL of the starter culture was inoculated into 100 mL of LB broth with the appropriate antibiotics and sugars added, if necessary. Common antibiotic markers were chloramphenicol for the pSIM5 plasmid, kanamycin for the X_2Cas9 plasmid, and spectinomycin for CREATE editing cassettes. 1 mL of 20% arabinose was added to the 100 mL culture if needed to induce expression of Cas9. The culture was grown in a shaking incubator at the appropriate temperature (30°C or 37°C), monitoring optical density until an absorbance of 0.4 with 600nm light was reached. If needed for expression of recombination proteins, cells were temperature induced in a shaking water bath at 42 °C for 15 minutes. The cells were placed on ice until chilled through, about 15 minutes. The cells were spun down at 8228 RCF for 3 minutes at 2 °C to form a pellet. Successive wash cycles followed: 3 washes with 25 mL of chilled sterile water, followed by one 15 mL wash with 10% chilled glycerol, 90% sterile water. After decanting off the excess fluid, the cells were re-suspended in 10% glycerol to a total volume of 1 mL, ready for transformation.

CLONING STRATEGIES TO GENERATE A LIBRARY OF UNIQUE EDITING CASSETTES

Circular Polymerase Extension Cloning (CPEC) (Quan & Tian, 2009) was used to generate a library of 50 unique guide RNAs. The process started with amplifying the insert (primers 1 and 2, Table 5) and the backbone (primers 3 and 4, Table 5) using the polymerase chain reaction (Figure 5) using the manufacturer recommended thermocycler protocol and annealing temperature for New England Biolabs Q5 High fidelity 2X Master Mix. The insert of 300 base pairs contained a priming region, homology arms, a synonymous PAM mutation, a target mutation, and a constitutive promoter expressing the 20 base pair gRNA. The backbone contained a spectinomycin antibiotic resistance marker, used as a selection for colonies that successfully took up the plasmid.

The amplified insert and backbone were gel purified following the QIAGEN protocol. The protocol for CPEC has been previously characterized in the literature, and was followed to combine insert and backbone into a single plasmid (Quan & Tian, 2009). An illustration of this process is provided in Figure 5, for clarity.

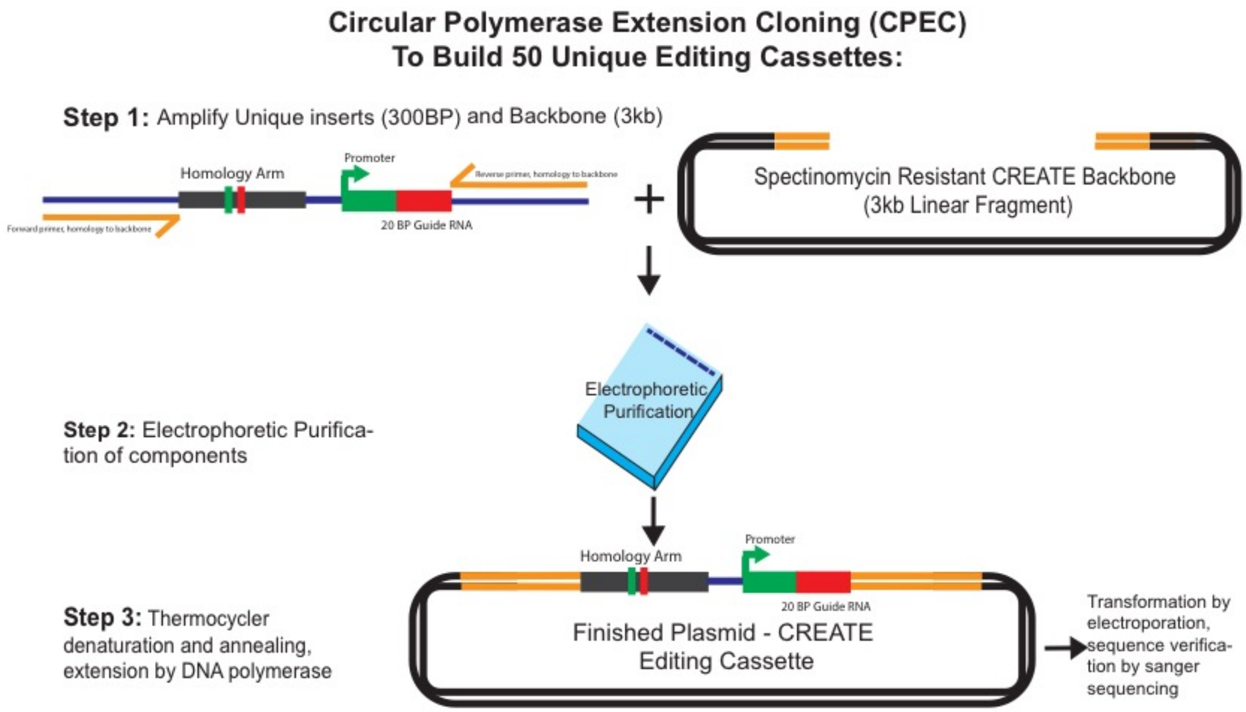


***Figure 5.** Circular polymerase extension cloning workflow*

Orange: homologous regions between backbone and insert. Each editing cassette contained a constitutive promoter expressing a unique gRNA sequence and its respective homology arm, including synonymous pam mutation and target mutation.

The recombinant DNA product was transformed into competent cells and sequence verified. 10 µL of recombinant plasmid was transformed into 50 µL of competent cells using a 1 mm gap cuvette in an electroporator set to 1800 V. The transformed cells were then recovered for one hour at 37 °C in LB media (10 g/L peptone, 5 g/L NaCl, 5 g/L yeast extract), and subsequently plated on LB agar plates (10 g/L peptone, 5 g/L yeast extract, 5 g/L NaCl, 12 g/L Agar) with spectinomycin antibiotic added. Plates were incubated overnight at 37 °C in a standing incubator. To ensure that the desired DNA product had been attained, and that no mutations arose during the cloning steps, the plasmids were amplified using colony PCR (primers 5 and 6, Table 5). This product was submitted to Quintara Biosciences for sequencing (primer 5, Table 5) After confirming the desired modifications using sequencing data, freezer stocks were made using 50% glycerol, and plasmids were extracted from overnight cultures using the QIAGEN spin miniprep protocol.

After the plasmids were built and sequence verified, they were tested for toxicity and editing efficiency. Figure 6 shows the workflow for testing the constructs for toxicity and editing efficiency. To test for toxicity, CREATE cassettes were transformed into *Escherichia coli* K-12, sub-strain MG1655 expressing Cas9. In the absence of homologous recombination proteins, any double-strand break in the *E. coli* genome should result in cell death, since *E. coli* cannot perform non-homologous end joining (NHEJ) to salvage broken DNA (Wilson, Topper, & Palmbos, 2003). It follows that plates with few colonies were the result of particularly effective guide RNAs - many DSBs were induced and few cells could survive. Colonies on each plate were counted to assign a quantitative toxicity score to each guide RNA. The equation was: $Transformation\ efficiency = \frac{Colony\ Forming\ Units\ (CFU)}{(ml\ cells\ plated)(mass\ DNA\ transformed)}$ thus the toxicity was normalized to both the amount of cells plated, and to the mass of DNA transformed during electroporation.

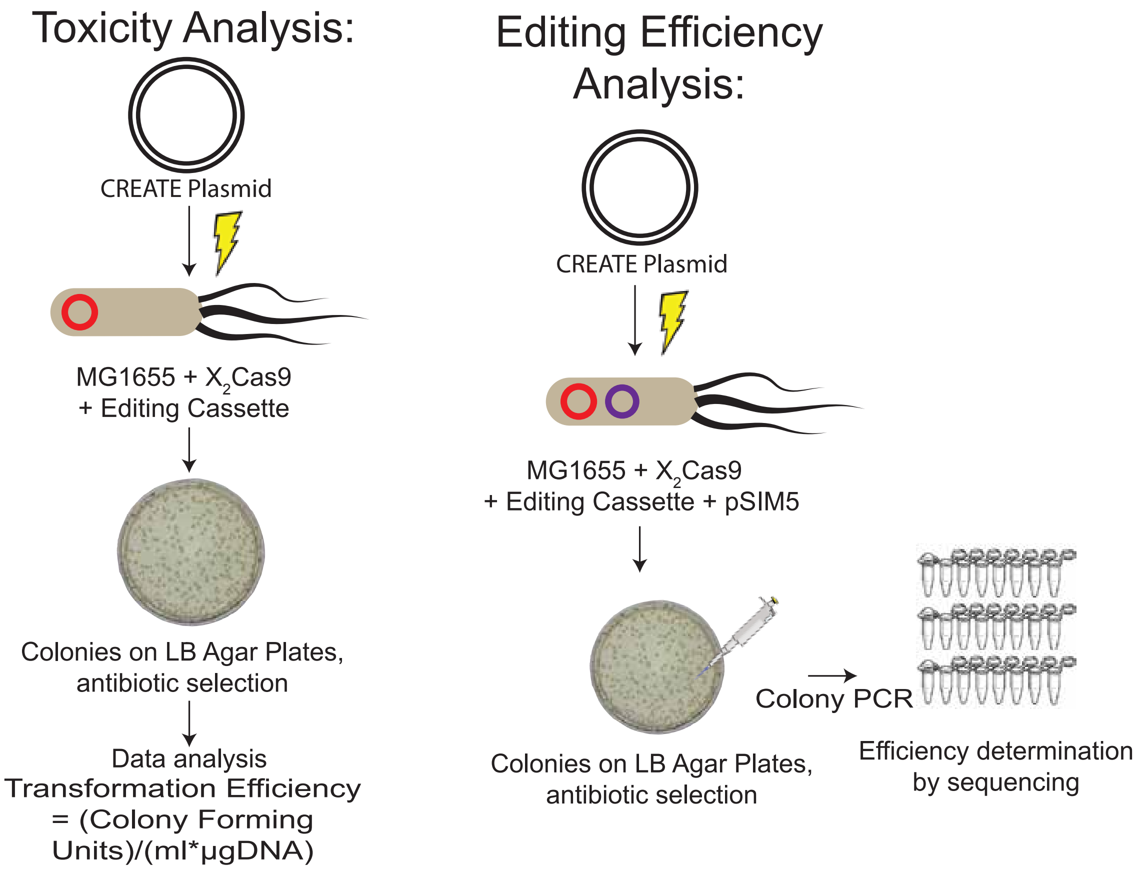


Figure 6. *Testing for toxicity and Editing Efficiency workflow.*

(Left) *the unique editing cassette and spectinomycin resistant backbone (CREATE plasmid) was transformed into MG1655 expressing Cas9, then plated on LB agar. Transformation efficiency was determined by counting up the number of colonies, then normalizing to the volume of cells plated and the mass of DNA transformed.* ***(Right)*** *Editing efficiency was determined by transforming a CREATE plasmid into cells expressing both Cas9 and recombination machinery (pSIM5). Editing efficiency was determined by colony PCR and sequencing for non-GalK targeting constructs, and through the MacConkey agar screen for GalK targeting constructs.*

The next variable tested was editing efficiency. Two different strategies were utilized to measure editing efficiency for the new constructs, based on which gene each gRNA targeted (Figure 6). The first strategy for determining editing efficiency was for those constructs which targeted any genes other than galactokinase (LysP, LdcC, PheA, DapF, etc.). Editing cassettes were transformed into competent cells expressing Cas9 and homologous recombination proteins. Transformed cells were plated on LB agar with spectinomycin and kanamycin antibiotics added. Colony PCR was performed on 10 colonies from each construct. The PCR product was sequenced, and the results were aligned to the target sequence to determine if the edit was made successfully or not.

SCREENING FOR EDITING EFFICIENCY (MACCONKEY)

The second strategy for determining editing efficiency was for those constructs which targeted the galactokinase (*galK*) gene. To measure editing efficiency of *galK* targeting constructs, the editing cassettes were transformed into competent cells containing both the X_2Cas9 plasmid and the pSIM5 plasmid. X_2Cas9 expressed the Cas9 protetin, and pSIM5 the recombination machinery. As such, when the guides created a double-strand break, the cells were saved from cell death via homologous recombination, introducing the two required mutations in the process. Transformed cells were recovered in LB media for three hours at 30°C, then plated on MacConkey agar (17.0 g/L peptone, 5.0 g/L NaCl, 3.0 g/L proteose peptone, 1.5 g/L bile salts, 30.0 mG/L neutral red, 1.0 mg/L crystal violet, 13.5 g/L agar, 10 g/L galactose) with kanamycin and spectinomycin antibiotics added. Plates were incubated overnight in a standing incubator at 37°C.

The mutation introduced into the *galK* gene was a premature stop codon, halting the cell's ability to metabolize galactose. On MacConkey agar, when cells are metabolizing galactose sugar (not successfully edited), they show up as red colonies on the plates. When the genomic edit has been made successfully, no red color is produced, and the cells grow as a translucent white colony. Thus, editing efficiency can be calculated by dividing the number of white colonies by the total number of colonies on each plate rather than performing colony PCR and Sanger sequencing.

NEXT GEN SEQUENCING

The number of samples being submitted for Sanger sequencing to determine editing efficiency became cost prohibitive after the positional dependence and Cas9 expression studies. Next generation Illumina sequencing technology was used to overcome this issue and to increase the statistical significance of results by increasing coverage and sample size. The workflow for sequencing the samples is illustrated in **Figure 7**.

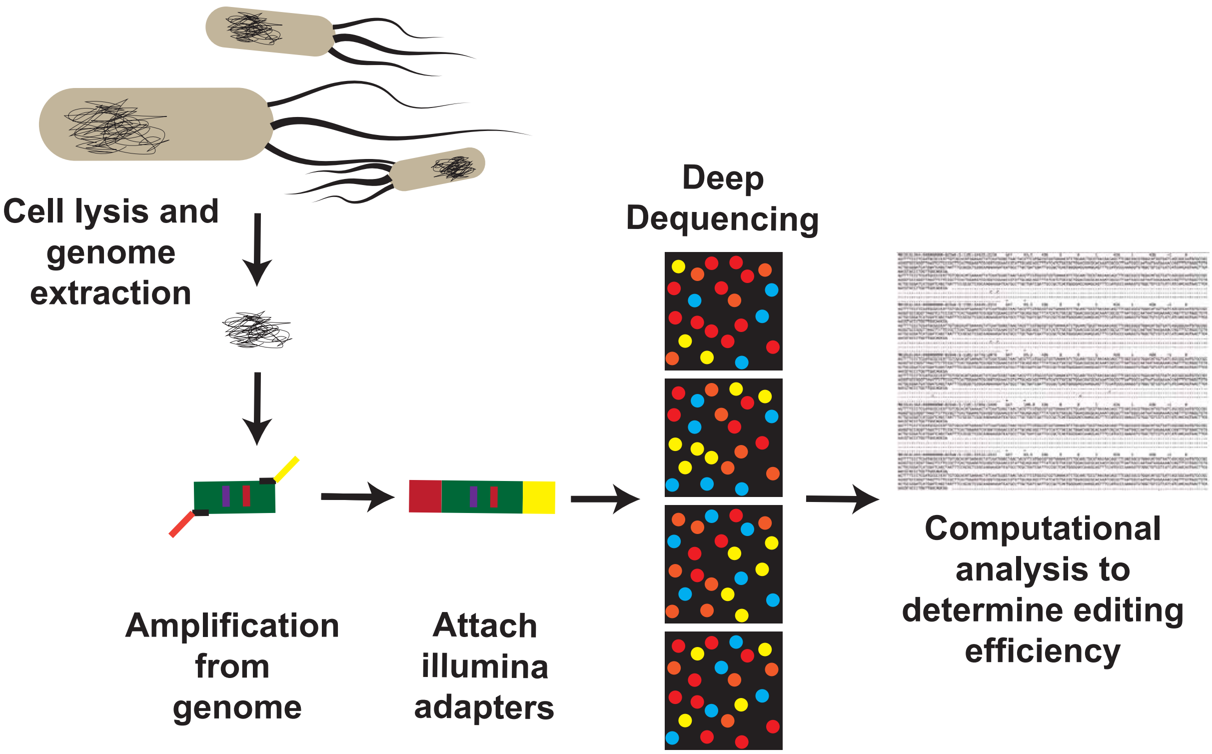


Figure 7. *Next generation Illumina sequencing workflow.*

Cells were lysed by boiling. The galK gene was isolated and amplified using PCR. The Illumina sequencing core performed PCR purification, attached adapters, and performed the sequencing reaction. Data analysis was performed using Python.

To prepare the samples for sequencing, colonies were scraped off of Agar plates using 1 mL of 10% PBS solution. 50 µL of cells were collected in a tube and pelleted at 13,000 rpm in a tabletop centrifuge for 10 minutes. The solution was decanted, and cells were resuspended in 50 µL of buffer TE (QIAGEN). Cells were boiled for 10 minutes at 100 °C, then diluted to a total volume of 1 mL. *galK* was amplified off the genome using primers 11 and 12 (Table 5) using the manufacturer recommended thermocycler protocol for 2X OneTaq HiFi polymerase.

Ultimately, 8 million reads were divided across 156 different experimental conditions. The quality control analysis of the sequencing run is in the supplements, Figure 26 . The data from this sequencing run was used in Figure 16 and Figure 20.

MAKING AND TESTING THE SAFE SITES

A functional deletion of the *galK* gene in MG1655 was generated using a cassette derived from the strain SW105, following a previously published method for incorporating metabolic pathways at specific genomic loci (Bassalo, et al., 2016). The *galK* gene was integrated at seven different genomic loci, or safe sites, in MG1655. Safe sites are genomic loci which have been previously demonstrated in the Gill lab to not influence the expression of native genes. Figure 8 illustrates the workflow behind the genomic incorporations.

The constructs Galk_0 through Galk_8 were tested for editing efficiency and transformation efficiency at the different regions. The results of the editing efficiency are shown in Figure 14. The results from the transformation efficiency experiments were not statistically significant and do not appear in the results section.

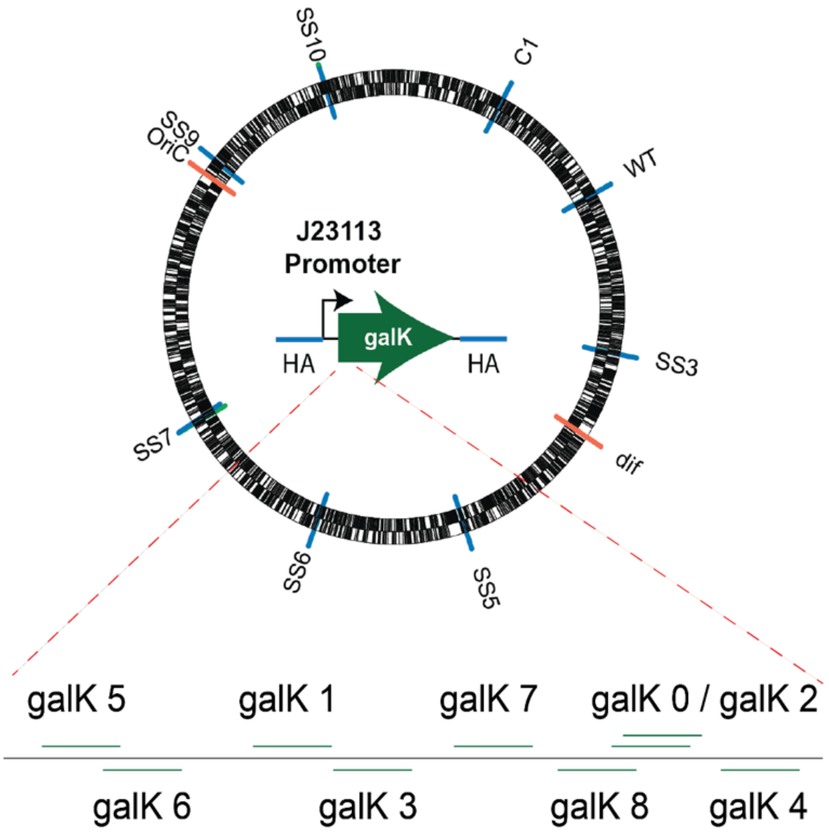


Figure 8. *Genomic positional dependence study design*

Outer ring: visual represenation of the E. coli genome, with relative positon of the safe sites shown. Starting with an MG1655 galK deletion strain, the safe sites were targeted with a guide specific to a desired site to cause a double-strand break. The galK gene was then inserted using sequence homology. The bottom of this figure illustrates the distribution of guides within the galK gene, notably, centered around the TSS of the gene.

BUILDING A SET OF VARYING STRENGTH PROMOTERS:

The promoter sequences identified by Davis, Rubin, and Sauer in 2011 were utilized in the development of strains with different Cas9 expression in this experiment (actual sequences cloned into Cas9 expression system listed in Table 3).

Table 3. *Promoter sequences in order of increasing expression used in this study*

Sequences adapted from Davis, Rubin, and Sauer. Nucleic Acids Research, 2011, Vol.39, No. 3

Promoter:	**-35 hexamer**	**-10 hexamer**	**Insulated Promoter Sequence:**

Pro1	tttacg	gtatct	ttctagagCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGAGTGGTTGCTGGATAAC**TTTACG**GGCATGCATAAGGCTCG**GTATCT**ATATTCAGGGAGACCACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTTtactagag
ProA	tttacg	taggct	ttctagagCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGAGTGGTTGCTGGATAAC**TTTACG**GGCATGCATAAGGCTCG**TAGGCT**ATATTCAGGGAGACCACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTTtactagag
ProC	tttacg	tatgat	ttctagagCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGAGTGGTTGCTGGATAAC**TTTACG**GGCATGCATAAGGCTCG**TATGAT**ATATTCAGGGAGACCACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTTtactagag
ProD	tttacg	tataat	ttctagagCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGAGTGGTTGCTGGATAAC**TTTACG**GGCATGCATAAGGCTCG**TATAAT**ATATTCAGGGAGACCACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTTtactagag

In total, three new constructs were generated with varying expression levels of Cas9 (Figure 9). The insulated promoter sequences were ordered as G-blocks from Quintara biosciences, amplified, then cloned into the backbone of a kanamycin-resistant Cas9 plasmid using CPEC and the manufacturer recommended thermocycler protocol using New England BioLabs Q5 High Fidelity 2X Master Mix. The weakest expression was the original arabinose inducible system. Pro1, ProC, and ProD were chosen as promoters since they provide the widest range of expression levels of the identified constitutive promoter sequences (0.009, 0.278, and 1.00 relative promoter units, respectively). 10 μL of the recombinant DNA product was transformed into 50 μL of E. cloni DH10B competent cells using a 1 mm gap cuvette in an electroporator set to 1800 V.

The transformed cells were recovered for one hour at 37 °C in LB media and plated on LB agar with kanamycin antibiotic added. Plates were incubated overnight at 37 °C. To ensure that the desired DNA product had been attained and that no mutations arose during the cloning steps, the plasmids were amplified using colony PCR. This product was sent to Qunitara Biosciences for sequencing. After confirming the desired modifications using sequencing data, freezer stocks were made in 50% glycerol, and plasmids were extracted from overnight cultures using the QIAGEN spin mini-prep protocol.

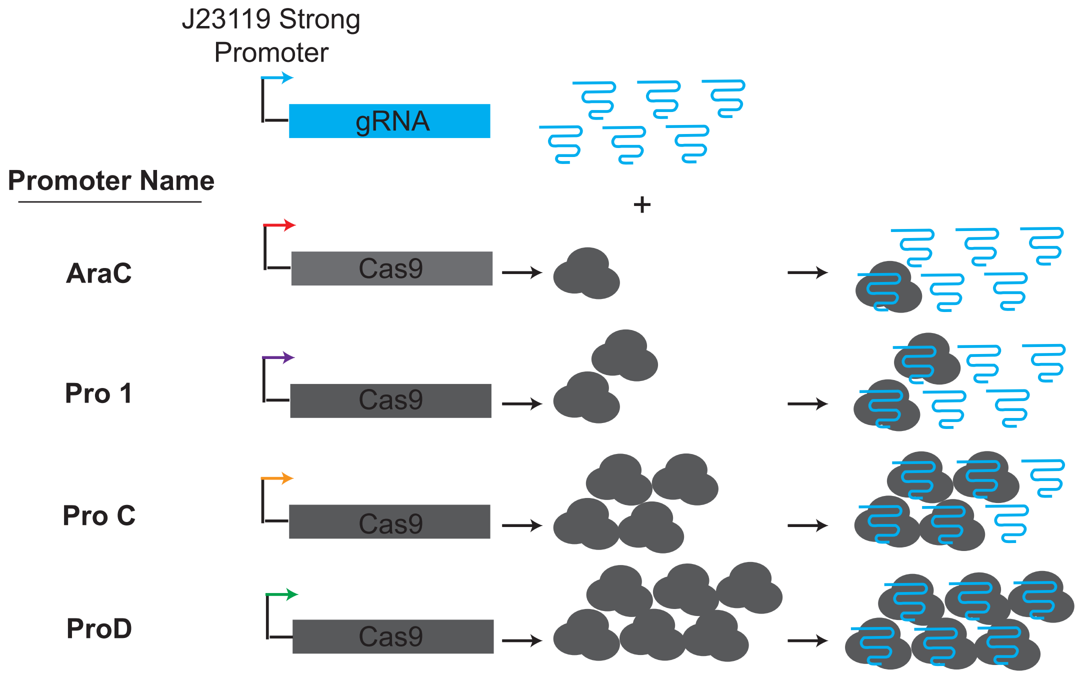


Figure 9. *Illustration for experimental design of Cas9 expression variants.*

Expression level of gRNA was kept constant across experiments. Moving down the figure shows the expected outcome (more active Cas9-gRNA complexes) of increasing cas9 expression moving from the weak arabinose inducible promoter to the strong constitutive ProD promoter.

To test the editing efficiency as function of expression of Cas9, each of the four Cas9 constructs had the pSIM5 plasmid introduced and verified using chloramphenicol antibiotic selection. Editing constructs Galk_0 through GalK_8 were transformed into each of the Cas9 expression variants expressing recombination proteins. The resulting editing efficiency was determined using the MacConkey Agar screen.

BUILDING AND TESTING THE LINEAR HOMOLOGY ARMS

The first parameter of the lambda Red recombination system (Figure 2) that was chosen for modification was altering the availability of recombination substrate. The experimental design for altering the availability of recombination substrate is illustrated in Figure 10. To generate the linear homology arms, primers 7 and 8, (Table 5) were used to amplify the homology arms off of the traditional CREATE plasmids. In another PCR reaction, primers 9 and 10 (Table 5) were used to remove the homology arm from the CREATE plasmid, such that its only function was to express a guide RNA. The product of this PCR reaction was circularized using DpnI digestion, gel extraction, PNK, and ligation reactions.

Three different quantities of homology arm were tested. The lowest quantity was an equimolar ratio between the homology arm and the gRNA only plasmid (1 mol HA : 1 mol gRNA only plasmid). The intermediate quantity of homology arm was 10 ng, and the highest quantity of HA was 100 ng. A constant two microliters of gRNA only plasmid was transformed with its respective homology arm, regardless of the quantity of homology arm transformed. Each of the three concentrations of homology arm was tested with editing constructs Galk_1, Galk_3, Galk_5, Galk_6, and Galk_7 at three genomic loci – SS3 (far from OriC), the native *galK* position (intermediate distance), and SS9 (close to OriC).

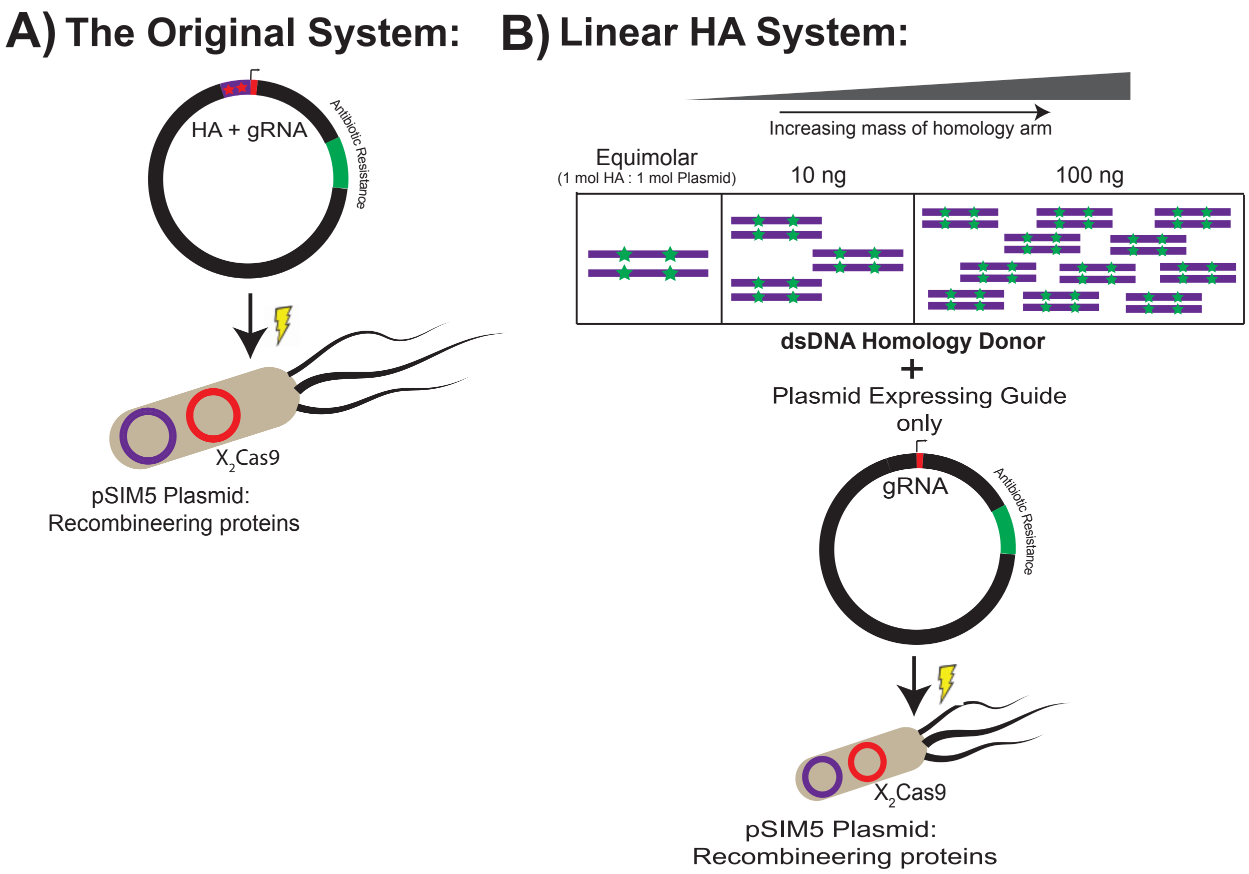


Figure 10. *Experimental design for altering the availability of homology arm.*

(A) *The original CREATE system; a CREATE cassette that contains both the homology arm and the guide RNA is transfromed into MG1655 containing both the* X_2*Cas9 plasmid and the pSIM5 plasmid.* ***(B)*** *The new system with variable quantity of homology arm. Three quantities of homology arm (equimolar (one mole HA to one mol gRNA only plasmid), 10ng, and 100ng) were co-transformed with a plasmid expressing gRNA only into MG1655 containing both the* X_2*Cas9 plasmid and the pSIM5 plasmid. Construct were tested with the GalK gene at three different genomic loci. Editing efficiency was analyzed using next generation sequencing.*

Altering the availability of the homology arm while it still existed in the plasmid would have required altering the copy number of the plasmid itself. Using linear homology arms made changing the quantity of homology arm simple: one could transform a variable amount of the homology arm with a constant amount of the plasmid expressing only the guide RNA. There was also temporal significance with this design; the plasmid proliferates throughout cycles of cellular replication and division, and thus is available to facilitate HDR for indefinite amounts of time.

The linear homology arms, on the other hand, get degraded by host exonucleases. This implies that to successfully make an edit, the cut and repair need to happen before the homology arm gets degraded. Testing a single construct at a time allows for mixing the guideRNA and the homology arm together as in the linear homology arm experiment. However, the homology arm and the guide RNA must, by design, be expressed from the same plasmid in a full scale CREATE experiment because many editing constructs are mixed together before electroporation. There needs to be a way to make sure that the right guide gets transformed with its respective homology arm into the same cell. Additionally, the plasmid stays with the edited population, which allows the edits to be tracked.

MAKING AND TESTING THE PRO1 PSIM5 AND VARIANTS

The second parameter of the lambda Red recombination system (Figure 2) that was chosen for modification was the availability of recombination protein. The objective was to build a construct with constitutive, yet conservative expression of Gam, Exo, and Beta to avoid issues associated with Gam toxicity (Sergueev, Yu, Austin, & Court, 2001).

Transcription of *Gam*, *Exo*, and *Beta* genes under the PL promoter is thermally regulated by the λ CI857 repressor (Lewis, 2011). Between 30°C and 32°C, the CI repressor gene is active and generates CI protein, which represses transcription of the recombination genes (Lewis, 2011). The repressor protein is inactive at 42 °C, allowing transcription of *gam,exo,* and *beta* to occur (Lewis, 2011). Mechanistically, repression of the PL promoter is caused by the formation of a loop in the DNA at CI binding sites, thus inhibiting transcription (Lewis, 2011). This mechanism is illustrated in Figure 11.

Three operator sites where the CI protein binds are responsible for repressing transcription of *gam, exo, and beta* (Lewis, 2011). These operator sites have a hierarchy of affinity for the CI Protein, where $O_R3 < O_R2 < O_R1$ (Lewis, 2011). This hierarchy was used to avoid introducing toxicity into the system by over-expressing the Gam protein. Pro1 (the weakest of the constitutive promoters tested) was cloned in front of the *gam, exo, beta* operon using primers 13 and 14 (Table 5), and sequence verified using primer 15 (Table 5). Operator sites O_R1 and O_R2, with the highest affinity for the CI repressor protein were removed in the same PCR reaction, leaving only O_R3, the operator site with the weakest affinity for the CI repressor protein, in place. The expected outcome was intermediate expression of recombination proteins, rather than very strong constitutive expression.

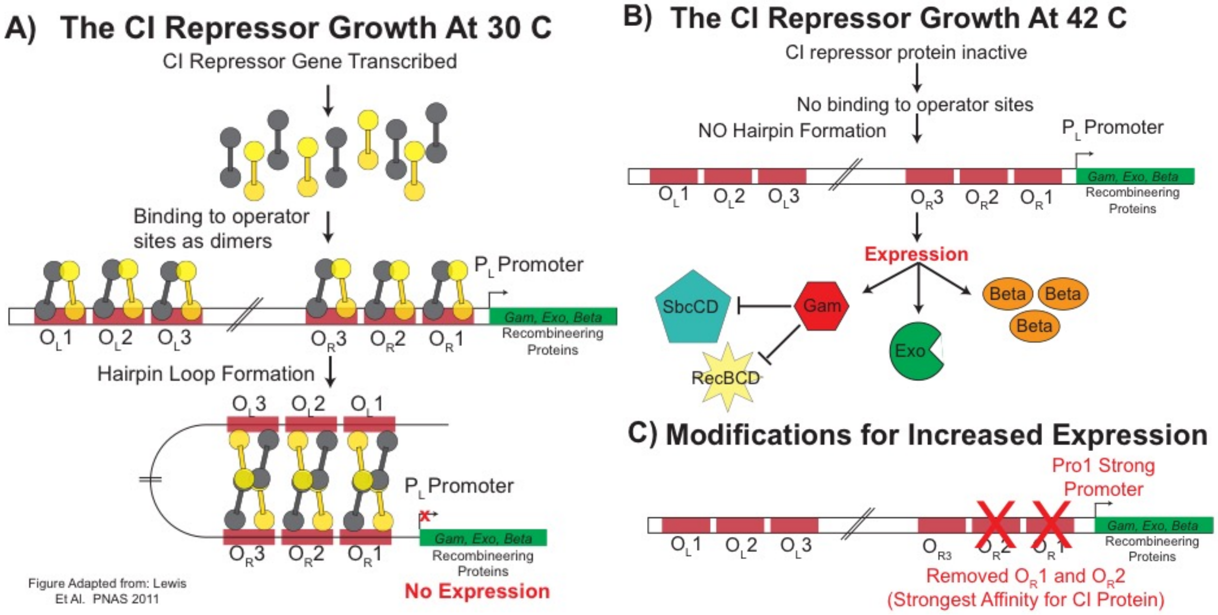


Figure 11. *Design of a tunable construct for expression of recombination proteins.*

(A) *Figure adapted from (Lewis, 2011). The phage based system for repression of recombination proteins. At 30℃, the CI protein is expressed, which binds to operator sites O_R and O_L 1, 2, and 3, repressing the operon.* ***(B)*** *At 42℃, the CI repressor protein becomes inactive and no longer binds to the operator sites. Recombineering proteins are transcribed* ***(C)*** *Note that Gam is toxic (Sergueev, Yu, Austin, & Court, 2001). Instead of removing all operator sites and cloning in ProD (the strongest constitutive promoter used in this study) an intermediate step was chosen. Operator sites O_R1 and O_R2, (highest affinity for the CI repressor protein) were removed, leaving only O_R3, the operator site with the weakest affinity for the CI repressor protein. This was accomplished in the same PCR reaction as cloning in the Pro1 constitutive promoter.*

During a typical CREATE experiment, cell cultures are induced at 42°C for 15 minutes before introduction of the editing cassette by electroporation, meaning the recombination proteins are only expressed for roughly one cell cycle. The implication of this short induction time is that once the editing cassette is transformed into the cell, there is only a small time frame during which Cas9 can complex with the gRNA, cause a double-strand break, and carry out homologous recombination to deliver the desired edit before the recombination proteins are no longer present in the cell to facilitate introduction of genomic modifications.

The workflow for expression of recombineering proteins in a typical CREATE experiment is illustrated in Figure 12.

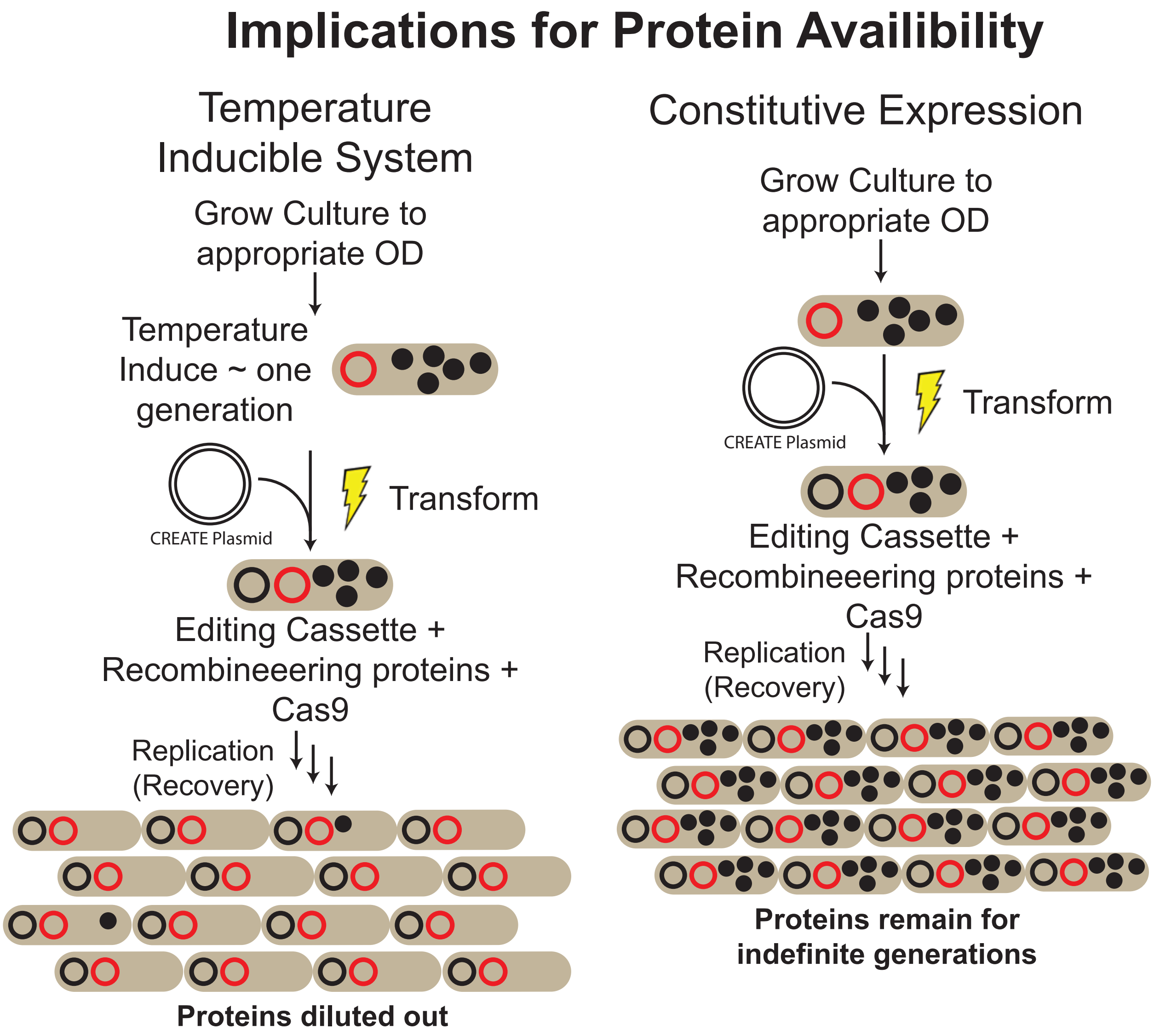


Figure 12. *Implications for Protein Availability in Dividing Cells with Constitutive Expression of Lambda Red Proteins.*

(Left) *The traditional temperature inducible expression system used in CREATE experiments. The recombination proteins are temperature induced for 15 minutes before transformation. During cellular replication and division, the proteins are successively diluted out. The cut must happen before the recombination proteins get diluted out, to allow for homologous recombination to occur.* ***(Right)*** *The new constitutive expression system. No temperature induction step is needed before transformation, and the proteins are expressed throughout cellular replication and division. If a guide takes a long time to find the target site and cause a DSB, then a broken genome can be salvaged further into the process of recovery – something the traditional system does not allow.*

In the traditional system, the recombination proteins get diluted out of the population while cells are replicating. Using a constitutive promoter to express the recombination protein combats this issue. When recombination proteins are expressed constitutively, each cell generation regenerates new recombination proteins, as opposed to being diluted out.

To generate the pSIM5 knockout strains used in Figure 18 and Figure 19, PCR was used to amplify the backbone, excluding the gene of interest, then DpnI digestion, agarose gel purification, PNK, and ligation reactions were performed as per the manufacturer instructions to circularize the linear amplification product.

The primers used to generate the “no Exo” construct were: 21 and 22 (Table 5)

The primers used to generate the “no Gam” construct were: 27 and 28 (Table 5)

The primers used to generate the “Partial Gam” construct were: 17 and 18 (Table 5)

The circular recombinant plasmid product was transformed into standard cloning E. cloni DH10B competent cells and sequence verified. Sequencing the plasmid involved the following primers: 16, 19, 20, 23, and 24 (Table 5). Using this collection of primers covered all of the *Gam, Exo,* and *Beta* genes to ensure none were perturbed during the cloning steps. Freezer stocks were made of verified constructs using 50% glycerol, and plasmids were extracted from overnight cultures using the QIAGEN spin miniprep protocol.

Recombinant plasmids were transformed into MG1655 cells containing the X_2Cas9 plasmid, then screened for successful transformation using chloramphenicol and kanamycin antibiotics on LB Agar plates. Successfully transformed cells were cultured and made electrically competent following the above protocol. The Galk_0 editing construct was then transformed into the cells. Editing efficiency was determined following the MacConkey agar screen protocol (above) and transformation efficiency was calculated by counting the number of

CFUs on each plate, then normalizing to the volume of cells plated and the mass of DNA transformed.

The Δ*RecA* strain was prepared using CREATE. Primers 25 and 26, from Table 5 were used to generate the gRNA, targeting the *RecA* gene, and a *RecA* deletion cassette, with sequence given in Table 6 was used to knock out the native gene.

CHAPTER 4

RESULTS AND DISCUSSION

TOXICITY VS EDITING EFFICIENCY – A SIGMOIDAL TREND

In total, 37 unique editing constructs were tested for both editing efficiency and transformation efficiency. After plotting the editing efficiency data against the toxicity data, a strong correlation was shown to exist between the two variables (Figure 13), taking the shape of a sigmoid curve. This would suggest, mechanistically, that the double-strand break created by Cas9 is very important to achieve high editing efficiency. Guide RNAs that effectively created DSBs (highly toxic guides) had high editing efficiency, and those that were poor at creating the DSB had a low editing efficiency.

The guides were subsequently segregated into three groups based on their behavior. Group one guides exhibited high toxicity and high editing efficiency – the effective cutters. Group two guides exhibited intermediate toxicity and intermediate editing efficiency. Group three guides were characterized by both low toxicity and low editing efficiency.

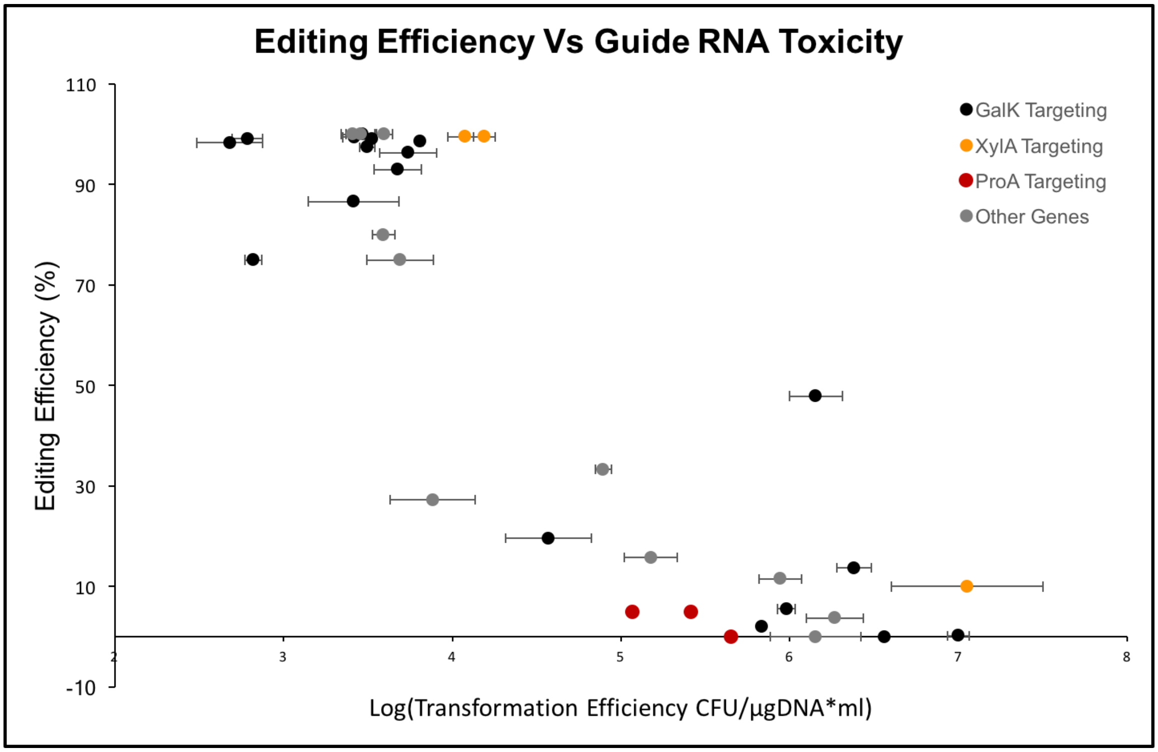


Figure 13. *Results comparing editing efficiency to guide RNA toxicity.*

Guides with high toxicity exhibited high editing efficiency, while guides with low toxicity exhibited low editing efficiency. This highlights the importance of creating a double-strand break in attaining high editing efficiency.

The most significant conclusion of this study is that high toxicity is related to high editing efficiency, and low toxicity is related to low editing efficiency - the ability to create a DSB in the genome is important to attaining high editing efficiency. This study did not, however, explain the variation in editing efficiency and the variation in toxicity. What features make one guide highly efficient at editing as opposed to an inefficient guide? The remaining objectives for this project were aimed at getting to the root of this question.

THE CUT SECTION ONE: POSITIONAL DEPENDENCE

Nine editing constructs (spanning the three groups) out of the 37 used in the editing efficiency vs transformation efficiency study (Figure 13) were selected as a representative sample and tested at each of seven genomic locations. A construct targeting the xylose isomerase (*xylA*) gene was used as a control to make sure that the observed variation in editing efficiency was a result of the position, not a result of variation in editing between the engineered strains. The results from studying the positional dependence of editing efficiency are shown in Figure 14.

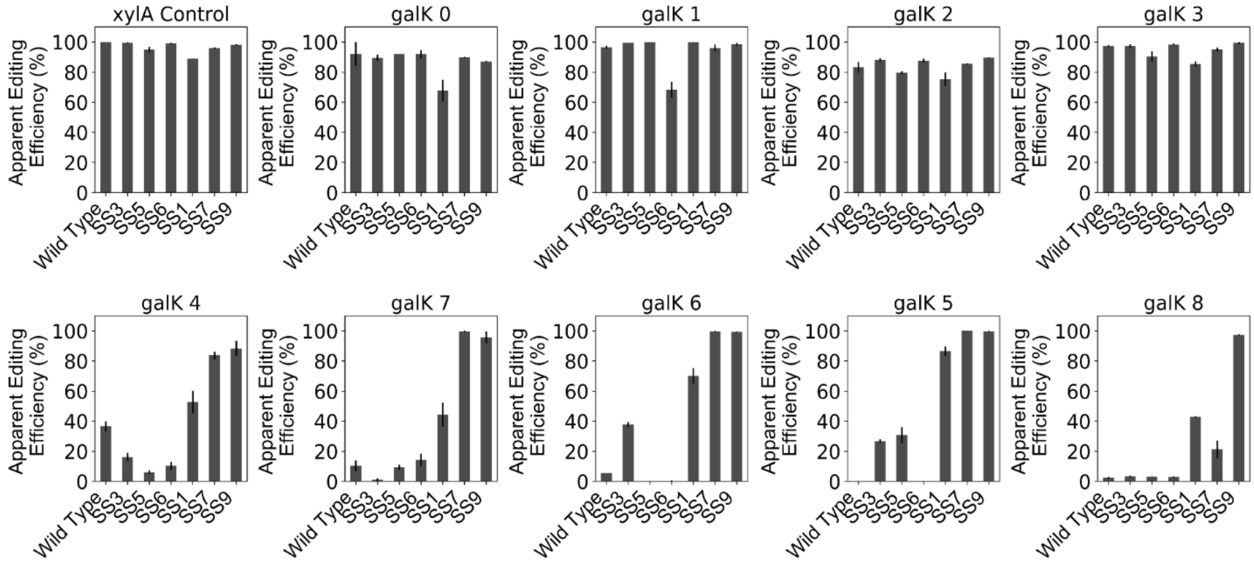


Figure 14. *Positional dependence of editing efficiency.*

Moving from left to right in each graph, the genomic target is closer to the chromosomal origin of replication. Group one guides (Galk_0, Galk_1, Galk_2, and Galk_3) exhibited uniform behavior; editing efficiency was not dependent upon genomic position. Group two guides (Galk_4, and Galk_7) showed a systematic increase in editing efficiency as the site being edited was moved closer to the chromosomal origin of replication. Group three guides (Galk_6, Galk_5, and Galk_8) demonstrated inconsistent behavior.

The conclusion from the positional dependence study is that a change in the activity of Cas9 and/or the activity of recombination is dependent upon the genomic position of the target site, leading to higher editing efficiency closer to the origin of replication.

THE CUT SECTION TWO: ALTERED EXPRESSION OF CAS9

The same editing constructs used in the positional dependence study were used to determine the impact of altering the expression levels of Cas9 on editing efficiency. Once again, a construct targeting the xylose isomerase (*xylA*) gene was used as a control to ensure that the observed behavior was not exhibited in a gene-specific fashion. The results of investigating the effects of expression levels of Cas9 on editing efficiency are shown in Figure 15.

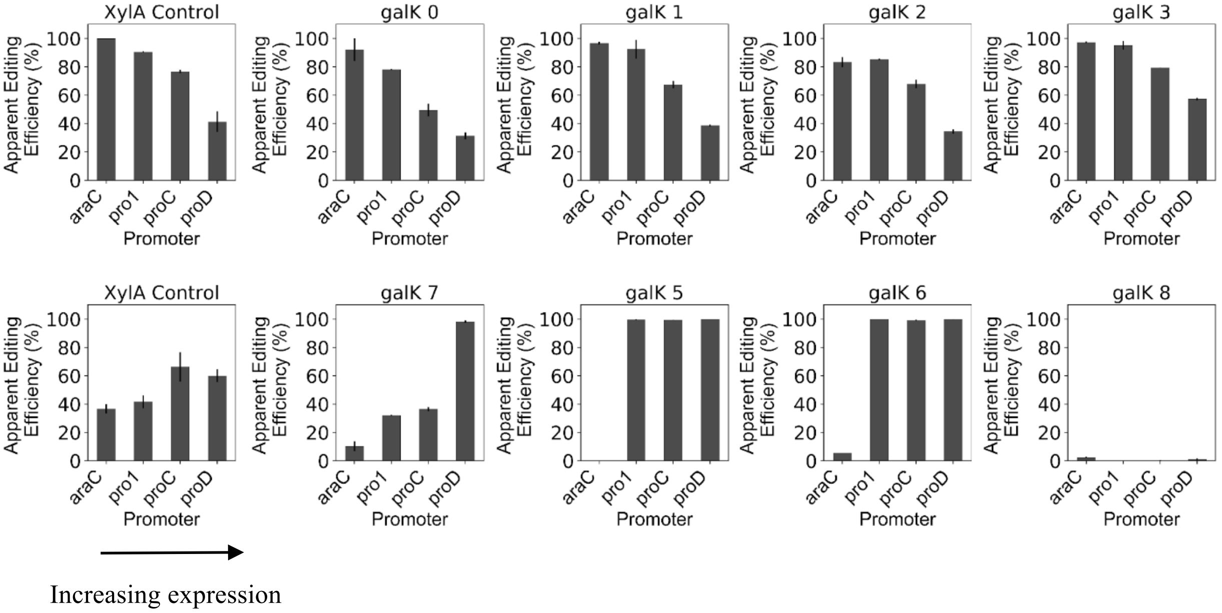


Figure 15. *Altering expression levels of Cas9 impacts editing efficiency.*

Moving from left to right in each graph, the expression level of Cas9 is increasing, with araC the lowest and ProD the highest expression. Group one guides (Galk_0, Galk_1, Galk_2, and Galk_3) showed a decrease in editing efficiency as promoter strength increases. Group two guides (Galk_4, and Galk_7) showed a systematic increase in editing efficiency as the site being edited was moved closer to the chromosomal origin of replication. Group three guides (Galk_6, Galk_5, and Galk_8) exhibited inconsistent behavior.

Altering the expression level of Cas9 changes editing efficiency in a guide-specific manner. Group one guides showed a systematic decrease in editing efficiency. Group two guides showed a systematic increase in editing efficiency. And group three guides showed inconsistent behavior. Because of this, increasing expression levels of Cas9 might not be the best decision when attempting to normalize guide behavior. It could, however, be useful in tuning the editing efficiency of group two guides to exhibit more desirable behavior.

THE PASTE SECTION ONE: ALTERED ABUNDANCE OF RECOMBINATION SUBSTRATE

Two parameters were identified with respect to modifying recombination. The first was altering the abundance of homology arm within the cell. The hypothesis here was that changing the quantity of homology arm would lead to variation in editing efficiency. The results from investigating how the abundance of homology arm impacts editing efficiency are shown in Figure 16.

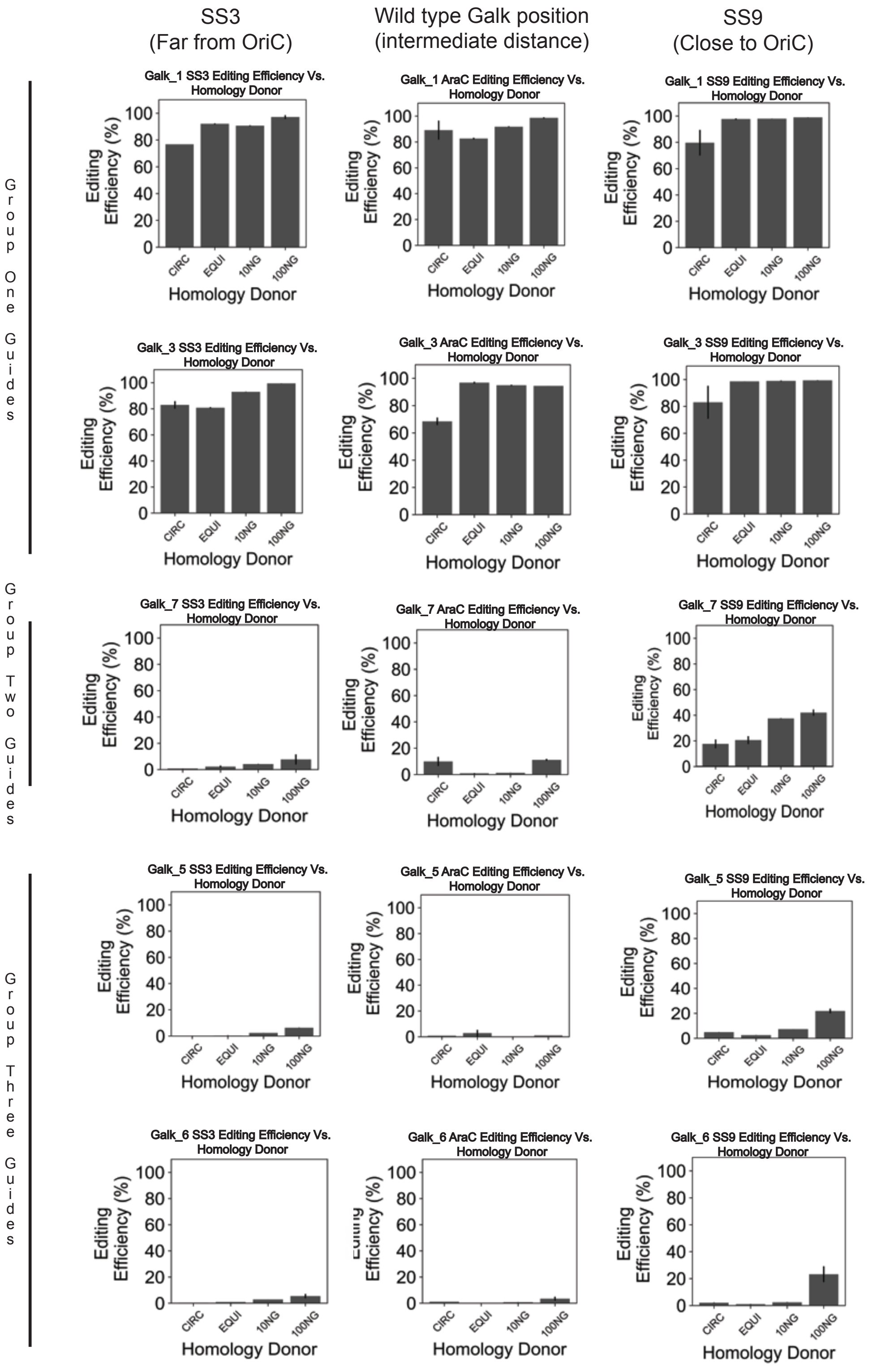


Figure 16. *Editing efficiency by changing availability of homology arm and genomic position.*

The first bar in each graph represents the editing efficiency using the traditional plasmid based editing system. Moving from left to right within each graph, the availability of linear homology *arm increases. Moving right across the page follows a trend of moving closer to OriC. Group one guides (Galk_1 and Galk_3) show similar editing efficiency, or perhaps a slight increase in editing efficiency compared to the plasmid based system. The group two and group three guides (Galk_5, Galk_6, and Galk_7) show an increase in editing efficiency with increasing availability of homology arm, with the most significant increase closest to the origin of replication.*

It is important to note that what is changing with each bar on the bar graph is the availability of homology arm. These results suggest that if editing efficiency is limited by the cut (group two and group three guides), increasing availability of homology arm can provide a modest improvement. One necessity for high editing efficiency is effective cutting of the DNA, however, these results suggest that with the poor cutters, editing efficiency could be limited by either the cut or the paste. These results also suggest that recombination is dependent upon position since the largest changes in editing efficiency occur closest to the origin of replication.

A plateau in editing efficiency was observed with the group 1 guides, but not with the group two or group three guides. This could indicate that editing efficiency could be improved even further by transforming more homology arm. A suggestion for the future would be to transform even more homology arm until the maximum improvement in editing efficiency is achieved (see “future studies”).

Transformation efficiency was tested alongside editing efficiency for each of the experimental conditions in Figure 16. The results of investigating how changing the availability of homology arm impacts transformation efficiency are given in Figure 17. The results of the transformation efficiency study parallelled the results from the editing efficiency study, shown in Figure 16. The decreased toxicity close to the origin of replication observed with group one guides is suggestive of more effective repair at that site. This suggests that there is a variation in recombination due to position. Another parallel between the editing efficiency and transformation efficiency graphs is that the largest magnitude of change in EE and TE with varying availability of homology arm were observed at SS9, closest to the origin of replication.

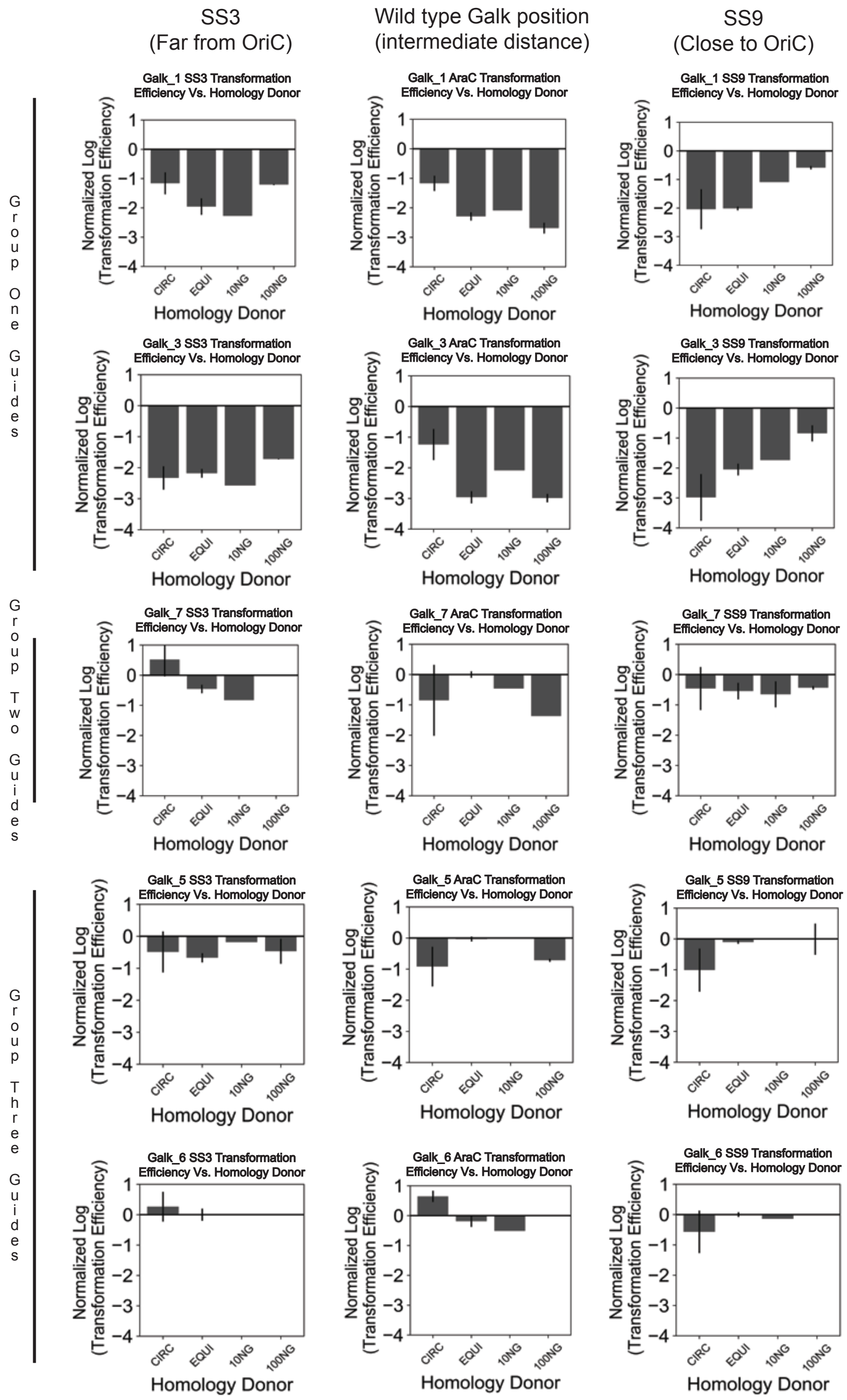


Figure 17. *Transformation efficiency compared to availability of homology arm.*

The first bar in each graph represents the transformation efficiency using the traditional plasmid based editing system. Moving from left to right within each graph, the availability of linear homology arm increases. Moving right across the page follows a trend of moving closer to OriC.

Group one guides (Galk_1 and Galk_3) show a decrease in transformation efficiency close to the origin of replication. At the intermediate genomic location, a slight increase in toxicity is observed. At the OriC proximal site there was a marked decrease in toxicity. The group two guide and group three guides (Galk_7, Galk_5, and Galk_6) yielded no statistically significant results.

THE PASTE SECTION TWO: ALTERED EXPRESSION OF RECOMBINATION PROTEINS

The second modification to recombination was with expression of recombination proteins.

This involved two phases of study. The first study was with modifying the expression of recombination proteins was an analysis of different loss-of-function variants of the lambda Red recombination system. The second study was the impact of constitutive expression of lambda Red recombination proteins on both editing efficiency and transformation efficiency. The following results were achieved from the first study when studying what Gam, Exo, and RecA proteins contribute to editing efficiency:

The *Exo* deletion construct showed a sharp decrease in editing efficiency by about 60% (Figure 18). I hypothesize that Exo is involved in resecting the DSB in the genome to make ssDNA overhangs. Thus, even though the recombination template is circular in CREATE, processing the genomic DSB with the Red system is essential for high editing efficiency

Preliminary results with the “no Gam” construct were surprising; it was expected to exhibit similar editing to the original construct since the Gam protein should be inessential when using a homology arm on a plasmid to repair a broken genome. The hypothesis was that deleting the entire *gam* gene had disrupted the entire *Gam, Exo, Beta* operon, and therefore no recombination proteins were being synthesized. To test this hypothesis, the *gam* knockout strain was redesigned, eliminating only the first 139 bases out of the 468 base pair ORF, rather than all 468. The new construct produced the expected outcome – comparable editing efficiency to the original system (Figure 18). This indicates that the Gam protein, while it is toxic to the cells, does not contribute to high editing efficiency. This suggests that working with *gam* deletion constructs in the future could help with increasing survivability of edited cells.

The *RecA* deletion construct had similar editing efficiency to wild type M1655 (Figure 18**).** This highlights the necessity of phage-derived recombination machinery in genome modification experiments; *E. coli* contribute little towards editing efficiency through native recombination systems.

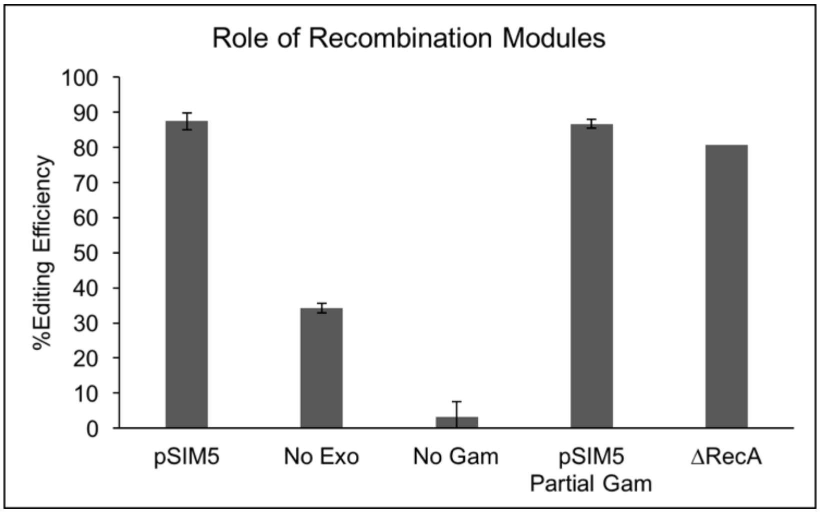


Figure 18. *Comparing recombination protein knockout strains to editing efficiency.*

Expression of pSIM5 has been observed to be essential for high editing efficiency. When the exo gene is inactivated from the lambda Red system, there is a decrease in editing efficiency. The first attempt at deactivating only the Gam *protein yielded unexpected results where the editing efficiency was nearly zero. The deletion was redesigned, and the new results suggested that* Gam *is in fact inessential to high editing efficiency. The RecA deletion strain also showed that* RecA *is inessential for high editing efficiency when the lambda Red system is expressed.*

The following results were achieved when studying what Gam, Exo, and RecA proteins contribute to toxicity:

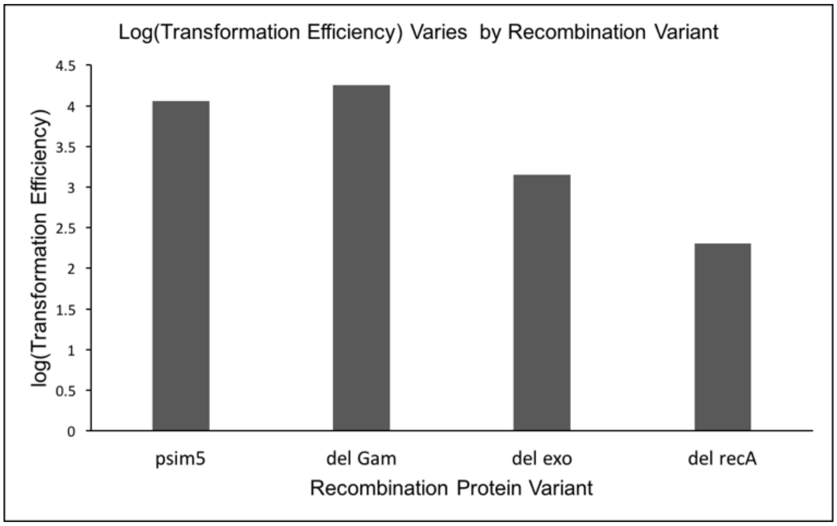


Figure 19. *Comparing recombination variants to toxicity.*

As expected, when compared to the traditional lambda Red system (pSIM5), the gam deletion strain showed decreased toxicity. The exo deletion strain exhibited increased toxicity, as did the RecA deletion strain. As expected, the partial gam construct showed slightly higher transformation efficiency, or rather, decreased toxicity when compared to the traditional pSIM5 system which expresses the Gam *protein.*

The partial *gam* construct showed less toxicity than traditional pSIM5 (Figure 19).This outcome was expected from the findings of (Sergueev, Yu, Austin, & Court, 2001), but serves as an interesting proof of concept nonetheless.

The *exo* deletion construct exhibited higher toxicity than the traditional system, suggesting that there was impaired repair of lethal double-strand breaks without Exo (Figure 19)

The *RecA* deletion strain showed higher toxicity than wild type MG1655 (Figure 19). This is likely due to the fact that RecA works in concert with LexA to initiate the entire SOS response to DNA damage (Voet & Voet, 2011). The SOS response is not limited to double-strand breaks; it responds to numerous sources of damage, including UV radiation, alkylation, and cross-linking (Voet & Voet, 2011). Inactivating RecA has larger impacts on cell viability than just repairing Cas9 induced DSBs.

The second parameter studied was the impact of constitutive expression of lambda Red recombination proteins on both editing efficiency and transformation efficiency. The results follow:

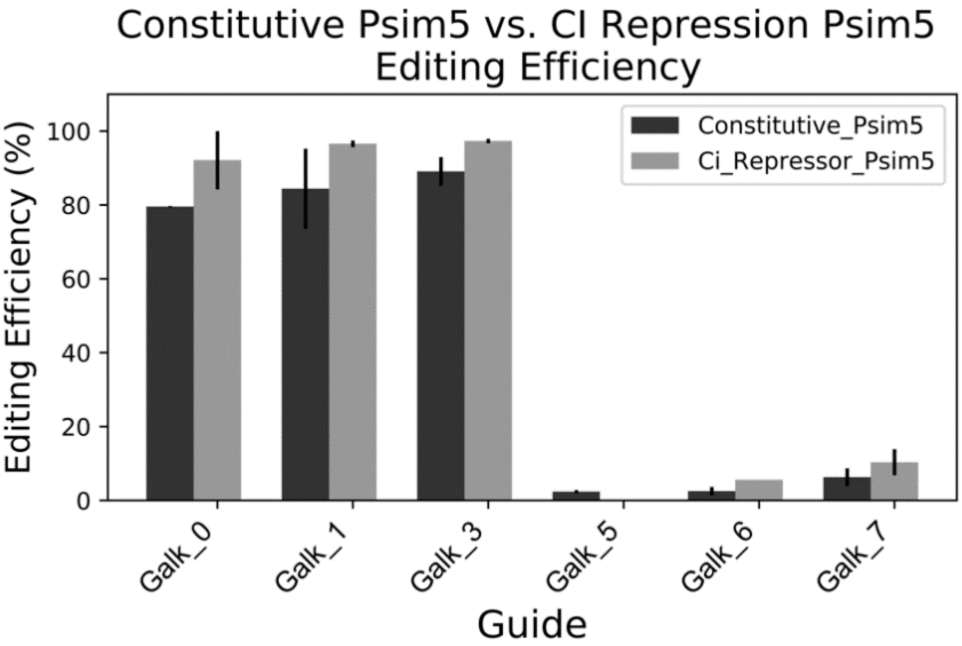


Figure 20. *Comparing editing efficiency of constitutive expression of lambda Red recombination proteins to the temperature inducible system.*

Ci_repressor_pSIM5 is the original temperature inducible system shown in grey, and constitutive_pSIM5 is the newer pro1 system shown in black. Constitutive expression of recombination proteins yielded no increase in editing efficiency.

A slight decrease in editing efficiency was observed with all guides (Figure 20). One potential cause for the decrease in editing efficiency is that the O_R3 operator site is still repressing the gene. Alternatively, the Pro1 constitutive promotor may be weaker in expression than the native P_L phage derived promoter. These two possibilities would contribute to decreased expression of recombination proteins, ultimately making the system recombination limited rather than limited by the cut. To answer this question, in the future one could delete the Gam protein since it is both toxic and inessential (Figure 18 & Figure 19) and increase the promoter strength for expression of Exo and Beta to determine if the new constitutive expression system is in fact recombination limited, and if the editing efficiency can be increased by further increasing expression of recombination proteins.

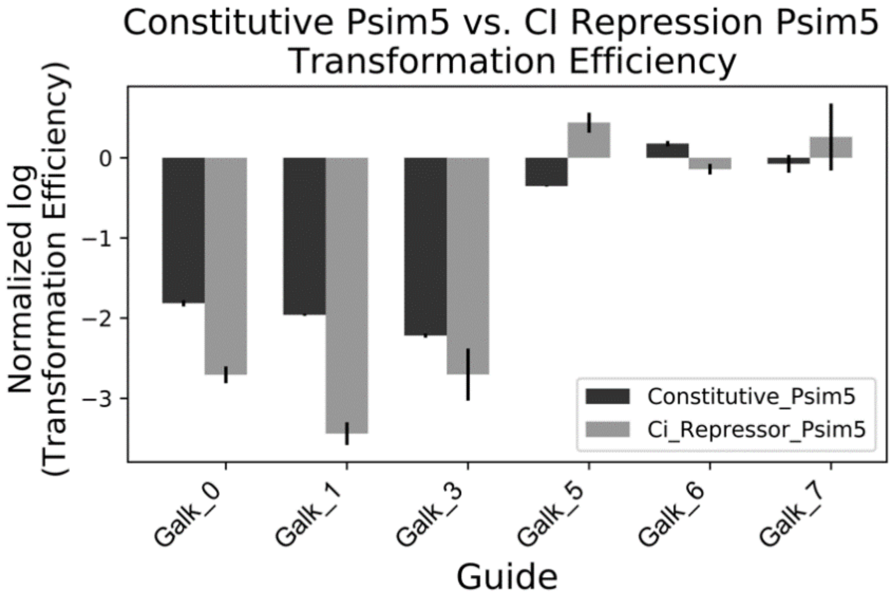


Figure 21. *Comparing transformation efficiency of constitutive expression of lambda Red recombination proteins to the temperature inducible system.*

Ci_repressor_pSIM5 is the original temperature inducible system shown in grey, and constitutive_pSIM5 is the newer pro1 system shown in black. The group one guides (Galk_0, Galk_1, and Galk_3) showed decreased toxicity with constitutive expression of recombination proteins. Constitutive expression showed no statistically significant impact on group two and group three guides (Galk_5, Galk_6, and Galk_7).

A decrease in toxicity was observed with constitutive expression of recombination proteins with all group one guides, statistically insignificant changes in transformation efficiency were observed with group two and group three guides (Figure 21).

Why was there a decrease in toxicity with the constitutive expression of recombination proteins? There could be two options. The first option is that lethal double-strand breaks are being repaired more efficiently than before. This seems unlikely, however, given that this construct exhibits decreased editing efficiency (Figure 20). The other, and more likely, alternative is that the expression of the constitutive system is still weaker than the traditional temperature inducible system. In that case, less Gam protein would be synthesized, a protein which is known to be toxic to the cell (Figure 19 and (Sergueev, Yu, Austin, & Court, 2001)).

The lack of change in toxicity with group two and group three guides could be explained by the fact that in this case, editing efficiency is limited by the cut. Altering the availability of proteins needed to salvage a broken genome is likely less important when few cuts are being made in the first place.

Ultimately, the study successfully generated a tunable construct that mimics the original system for recombination protein expression. This system can now be further tuned.

This specific design was not useful for increasing editing efficiency, but there is room for improvement. Constitutive expression of recombination proteins could prove useful for other applications where long lasting expression of recombination proteins is necessary

CHAPTER FIVE

CONCLUSIONS

The conclusions section begins with a highlight of the important findings from this study. The findings are then used to make suggestions for future experimental designs.

1) How does guide RNA toxicity impact editing efficiency?

It was observed that high toxicity is related to high editing efficiency, and low toxicity is related to low editing efficiency - the ability to create a DSB in the genome is important to attaining high editing efficiency. The cause of the variation in toxicity and editing efficiency was explored further in aims 2-5.

2) The cut: how does the genomic position of the gene being targeted impact editing efficiency?

The positional dependence study showed that the activity of Cas9 is dependent upon the position on the genome of the gene being targeted, and that there is an increase in editing efficiency closer to the origin of replication.

3) The cut: how does the expression level of cas9 impact editing efficiency, and can increasing expression of cas9 increase editing efficiency?

Altering the expression level of Cas9 changes editing efficiency in a guide-specific manner. Group one guides showed a systematic decrease in editing efficiency. The current hypothesis is that the observed decrease in editing efficiency is due to off target activity. If the complexes are binding to an off target site in the genome and creating a double-strand break, then there will not be an appropriate homology arm available for repair, and thus edited cells will be selected against by the toxicity caused by off-target effects. Group two guides showed a systematic increase in editing efficiency. Group three guides showed inconsistent behavior.

4) Recombination, the paste: how does altering the quantity of available homology arm impact editing efficiency?

The results of aim four suggest that if editing efficiency is limited by the cut (group two and group three guides), increasing availability of homology arm can provide a modest improvement in editing efficiency. One necessity for high editing efficiency is effective cutting of the DNA, however, these results suggest that with the poor cutters, editing efficiency could be limited by either the cut or the paste. These results also suggest that recombination is dependent upon position since the largest changes in editing efficiency occur closest to the origin of replication.

5) What do Gam, Exo, and RecA proteins contribute to toxicity and editing efficiency, and can altering the expression levels of recombination proteins lead to increased editing efficiency?

The results from aim five are:

- The Exo protein makes a significant contribution to achieving high editing efficiency.
- A partial gam construct produced the expected outcome – comparable editing efficiency to the original system.
- The Gam protein, while it is toxic to the cells, does not contribute to high editing efficiency.
- The *RecA* deletion strain highlighted the necessity of phage-derived recombination machinery in genome modification experiments.
- The partial *Gam* construct showed less toxicity than traditional pSIM5, as expected.
- The *Exo* deletion construct exhibited higher toxicity than the traditional system.

- The *RecA* deletion strain showed higher toxicity than wild type MG1655, highlighting its significance in the response to a wide range of DNA damage.

Ultimately, the study successfully generated a tunable construct that mimics the original system for recombination protein expression. This system can now be further tuned. This specific design was not useful for increasing editing efficiency, but there is room for improvement.

The best strategy for overcoming current problems associated with CREATE could involve altering some or all of the kinetic parameters that have been experimented with thus far, in an attempt to normalize the guides. The ideal guide will have both high editing efficiency and transformation efficiency. How can the results from aim one through five be used in the future to optimize guide RNA design for high editing efficiency?

1) If possible, choose guides that are known to exhibit high toxicity. Choosing guides from a databank of guides previously experimentally determined to be efficient would be help with normalizing behavior.

2) Little can be done about the position of a gene within the host organism's genome. It is feasible, however, to group libraries of mutants based on genomic position. One might divide a full size CREATE library into multiple sub-libraries based on genomic position, and then perform steps 2 through 5, Figure 3, such that guide behavior within each sub-pool is more uniform than if they were all cloned in multiplex at once.

3) Increasing expression levels of Cas9 might not be the best decision when attempting to normalize guide behavior at the full scale CREATE experiment level, since the response to expression varies by guide group. Varying Cas9 expression could, however, be useful in tuning the editing efficiency of editing constructs with known toxicity behavior. For example, if a guide

is known to exhibit group two-like behavior, Cas9 could be expressed heavily to attain high editing efficiency.

4) The homology arm and the guide RNA must, by design, be expressed from the same plasmid in a full scale CREATE experiment because many editing constructs are mixed together before electroporation. There needs to be a way to make sure that the right guide gets transformed with its respective homology arm, which is accomplished by introducing them simultaneously on one plasmid. Additionally, the plasmid stays with the edited population, which allows the edits to be tracked. Thus, increasing the availability of homology arm in a CREATE experiment could be achieved by increasing the copy number of the plasmid, which could increase editing efficiency in a position-dependent manner.

5) Use a constitutive promoter stronger than Pro1 and remove both O_R3 and the functionality of the *gam* gene to yield high expression of Exo and Beta. This was not proven to be effective in this study, however the results would indicate that is likely the best strategy to go about modifying recombination proteins.

FUTURE STUDIES

Kinetics study

Determine the K_d of associating to the genome and the K_c, actually causing the DSB. Try to replicate the observed behavior of group 2 guides by introducing mismatches to the guides – this would evidence for guide activity depending on K_d.

Explain the decrease in editing efficiency of group one guides with increased Cas9 expression

The current hypothesis is that the observed decrease in editing efficiency is due to off target activity. If the complexes are binding to an off target site in the genome and creating a double-strand break, then there will not be an appropriate homology arm available for repair, and thus edited cells will be selected against by the toxicity caused by off-target effects. To test this hypothesis, delete the major off target site and see if high editing efficiency is maintained in the presence of high expression levels of Cas9.

Improving computational tools

Employing computational tools for predicting editing efficiency would be extremely useful for improving overall CREATE editing efficiency. I decided to put the most commonly used gRNA efficiency prediction algorithm to the test against the toxicity data used in Figure 13. The code was developed by Doench et al. in 2016, and is found on Benchling's web platform for designing guides (Doench, et al., 2016). This code uses position of the cut within the translated gene, and the sequence of the guide itself to assign a score – two somewhat simple analyses.

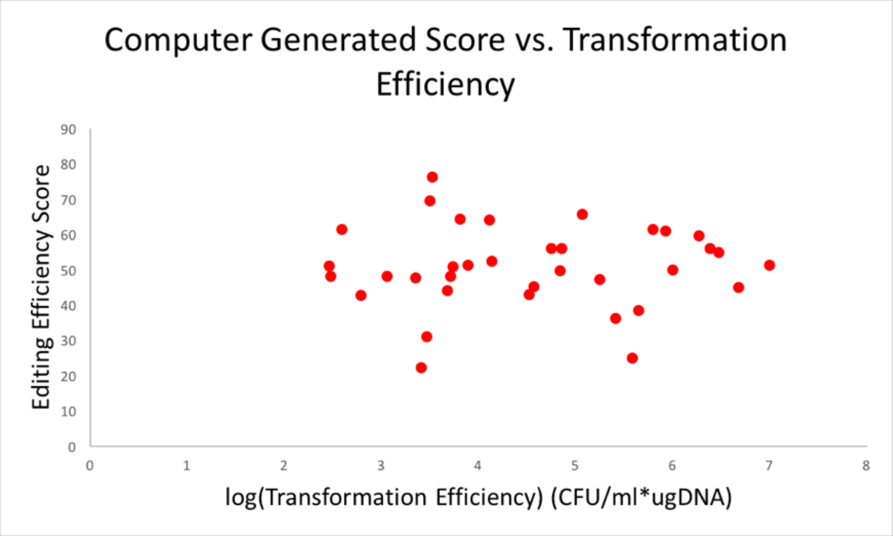


Figure 22. *Transformation Efficiency vs. Computer Predicted Editing Efficiency.*

The predicted editing efficiency of a guide RNA showed no correlation to the experimentally determined toxicity values. Computational algorithms for predicting editing efficiency need to be improved. Scores generated using (Doench, et al., 2016).

The lack of correlation suggests understanding of the system and all the factors that contribute to guide RNA toxicity are not fully understood. The position of the cut within a gene, and the sequence of the guide are not sufficient metrics for predicting guide efficiency. These models were trained using eukaryotic DNA data, which could explain the poor predictions, however this still suggests that there are no sufficient models for editing the bacterial genome to date.

More research needs to be done to probe what the primary determinants of gRNA toxicity are. Predictive software needs to be advanced by taking even more factors (including but not limited to those identified in this study) that impact editing efficiency into account to make better predictions. As identified in this study, guide RNA toxicity, genomic position of the target site, Cas9 expression level in the system of interest, availability of homology arm, and expression of recombination proteins could all be used to make a more biologically informed and accurate predictive model.

Full scale CREATE experiment

Take what this study demonstrates about the significant kinetic parameters and build an improved model for predicting editing efficiency. Design a new CREATE library using this model and perform a depletion analysis study (measure the change in frequency of guides before and after introduction of selective pressures). Choose some of the guides from across the depletion distribution and analyze them independently for toxicity/ editing efficiency. Based on the results of the individual characterization, the performance of the model can be determined – how well did the model predict the performance of guides?

Deletion of Gam protein

Delete the Gam protein since it is both toxic and inessential (Figure 18, Figure 19, (Sergueev, Yu, Austin, & Court, 2001)) and increase the promoter strength for expression of Exo and Beta to determine if constitutive expression system is in fact recombination limited, and if the editing efficiency can be increased by further increasing expression of recombination proteins beyond what was accomplished in this study. Perform a western blot to verify the increased expression of recombination machinery.

Further transformation of homology arm

A plateau in editing efficiency was observed with the group 1 guides, but not with the group two or group three guides. This could indicate that editing efficiency could be improved even further by transforming more homology arm. A suggestion for the future would be to transform even more homology arm until the maximum improvement in editing efficiency is determined.

BIBLIOGRAPHY

Amin, F., Khalid, H., Zhang, H., Rahman, S., Zhang, R., Liu, G., & Chen, C. (2017). Pretreatment methods of lignocellulosic biomass for anaerobic digestion. *AMB Express, 7*(72), 1-12.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., . . . Horvath, P. (2007, March). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science, 315*, 1709-1712.

Bassalo, M. C., Garst, A. D., Halweg-Edwards, A. L., Grau, W. C., Domaille, D. W., Mutalik, V. K., . . . Gill, R. T. (2016). Rapid and Efficient One-Step Metabolic Pathway Integration in E. Coli. *ACS Synthetic Biology, 5*(7), 561-568.

Datta, S., Costantino, N., & Court, D. (2006). A set of recombineering plasmids for gram-negative bacteria. *Gene, 379*, 109-115.

Davis, J. H., Rubin, A. J., & Sauer, R. T. (2011, February). Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Research, 39*(3), 1131-1141. doi:https://doi.org/10.1093/nar/gkq810

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., . . . Root, D. E. (2016). Optimized SgRNA Design to Maximize Activity and Minimize Off-target Effects of CRISPR-Cas9. *Nature Biotechnology*, 184-191. Retrieved July 25, 2017

Farasat, I., & Salis, H. (2016). A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLOS Computational Biology*, 1-33.

Garst, A. D., Bassalo M. C., Pines, G., Lynch, S. A., Halweg-Edwards, A. L., Liu, R., . . . Gill, R. T. (2017). Genome-wide Mapping of Mutations at Single-nucleotide Resolution for Protein, Metabolic and Genome Engineering. *Nature Biotechnology, 35*(1), 48-55. Retrieved July 25, 2017

Jiang, W., Bikard, D., Cox, D., Zhang, F., & Marrafini, L. A. (2013). RNA-Guided Editing of Bacterial Genomes Using CRISPR-Cas Systems. *Nature Biotechnology, 31*(3), 233-239.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012, August). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science, 337*(6096), 816-821.

Klemm, D., Heublein, B., Fink, H.-P., & Bohn, A. (2005). Cellulose: Fascinating Biopolymer and Sustainable Raw Material. (Angew, Ed.) *Chem. Int., 44*(22), 3358–3393.

Lal, A., Dhar, A., Trostel, A., Kouzine, F., Seshasayee, A., & Adhya, S. (2016). Genome Scale Patterns of Supercoiling in a Bacterial Chromosome. *Nature Communications*, 1-8.

Lewis, D. e. (2011). Multilevel Autoregulation of Λ Repressor Protein CI by DNA Looping in Vitro. *Proceedings of the National Academy of Sciences of the United States of America, 108*(36), 14807-14812. Retrieved Feb 12, 2018

Liang, L., Liu, R., Garst, A. D., Lee, T., Nogué, V. S., Beckham, G. T., & Gill, R. T. (2017). CRISPR enabled trackable genome engineering for isopropanol production in Escherichia coli. *Metabolic Engineering, 5*, 1-10.

Liu, R., Liang, L., Garst, A. D., Choudhury, A., Nogué, V. S., Beckham, R. T., & Gill, R. T. (2018, March). Directed combinatorial mutagenesis of Escherichia Coli for Complex Phenotype Engineering. *Metabolic Engineering, 47*, 10-20.

Mosberg, J. e. (2010). Lambda Red Recombineering in Escherichia Coli Occurs Through a Fully Single-Stranded Intermediate. *Genetics, 186*(3), 791-799. doi:doi:10.1534/genetics.110.120782

Nishimasu, H., Ran, R. A., Hsu, P. D., Konerman, S., Shehata, S. I., Dohmae, N., . . . Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell, 156*(5), 935-949.

Pines, G., Freed, E., Winkler, J., & Gill, R. (2015). Bacterial Recombineering: Genome Engineering via Phage-Based Homologous Recombination. 2015. *ACS Synthetic Biology*.

Prior, J., Lynch, M., & Gill, R. (2010). Broad-host-range vectors for protein expression across gram negative hosts. *Biotechnology Bioengineering, 106*, 326-332.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell, 152*(5), 1173-1183.

Quan, J., & Tian, J. (2009, July 30). Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. *PLOS ONE, 4*(7). doi:https://doi.org/10.1371/journal.pone.0006441

Sergueev, K., Yu, D., Austin, S., & Court, D. (2001, July 11). Cell toxicity caused by products of the pL operon of bacteriophage lambda. *Gene, 272*(1-2), 227-235.

Sharan, S. K., Thomason, L. C., Kuznetsov, S. G., & Court, D. L. (2009). Recombineering: A Homologous Recombination-based Method of Genetic Engineering. *Nature Protocols, 4*(2), 206-223.

Voet, D., & Voet, J. (2011). *Biochemistry* (Fourth Edition ed.). Hoboken, New Jersey: John Wiley and Sons, Inc.

Wang, H., La Russa, M., & Qi, L. S. (2016, April 25). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry, 85*(1), 227. Retrieved October 13, 2016

Wilson, T. E., Topper, L., & Palmbos, P. (2003). Non-homologous end-joining: bacteria join the chromosome breakdance. *Trends in biochemical sciences, 28*(2), 62-66.

SUPPLEMENTS

DCAS9 AND THE REPRESSION HYPOTHESIS, MOTIVATION FOR NEXT GENERATION SEQUENCING

During the Cas9 expression variation study, some large patches of completely edited or unedited cells started showing up on MacConkey Agar plates (*Figure 23*). A more random distribution of edited and unedited cells is expected (unless working with a non-targeting guide RNA). The genotype of the cells in these patches was investigated using Sanger sequencing. White colonies (those that should have been edited, shutting off galactokinase) were sequenced and surprisingly all of them matched the wild type sequence. Thus the hypothesis that Cas9 was associating with the genome and repressing the genes, but not actually causing a double-strand break, was born. This was enough evidence to suggest that the MacConkey Agar screen is unreliable and that there are better methods to determine genomic editing efficiency.

The first step in investigation was using plasmids expressing gRNA only (synthesis of these constructs described in methods section *BUILDING AND TESTING THE LINEAR HOMOLOGY ARMS*). Expressing a guide, the complex can associate to the genome and cause either binding or a double-strand break. Without a homology arm, a targeted genomic edit is an impossibility. The gRNA only constructs replicated the observed phenotype from the plasmids that did contain a homology arm when transformed into cells, thus suggesting that no DSB followed by homologous recombination was occurring (Figure 23).

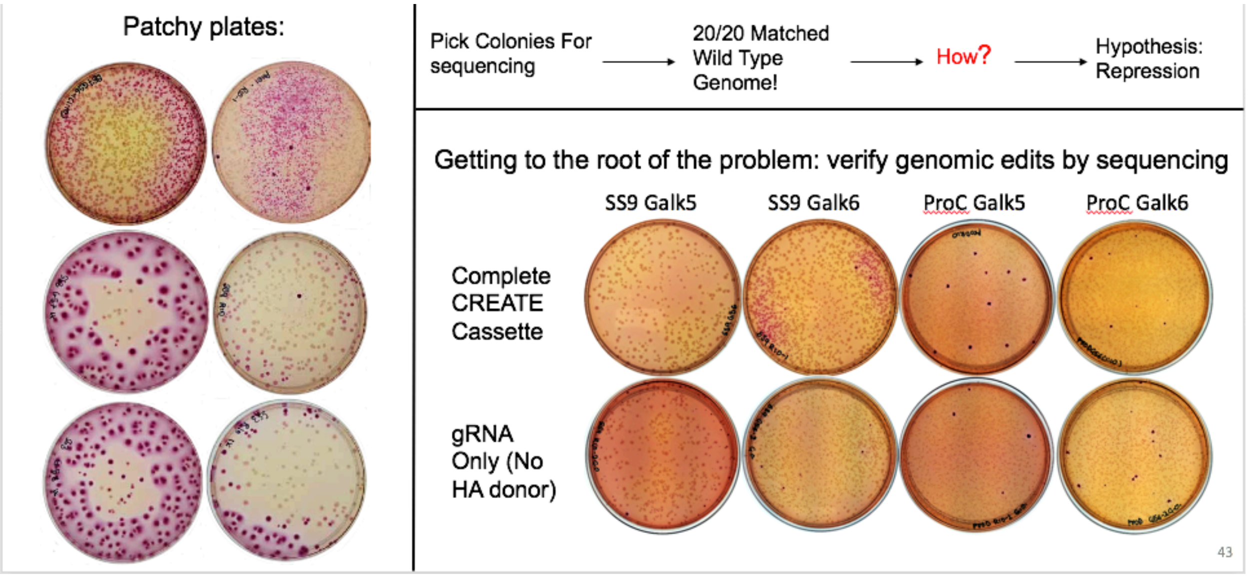


Figure 23. *Problems with the MacConkey Agar screen and the repression hypothesis.*

(Left) Patchy plates – suggestive of an unreliable screen. (Top right) workflow for arriving at the repression hypothesis. (bottom right) first steps to testing the hypothesis; used plasmids expressing gRNA only in cells expressing high levels of Cas9. The phenotype was replicated, even in the absence of a homology arm: no mechanism for making a genomic edit was provided.

To further investigate the repression hypothesis nuclease deficient, or dCas9, was expressed using the same promoters from the Cas9 expression variation study.

The dCas9 constructs were made by introducing two point mutations, D10A and H840A, into the Cas9 expression variants from an earlier study (see methods section, "building a set of varying strength promoters"). The protocol for making the two point mutations can be found in *Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression* (Qi, et al., 2013).

Two gRNA only plasmids from each group (one, two, and three) were transformed into the dCas9 strains (each of the varying promoter strengths) and plated on MacConkey agar. Group one and group three guides (Galk_1, Galk_3, Galk_5 and Galk_6) complexed to dCas9 replicated the trends observed with increasing promoter strength observed before using active cas9, supporting the repression hypothesis; the cells went unedited at the genomic level, yet phenotypically, appeared edited. The group two guides (Galk_4 and Galk_7) however, did not replicate the active cas9 trends. This was found later to be due to bias in repression for constructs targeting the non-template strand (Figure 25).

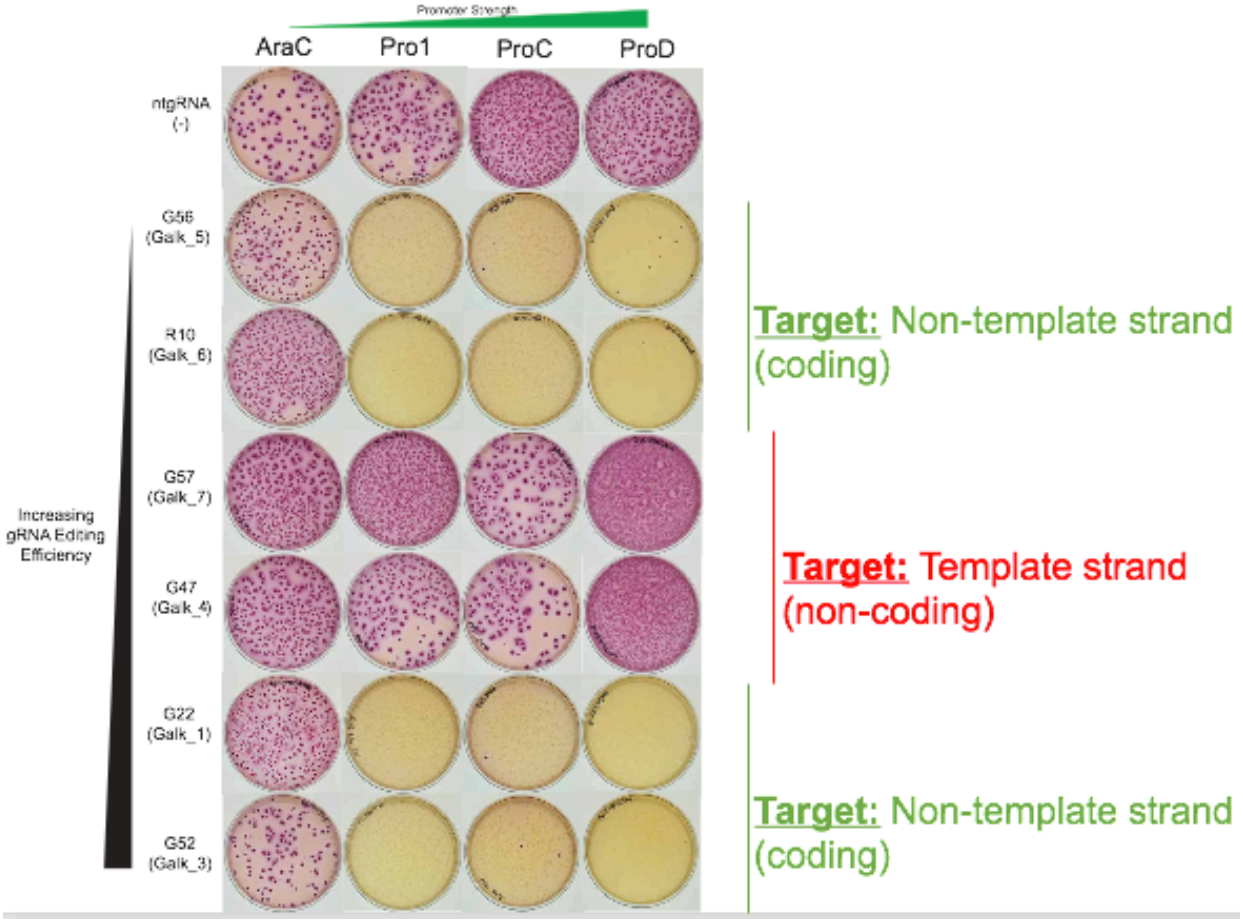


Figure 24. *Further investigation of the repression hypothesis using dCas9*

Group one and group three guides (Galk_1, Galk_3, Galk_5 and Galk_6) complexed to dCas9 replicated the active cas9 trends as promoter strength increased, supporting the repression hypothesis. The group two guides (Galk_4 and Galk_7) however, did not replicate the active cas9 trends. This is explained by Figure 25.

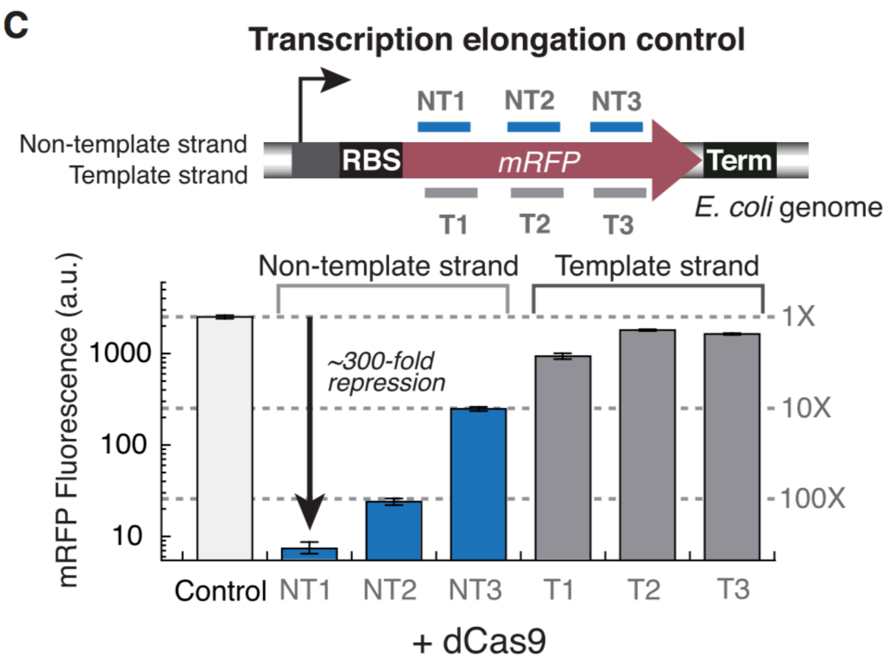
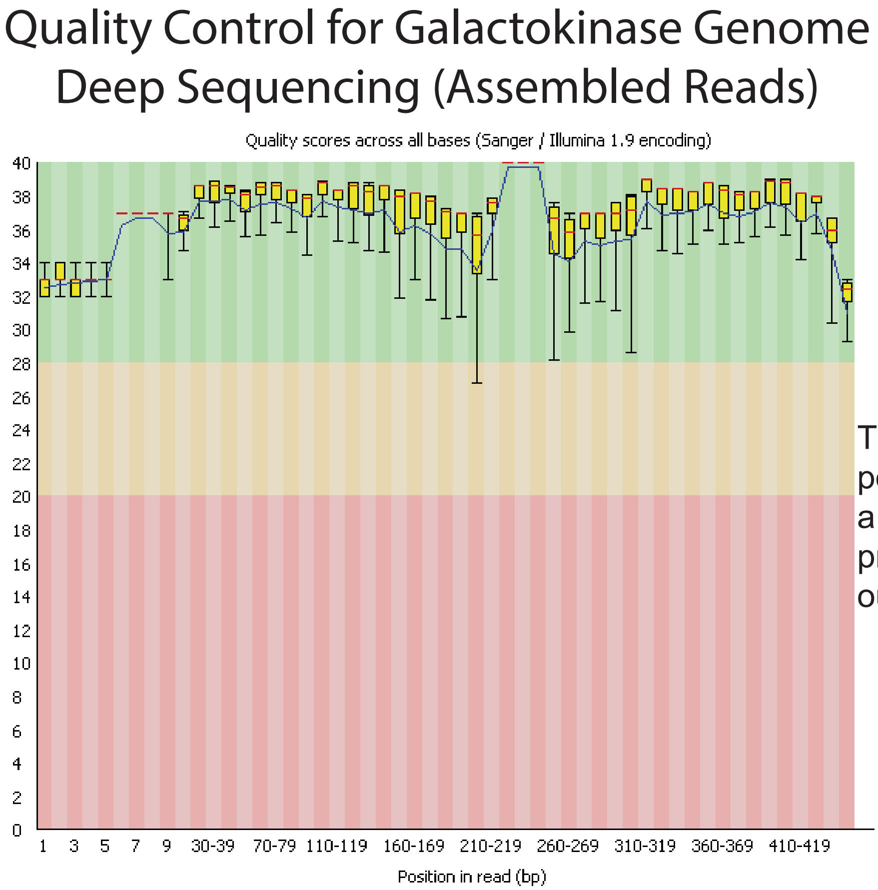


Figure 25. *Bias in dCas9 gene repression for guides targeting the non-template strand*

This figure comes from (Qi, et al., 2013) and can be use to explain the results in Figure 24; dCas9 represses genes more efficiently when targeting the non-template strand.

NEXT GENERATION SEQUENCING QUALITY CONTROL

Due to the observed difficulties with the MacConkey Agar screen for editing efficiency mentioned in "DCAS9 AND THE REPRESSION HYPOTHESIS, MOTIVATION FOR NEXT GENERATION SEQUENCING" and the high cost of Sanger sequencing individual colonies, I resorted to next generation sequencing to determine a value for actual genomic editing efficiency, rather than an apparent value using the MacConkey screen. This strategy had the added benefit of increasing the confidence of editing efficiency data.



$P_{Error} = 10^{-Q/10}$

Thus, in this case most positions have between a 1/4000 and 1/6250 probability of erroneously calling a base.

Figure 26. *Quality control from the next generation sequencing run.*

Y-axis: Phred score. X-axis: position in read. 8 million reads were divided across 156 different experimental conditions to generate the editing efficiency data in Figure 16 and Figure 20. Base calls with a Phred score lower than 30 were excluded from the dataset (low reliability).

SUPPLEMENTARY TABLES

Table 4*. Editing Cassettes*

Construct Name	GuideRNA Sequence	Strand	Homology Arm
G1 ArgP E246Q	CCAGCTCTTTCTCGATTTGC	Non-Template (Coding)	GAAGCGTTCGTACAACTTGCTCGCCAGGGCACCACCTGCTGTATGATCCCGCACttaCAGATTcagAAAGAGCTGGCCAGCGGTGAACTGATTGACTTAACGCCTGGGCTATTTCAAC

Construct Name	**GuideRNA Sequence**	**Strand**	**Homology Arm**
G2 LysP Q219I	ATCGGCGTAGCTATGATTGT	Template (Non-Coding)	CGTTTGCTGGTGGTTTTGCGGCGATGATCGGCGTAGCTATGATTGTCGGCTTCTCTTTCattGGCACGGAGCTGATCGGTATTGCTGCAGGCGAGTCCGAAGATCCGGCGAAAAACAT
G3 LysP I31K	AACAAAAAGACCTGTACCGA	Non-Template (Coding)	CGGGCTTACGCCGTGAATTAAAGGCGCGTCACCTGACGATGATTGCCATTGGCGGTagcaaaGGCACAGGTCTTTTTGTTGCCTCTGGCGCAACGATTTCTCAGGCAGGTCCGGGCGG
G4 DapF M260Y	GGTCACCCGTTATATATGAC	Template (Non-Coding)	GAACTCCCCGGCGGTCGTCTTGATATCGCCTGGAAAGGTCCGGGTCACCCGTTATATtataccggtCCGGCGGTACATGTCTACGACGGATTTATTCATCTATGAAGCAACCAGGGGA
G5 LdcC S376K	AATGCTGGCGGCGTTATCGC	Template (Non-Coding)	GGGAAAAGTGATCTTCGAAACGCAATCGACCCACAAAATGCTGGCGGCGctgagccaaGCTaaaCTGATCCACATTAAAGGCGAGTATGACGAAGAGGCCTTTAACGAAGCCTTTATG
G6 LysR S198P	CTGGCGATAGCTGTCAGTAC	Non-Template (Coding)	TGGCGGTAAAAAAGGTATTAACGCCGGATGATTTTCAGGGTGAGAACTACATCAGCCTTccgAGAACCGACAGCTATCGCCAGTTGCTGGATCAGCTATTTACTGAACATCAGGTTAA
G7 DapF G210P	CGTTTACGCGTTTATGAGCG	Template (Non-Coding)	ATATCGGTTTTATGCAAGTGGTTAAGCGCGAGCATATTCGTTTACGCGTTTATgaaagagatGCAGGAGAAACCCAGGCCTGCGGCAGCGGCGCGTGTGCGGCGGTTGCAGTAGGGAT
G8 LysC T229G	TGCGGAAACTACGCGTGGAT	Non-Template (Coding)	CGGAGGCTTTACACGCATCTCGTGTTGATATCTGGACCGACGTCCCGGGCATCTACACCggcGACCCGCGCGTAGTTTCCGCAGCAAAACGCATTGATGAAATCGCGTTTGCCGAAGC
G9 LysR T151H	GTTCTGTTCCCGCAGGCGTA	Non-Template (Coding)	TCACCGCTACTTGAAGAGTGGCTCTCGGCCCAGCGTCATGATTTAGGACTCACTGAAcatCTCcacACCCCGGCGGGAACAGAACGTACCGAATTACTCTCTTTAGATGAAGTGTGTG
G10 LysR S36R	TCTCAAAGCGCGCAAGTTCG	Non-Template (Coding)	TGACCGCCGGAAGCCTGACTGAGGCGGCACACCTGCTACACACCTCACAGCCAACCGTCcgtcgtGAGTTAGCGCGCTTTGAGAAGGTGATCGGGCTGAAATTGTTTGAGCGCGTACG
G11 LysS Y280F			
G18 S1-1	TGCGACAATGGGCGCATCGA	Non-Template (Coding)	ATGACCGTAAAGTTCGCGTGATGGCAGCCGATTATGAAAATCAGCTCGACGAGTTTagctaataaGCGCCCATTGTCGCACATGAAAACTATCAATGGGCTAACTACGTTCGTGGCGT

Construct Name	**GuideRNA Sequence**	**Strand**	**Homology Arm**
G19 S1-7	ATGATCTTTCTTGCCGAGCG	Non-Template (Coding)	GGAAGCAGAAAACCAGTTTGTAGGCTGTAACTGCGGGATCATGGATCAGCTAATTagcGCCtaataaAAGAAAGATCATGCCTTGCTGATCGATTGCCGCTCACTGGGGACCAAAGCA
G22 Galk D60	CGAACTTTACGGTCATCGCG	Non-Template (Coding)	CGACGGTTTCGTTCTGCCCTGCGCGATTGATTATCAAACCGTGATCAGTTGTGCAccgAGAtaataaCGTAAAGTTCGCGTGATGGCAGCCGATTATGAAAATCAGCTCGACGAGTTT
R10 Galk N30	TGAAGCGGAAGAACTTAACC	Non-Template (Coding)	ATCTCTGTTTGCCAACGCATTTGGCTACCCTGCCACTCACACCATTCAGGCGccgGGTAGAGTGtaataaATTGGTGAACACACCGACTACAACGACGGTTTCGTTCTGCCCTGCGCG
G27 L399C ProA	CGCGCGTGGCCCAATGGGGC	Template (Non-Coding)	TCTGGGTGCGGCAGTGGCGGTAAGCACACAAAAACTCCACGCGCGTGGCCCAATGGGCTGCGAAGCACTGACCACTTACAAGTGGATCGGCATTGGTGATTACACCATTCGTGCGTA
G28 ProA A368M	GCCGTCGGTAAAACGCGTAG	Non-Template (Coding)	TATGCGCAACGCCCAGCGTTTTGTTAACGAAGTGGATTCGTCCGCTGTTTACGTTAACATGAGCACCCGTTTTACCGACGGCGGCCAGTTTGGTCTGGGTGCGGCAGTGGCGGTAAG
G29 Pro A P169C	CCTGAAATCCTGCGGCTTAC	Template (Non-Coding)	TCGCACTAACGCTGCAACGGTGGCGGTGATTCAGGACGCCCTGAAATCCTGCGGTCTGGAAGCGGGTGCCGTGCAGGCGATTGATAATCCTGACCGTGCGCTGGTCAGTGAAATGCT
G39 XylA Q70 KO	GCCTCACCAGGCTGCTGCCA	Non-Template (Coding)	tcctctggcggaaagcctagaatgcactagGCAATAAAATGGCACATGTAACTTGTGGAAAAACTCAAATGCGACATCTGCTTTACGCTTCGCCAACGCCAGTGCCTCACCAGGttattaCCAtggACGATTAAACGCCCCCACACCAAACATATCCGCCCCGTTCCAGCAGAAGGTGTGCCAGTAGCAGGCGGCAAAACGCAAGTGCTCTTCCA
G40 Xyla F145 KO	GCGAGGGTTTGTAAAGCAGT	Non-Template (Coding)	tcctctggcggaaagcctactgtacgactgATCGCTGTAACAACTTGCGTTGCCGCCCAGCTGAAGACTTCAGGATCTGGGTTCGTCGCCGCACCCGCGCCGTAGCGAGGGTTttattaGCAGTTcgcCGTTCCCCACAGCAGCTTCACGCCGCTCTCTTCTTGCTTGCCTGCCAGGACATCAACCATTTGCGCAAAATTATTGATGTACTCTTT
G 41 Galk Y145 KO	ATGATAAAGCTGCTGCAATA	Non-Template (Coding)	tcctctggcggaaagccttgcttcactacgGATCCCGCAGTTACAGCCTACAAACTGGTTTTCTGCTTCCTGACCGTTAAGCGCGATTTGTGCGCCGTCCAGCGGCAGttattaAAGCTGCTGCAATACcgtTCCGACCGCGACTTCCAGTGAAGCGGAAGAACTTAACCCGGCACCCTGCGGCACATTGCCGCTGATCACCATGTCCACGCCGC

Construct Name	**GuideRNA Sequence**	**Strand**	**Homology Arm**
G42 XylA R10_KO	GCGATCGAGCTGGTCAAAAT	Non-Template (Coding)	tcctctggcggaaagcctactatgcaacgtCGCTTACCCAACACCAGTTCGTCGGGATTGTAGTGACGGAATGCTAACGGGTTTGAGGATTTTGAGCCTTCATAACGttattaATCGAGCTGGTCAAAATAcgcTTGCATATTGAACTCCATAATCAGGTAATGCCGCGGGTGATGGATGATGTCGTAATATTGGGCACTCCCTTTCAGTTGCTC
G43 Galk D70 KO	GTCGAGCTGATTTTCATAAT	Non-Template (Coding)	TCCTCTGGCGGAAAGCCTGGACTAGCGATGACCACGCCACGAACGTAGTTAGCCCATTGATAGTTTTCATGTGCGACAATGGGCGCATCGAGGGAAAACTCGTCGAGCTGATTTTCTTATTACGCTGCCATCACGCGAACTTTACGGTCATCGCGTGGTGCACAACTGATCACGGTTTGATAATCAATCGCGCAGGGCAGAACGAAACCGTCGTT
G44 GalkD60	CTGCCATCACGCGAACTTTA	Non-Template (Coding)	CGACGGTTTCGTTCTGCCCTGCGCGATTGATTATCAAACCGTGATCAGTTGTGCAccgAGAtaataaCGTAAAGTTCGCGTGATGGCAGCCGATTATGAAAATCAGCTCGACGAGTTT
G45 GalkD150	GCAGCTTTATCATCTGCCGC	Template (Non-Coding)	TACGAGAGGATCCCGCTTCACTGGAAGTCGCGGTCGGAACCGTATTGCAGCAGCTTTATCATTTACCATTATAATAAGCACAAATCGCGCTTAACGGTCAGGAAGCAGAAAACCAGTTTGTAGGCTGTAA
G46 GalkD40	GAACACACCGACTACAACGA	Template (Non-Coding)	CATGAGTCATGCCTCACACCATTCAGGCGCCTGGCCGCGTGAATTTGATTGGTGAACACACCGACTACAATTAATAATTCGTTCTGCCCTGCGCGATTGATTATCAAACCGTGATCAGTTGTGCACCACG
G47 GalkE160	ACAAATCGCGCTTAACGGTC	Template (Non-Coding)	GAGTGACTGGTATATTGCAGCAGCTTTATCATCTGCCGCTGGACGGCGCACAAATCGCGCTTAATGGCCAATAATAAGAAAACCAGTTTGTAGGCTGTAACTGCGGGATCATGGATCAGCTAATTTCCGC
G48 C170	GAAGCAGAAAACCAGTTTGT	Template (Non-Coding)	AGCGCTAGAGGTCTGGACGGCGCACAAATCGCGCTTAACGGTCAGGAAGCAGAAAACCAGTTCGTGGGTTGTAACTAATAAATCATGGATCAGCTAATTTCCGCGCTCGGCAAGAAAGATCATGCCTTGC
G52 Galk L80	TGCGACAATGGGCGCATCGA	Non-Template (Coding)	ATGTATCCAGTGATGACCGTAAAGTTCGCGTGATGGCAGCCGATTATGAAAATCAGCTCGACGAGTTTAGCTAATAAGCGCCCATTGTCGCACATGAAAACTATCAATGGGCTAACTACGTTCGTGGCGT
G53 Galk N120	GATCAGCGGCAATGTGCCGC	Template (Non-Coding)	TGTCGAGATAGGACATCTGCAACTGCGTAACAACAGCTTCGGCGGCGTGGACATGGTGATCAGCGGCTAATAACCACAAGGTGCCGGGTTAAGTTCTTCCGCTTCACTGGAAGTCGCGGTCGGAACCGTA
G54 Galk N30	GTTCACCAATCAAATTCACG	Non-Template (Coding)	AGATCTGTAGTCCTCTGTTTGCCAACGCATTTGGCTACCCTGCCACTCACACCATTCAGGCGCCTGGCCGTGTTTAATAAATTGGTGAACACACCGACTACAACGACGGTTTCGTTCTGCCCTGCGCGAT

Construct Name	**GuideRNA Sequence**	**Strand**	**Homology Arm**
G55 Galk S130	GACCGCGACTTCCAGTGAAG	Non-Template (Coding)	GCTAAGTCTGACTCGGCGGCGTGGACATGGTGATCAGCGGCAATGTGCCGCAGGGTGCCGGGTTAAGTTCTTAATAAAGCCTGGAAGTCGCGGTCGGAACCGTATTGCAGCAGCTTTATCATCTGCCGCT
G56 Galk T20	AGGCGCCTGAATGGTGTGAG	Non-Template (Coding)	GCGAAGTGTACGAAATGAGTCTGAAAGAAAAAACACAATCTCTGTTTGCCAACGCATTTGGCTACCCTGCGTAATAAACCATTCAGGCGCCTGGCCGCGTGAATTTGATTGGTGAACACACCGACTACAA
G57 Galk V100	GGCTAACTACGTTCGTGGCG	Template (Non-Coding)	CACGCAAGACAGTCGATGCGCCCATTGTCGCACATGAAAACTATCAATGGGCTAACTACGTTAGAGGTGTTTAATAACATCTGCAACTGCGTAACAACAGCTTCGGCGGCGTGGACATGGTGATCAGCGG
G58 Galk V140	TTCCGCTTCACTGGAAGTCG	Template (Non-Coding)	TGTCCATCGCATCGGCAATGTGCCGCAGGGTGCCGGGTTAAGTTCTTCCGCTTCACTGGAGGTGGCCGTCGGAACCTAATAACAGCAGCTTTATCATCTGCCGCTGGACGGCGCACAAATCGCGCTTAAC
G59 Galk_Off	ATGATAAAGCTGCTGCAATA	Non-Template (Coding)	TCCGCTTCACTGGAAGTCGCGGTCGGAACGGTATTGCAGCAGCTTTAACATCTGCCGCTGGACGGCGCACAAATCGCGCTTAACGGTCAGGAAGCAGATC

Table 5. *Primer Sequences.*

Primer number (referenced in text):	**Primer Name:**	**Primer Sequence (5' to 3'):**
1	HA_Insert_Amp_F	TAGCTTTCGCTAAGGATGAT
2	HA_Insert_Amp_R	ACTAGTATTATACCTAGGAC
3	CREATE_BB_F	GTTTTAGAGCTAGAAATAGCAAGTTAA
4	CREATE_BB_R	CCAGAAATCATCCTTAGCGAAAGCTAA
5	Amp_F_Seq_For	ATAAGGGCGACACGGAAATGTTGAATACTCATACT
6	Amp_F_Seq_Rev	CCAAAACAGCCAAGCTGGAGACC
7	HA_Iso_For	TTCTGGAATTCTAAAGATCT
8	HA_Iso_Rev	TTCTGGAATTCTAAAGATCT
9	HA_BB_AMP_F	TTGACAGCTAGCTCAGTCCT
10	HA_BB_AMP_R	AGATCTTTAGAATTCCAGAA
11	GalK_45_for	CGTCAGATGTGTATAAGAGACA
12	GalK_45_rev	ACCGTTAAGCGCGATTTGTG
13	Pro1Psim_for	ATCTATATTCAGGGAGACCAACGCACTGACCACCATGAAG

14	Pro1psim_rev	ACCGAGCCTTATGCATGCCCGTAAAGTTATTATCACCGCAGATGGTT
15	Pro1Psim_Seq_F	CATAGCAATTCAGATCTCTC
16	Beta Seq_F	GATGTTGAGTTTCAGGAGCG
17	Gam_par_for	GAATCGCTATGCATCGATC
18	Gam_Par_rev	CTCCTGAAATAGCTGTGA
19	Gam_Seq_F	GATCTCTCACCTACCAAACAATG
20	Gam_Seq_R	ATGCGATGAACTGCGCAT
21	Exo_Knockout_for	CGCATCCTCACGATAATATC
22	Exo_Knockout_Rev	TCATGCTGCCACCTTCTGCT
23	Exo_Knockout_Seq_For	CATGGGATGACGACTTATTG
24	Exo_Knockout_Seq_Rev	GTCCAGCTGTCTCTTATACA
25	recA_gRNA_f	TTGACAGCTAGCTCAGTCCTAGGTATAATACTAGTacggatgtcgagacgaacag
26	recA_gRNA_r	CCTTATTTTAACTTGCTATTTCTAGCTCTAAAACctgttcgtctcgacatccgt
27	Gam_knockout_for	AACGAATGAGTACTGCACTC
28	Gam_Knockout_rev	GGTGGTCAGTGCGTCCTGCT

Table 6. *RecA Deletion Cassette Sequence.*

CAATTTCTACAAAACACTTGATACTGTATGAGCATACAGTATAATTGCTTCAACAGAACATATTGACTATCCGGTATTACCCGGCATGACAGGAGTAAAATCGTCTTGTTTGATACACAAGGGTCGCATCTGCGGCCCTTTTGCTTTTTTAAGTTGTAAGGATATGCCATGACAGAATCAACATCCCGTCGCCCGGCATA