COVER SHEET

TITLE: The RecA-dependent endonuclease Ref: characterization of DNA binding, nuclease activity, and use for *in vivo* genome editing in *Escherichia coli*

AUTHOR'S NAME: Tayla M. Olsen

MAJOR: Biochemistry

DEPARTMENT: Biochemistry

MENTOR: Michael M. Cox

DEPARTMENT: Biochemistry

YEAR: Senior, 2014

ABSTRACT

The RecA-dependent endonuclease Ref: characterization of DNA binding, nuclease activity, and use for *in vivo* genome editing in *Escherichia coli*

Bacteriophage Ref protein is an HNH class, RecA-dependent endonuclease. RecA-bound oligonucleotides complementary to a region on circular dsDNA will catalyze RecA strand-invasion, creating a displacement loop (D-loop). Ref can be directed to create double-strand breaks within the D-loop based only on the sequence of the oligonucleotide used. The Ref protein consists of a globular domain containing the active-site and 76 disordered N-terminal residues responsible for DNA binding. Characterization of N-terminal truncations revealed modulated activity for DNA binding and double-strand break formation, this system has been explored *in vivo*. *Escherichia coli* containing RecA, Ref, and a targeting oligonucleotide demonstrate oligo-recombination induced by DSBs created by Ref; this has potential for being implemented in genome editing.

Tayla M. Olsen, Biochemistry
Author Name/Major

Author Signature

Michael M. Cox, Biochemistry
Mentor Name/Department

Mentor Signature

May 1, 2014
Date

The RecA-dependent endonuclease Ref: characterization of DNA binding, nuclease activity, and use for *in vivo* genome editing in *Escherichia coli*

Tayla M. Olsen, Angela J. Gruber, Rachel H. Dvorak, and Michael M. Cox
Department of Biochemistry, University of Wisconsin-Madison

INTRODUCTION

Recombination enhancement function, Ref, is a 21-kDa, HNH class, RecA-dependent endonuclease originally studied from bacteriophage P1 in the 1980s. During this time it was determined that bacteria lysogenic for phage containing *ref* resulted in increased RecA- and RecBCD-dependent recombination events via an unknown mechanism (1, 2). This led to the initial hypothesis that Ref was a positive regulator of RecA. Upon further investigation, it was discovered that Ref was capable of binding single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA) irrespective of cofactors. In the presence of RecA, ATP, and Mg^{2+}, which forms an active RecA nucleoprotein filament, Ref produces extensive degradation of ssDNA, but not dsDNA (3). Gruenig *et al.* described that RecA-bound oligonucleotides (oligos) complementary to a specific region on circular dsDNA (cdsDNA) catalyzed RecA strand-invasion creating a displacement loop (D-loop) (3). During strand-invasion, the RecA-bound oligo base pairs with its complement (the paired strand) and the non-complementary strand becomes distal (the displaced strand). Upon Ref addition, double-stranded breaks (DSBs) are created within the D-loop. Thus, Ref is a programmable endonuclease via the targeting oligo sequence and can be directed to create a double-strand break within a D-loop formed by RecA.

Ref creates DSBs via a two-step nicking mechanism in a RecA-dependent manner (4). Ref first nicks the DNA strand that is complementary to the oligo (the paired strand), then with some delay, nicks the strand homologous to the oligo (the displaced strand), creating a DSB (4). These nicking steps are distinct from one another and are optimized for different reaction conditions (4). The second cleavage event on the displaced strand requires ATP hydrolysis

(ATPase) activity of RecA. This implicates that disassembly of RecA from DNA is necessary for Ref to access the displaced strand (4). Interactions with RecA are usually dependent on having an active RecA nucleoprotein filament, i.e. one that is activity hydrolyzing ATP, such that interacting proteins can bind in the RecA filament groove or at filament ends. In the early characterization of Ref, it was found that a non-cleavable form of LexA known to interact in the filament groove inhibited Ref nuclease activity suggesting an interaction of Ref within the groove (3, 4). The dependence of Ref on the ATPase activity of RecA implies a direct interaction between Ref and RecA, although the direct interaction has not been demonstrated.

The crystal structure of the C-terminal residues (77-186 amino acids) containing the active site of P1 Ref has been determined to 1.4 Å resolution. The N-terminal domain is disordered, though the crystallized protein is full-length (PDB ID: 3PLW) (3). The crystal structure of Ref shows an asymmetric monomer that contains two stably bound Zn^{2+} ions (3). Structural analysis of Ref shows that it is similar to HNH endonucleases and contains the signature ββα metal-binding core (4). Other HNH homing endonucleases have similar catalytic mechanism as Ref. Ref does not exhibit any sequence specificity for DNA binding or nuclease activity, though Ref preferentially cleaves DNA upstream of a pyrimidine base (4).

P1 Ref protein missing the N-terminal 76 amino acids (ΔN76) is incapable of DNA binding and is deficient in circular ssDNA (cssDNA) cleavage activity (3). A 10-fold increase in P1 Ref ΔN76 concentration restores nuclease activity on cssDNA and targeted-DSB nuclease activity (unpublished). The unstructured N-terminal domain is positively charged (25/76 amino acids) supporting its role in DNA binding. Visual inspection of the sequence revealed a possible signature charge distribution motif, which was used to guide the design of N-terminal truncation

variants. We hypothesize that the N-terminal domain modulates nuclease activity via control of N-terminal DNA binding.

Because Ref can be directed to cleave a specific DNA structure generated by RecA and a complementary oligo, the Ref/RecA system has potential to be used for *in vivo* genome editing. Strategies of genome editing often utilize DSBs to direct repair and RecA-mediated homologous recombination with a fragment of DNA containing a desired mutation. A similar technique, CRISPR/Cas9, has become a popular mode of *in vivo* genome editing in recent years. CRISPR/Cas9 uses an RNA guide rather than a DNA guide to direct induced incorporation of a desired mutation (6). The CRISPR/Cas9 system has some limitations including the requirement of a protospacer adjacent motif (PAM) that limits editing possibilities to ~40% of the genome (4). A short homology-matching of the RNA guide (20-40 bases) means increased potential for off-target effects, which has surfaced as problematic.

Our investigation of N-terminal truncations variants shows this region is required for DNA binding and is necessary for efficient Ref-mediated DNA cleavage. Specifically, Ref N-terminal truncation variants lacking between 21 and 47 amino acids are more effective RecA-dependent targeting nucleases than full length Ref. We also provide evidence that the Ref/RecA system is functional for *in vivo* genome editing in *Escherichia coli*.

METHODS AND MATERIALS

DNA substrates – The M13mp18 cssDNA (7249 nucleotides) was prepared as described previously (7, 8). The concentration of circular ssDNA was determined by absorbance at 260-nm using 108 μM nt ml^{-1} A_{260} $^{-1}$ as the conversion factor. The M13mp18 circular dsDNA was prepared as described previously (7 – 9). The concentration of circular dsDNA was determined

by absorbance at 260-nm using 151 μM nt ml^{-1} A_{260} $^{-1}$ as the conversion factor. All DNA concentrations are given in μM nucleotides (μM nt). Oligonucleotides were purchased from Integrated DNA technologies. Sequences of oligonucleotides used are represented in Table 1.

Proteins – The *E. coli* RecA E38K protein was purified as described previously (4). The protein was flash frozen in liquid N_2 and stored at -80°C. The concentration of the purified RecA protein was determined from the absorbance at 280 nm using the extinction coefficients of 2.23 x 10^4 M^{-1} cm^{-1}.

The P1 WT Ref and P1 ΔN76 proteins were purified as described previously and concentration was determined using an extinction coefficient of 2.851 x 10^4 M^{-1} cm^{-1} and 1.5220 x 10^4 M^{-1} cm^{-1}, respectively (3).

The φW39 WT Ref and φW39 ΔN59 Ref were purified with procedures using similar growth, induction, cell harvesting, and early fractionation steps as P1 WT Ref. Both of these over-expression vectors were transformed into BL21 (DE3) cells. Purification entailed polyethyleneimine precipitation, precipitation with $(NH_4)_2SO_4$, and chromatography successively using butyl-Sepharose (120 mL CV), Source 15 Q, Source 15S, and Sephacryl S-100 gel filtration columns. This was followed by another butyl-Sepharose (20 mL CV) chromatography step. Before cell lysis, the protease inhibitor Pefabloc SC (Sigma) was added to .1 mg/ml final concentration. Concentration was determined using an extinction coefficient of 2.788 x 10^4 M^{-1} cm^{-1} for φW39 WT Ref and 1.5220 x 10^4 M^{-1} cm^{-1} for φW39 ΔN59.

The φW39 ΔN21 Ref, φW39 ΔN47 Ref, φW39 ΔN66 Ref, and φW39 ΔN74 Ref were purified similarly to the above proteins with one major modification: the overexpression vectors were transformed into BLR (DE3) cells which prevents RecA contamination during the purification. Concentration was determined using an extinction coefficient of 2.219 x 10^4 M^{-1}

cm^{-1} for φW39 ΔN21, 1.5220 x 10^4 M^{-1} cm^{-1} for φW39 ΔN47, φW39 ΔN66, and φW39 ΔN74. All proteins were rigorously tested for exonuclease and endonuclease contamination and all were free from detectable nuclease activity in the absence of RecA protein.

The TMO Hybrid Ref was purified with procedures using similar growth, induction, cell harvesting, and early fractionation steps as P1 WT Ref. Similarly, the over-expression vector was transformed into BL21 (DE3) cells. Purification included polyethyleneimine precipitation, precipitation with $(NH_4)_2SO_4$, and chromatography successively using butyl-Sepharose (120 mL CV), hydroxyapatite, Source 15Q, and SP Fast Flow (GE17-5157-01) columns. These steps were followed by another butyl-Sepharose (20 mL CV) chromatography step. Significant degradation of the protein occurred during the hydroxyapatite step and successive chromatography steps were performed to purify the specific-sized degradation product. Electrospray ionization and tandem mass spectrometry were used verify the amino acid position of degradation. Concentration was determined using an estimated extinction coefficient of 1.6500 x 10^4 M^{-1} cm^{-1} inferred by the mass spectrometry data.

Cloning of Ref and Ref Variants – φW39 WT Ref was made by mutagenesis of pEAW584, which is WT P1 Ref in protein expression vector pET21A (Novagen). The mutations were I8L, C11R, K13E, V34A, T46S, L47I, R86T, A93T, N107E, and A115S. Silent mutations to improve codon usage were also made and the resulting plasmid is pTMO9.

φ*W39 ΔN59 Ref* - pTMO9 was used as the template in a PCR with an upstream primer consisting of an NdeI site followed by bases 178-202 of φW39 Ref, with a G to T change at base 192, and a T to C change at base 198 to make silent mutations for better codon use at amino acid 64 and 66. The downstream primer consists of a BamHI site followed by bases 544-561 of the φW39 Ref gene in 5' to 3' orientation. The PCR product was digested with NdeI and BamHI and

inserted into pET21A digested with the same. The resulting plasmid, pEAW920 was confirmed by sequencing.

φ*W39 ΔN21 Ref*- pEAW886 was constructed in the same manner as pEAW920 except the upstream primer consisted of bases 64-101 of φW39 Ref, with several changes for improved codon usage.

φ*W39 ΔN47 Ref* -pEAW865 was constructed in the same manner as pEAW920 except the upstream primer consisted of bases 142-154 of φW39 Ref, with a C to T change at base 147 to make a silent mutation for better codon use at amino acid 49.

φ*W39 ΔN66 Ref* – pEAW919 was constructed in the same manner as pEAW920 except the upstream primer consisted of bases 199-222 of φW39 Ref, with an A to C change at base 210 to make a silent mutation for better codon use at amino acid 70.

φ*W39 ΔN74 Ref* – pEAW971 was constructed in the same manner as pEAW920 except the upstream primer consisted of 223-248 of φW39 Ref, with a T to C change at base 223, a G to T change at base 231, and an A to T change at base 234 to make silent mutations for better codon use at amino acid 75, 77, and 78.

TMO hybrid Ref vector is pTMO8, which is P1 Ref, in pET21A, with mutations at I8L, C11R, K13E, V34A, T46S, L47I, N107E, and A115S. It lacks the R86T and A93T mutations found in φW39 WT Ref.

Targeted-DSB Nuclease Assay – RecA-bound oligonucleotides complementary to a specific region on cdsDNA catalyze RecA strand-invasion creating a displacement loop (D-loop); Ref creates DSBs within the D-loop. The amount of ldsDNA formed over time by each Ref construct was determined. Procedure adapted from Ronayne (4).

Reactions were incubated with Buffer A, the ATP regeneration system (10 U/mL pyruvate kinase and 3.5 mM phosphoenolpyruvate), 4 μM n.t. 150-n.t. long oligonucleotide (rlb1) and 1.33 μM RecA E38K at 37ºC for 10 min. Reactions were incubated an addition 20 min with 3 mM ATP to allow an active RecA nucleoprotein filament to form. 8 μM n.t. M13mp18 cdsDNA is followed by another 20 min incubation. After incubation Ref is added at 0.05, 0.1, and 0.2 μM. Time points are taken at 0, 10, 30, 60, 120, and 180 min by adding 20 μL of reaction mixture to 10 μL of stop solution. Samples are incubation at 37 °C for another 30 min. Samples are analyzed by electrophoresis in 0.8% agarose with TAE buffer (40 mM Tris-Acetate 1 mM EDTA), stained with SYBR-Gold nucleic acid stain and visualized. Band intensity is quantified by ImageQuant software to numerically compare Ref constructs in assay.

Fluorescence Polarization Assay – To determine DNA binding affinity for each Ref construct, Ref were titrated into a solution containing either linear ssDNA (lssDNA) or ldsDNA and read in a 96-well, black polystyrene, assay plates (Corning).

Ref proteins at 0.5-10,000 nM were incubated with 2 nM AJM25 71mer ssDNA or AJM25 annealed with AJG52 71mer to make dsDNA substrate in 25 mM Tris-acetate (pH 8.5), 3 mM potassium glutamate, 15 mM magnesium acetate, and 5% w/v glycerol at room temperature for 15 minutes. Fluorescence anisotropy (FA) was measured at 25°C, using a Tecan Infinite M1000 instrument with 470-nm excitation and 535-nm emission wavelengths for at least three replicates. The average FA values were plotted with one standard deviation of the mean shown as error. GraphPad Prism software was used to convert FA values to the fraction of DNA bound and apparent dissociation constants were the Ref concentration needed for 50% binding saturation indicated as $K_{d.}$

Electrocompetent cells for *in vivo* targeting – Plasmids containing the empty vector (pET21A), P1 Ref only (pEAW584), RecA only (pEAW260), and RecA and P1 Ref (pEAW692) were transformed into chemically competent MG1655 (DE3) *ΔrecA* (strain EAW20 (DE3)). Each of these cell types were made electrocompetent by growing up to an OD_{600} of ~0.80 and stopping growth on ice for 15 minutes. Cells were pelleted between washing by centrifugation for 10 minutes at 6000 rpm; cells were washed three times with ice cold water. Final pellet was resuspended in 0.5 mL cold 10% glycerol, 40 μL aliquoted were made and cells were frozen in liquid N_2 before storing in a -80°C freezer. When cells were needed, they were thawed on ice.

***In vivo* targeting with the RecA/Ref system** – To determine if RecA/Ref can be used to incorporate a desired mutation into the *E. coli* MG1655 genome, electrocompetent cells were made as described above, an oligo was introduced by electroporation, and cells were recovered overnight and plated appropriately. Colony forming units (cfu) were counted and the fraction of streptomycin resistance cells per total cfu was determined as efficiency. Normalized efficiency was calculated as efficiency of cell type expressing RecA/Ref divided by efficiency of the strain expressing only RecA.

Appropriate cell aliquots were thawed on ice and electroporated at 2.5 kV after adding 10 μL of appropriate oligo at 100 μM (all oligos resuspended in water from lyophilized pellets). 1 mL of room temperature LB was added immediately to electroporation cuvette and cell-suspension was transferred to 4 mL room temperature LB and recovered with aeration overnight at 37°C. Dilutions of overnight were made and 100 μL of 10^{-6} cells were plated on LB, 1 mL of the overnight was plated on streptomycin (100 μg/mL) and grown overnight at 37°C. Colony forming units (cfu) were counted the next day. Error bars represented as standard deviation of 3 or more independent experiments.

RESULTS

A repeating pattern in the charge distribution in the N-terminal domain – An examination of sequence databases revealed very few homologues of the P1 Ref protein, essentially all of them were encoded by bacteriophages or bacteriophage remnants; bacteriophage φW39 and P7 Refs were the most closely related (3). Nonexistent laboratory stocks of φW39 and P7 phage necessitated the use of DNA cloning techniques to mutate P1 Ref to φW39 and P7 Ref. Sequence identity of >90% between P1 Ref and P7 and φW39 Ref made this feasible. Unfortunately, P7 Ref, containing an additional 30 amino acids on the C-terminus as well as 13 amino acids dissimilar from P1 Ref, proved to be insoluble in several conditions tested and could not be analyzed (unpublished). Purification of φW39 Ref and subsequent characterization of the protein showed RecA-dependence identical to P1 Ref and was at least as active in targeted nuclease assays. Because of this, the two proteins are used interchangeably in this study as "Ref," any discrimination will be indicated.

The φW39 Ref has ten amino acid differences in comparison to the P1 Ref (six of these in the unstructured N-terminal domain); sequence alignment of these two proteins identifies these differences (Figure 1A). Also included in the alignment is a variant we refer to as the TMO hybrid Ref protein, which was serendipitously cloned and purified. The TMO hybrid Ref was an intermediate in the construction of the φW39 Ref. Of the four amino acid differences in the C-terminal domain, only two of the changes characteristic of the conversion from P1 Ref to φW39 Ref was completed: N107E and A115S (Figure 1A). The TMO hybrid Ref is also missing 47 amino acid residues from the N-terminus due to a specific but apparently spontaneous degradation that occurred during purification.

After the purification of the degraded TMO hybrid Ref a closer look at the N-terminus region revealed a signature charge pattern, (++..+.+ - +), repeated three times as shown in the N-terminal segment in Figure 1B. The interpretation was that this pattern may have influenced the precise degradation of the N-terminal domain of TMO hybrid Ref. Using this pattern as a guide, we cloned a series of specific truncation variants that removed one or more of these motifs (Figure 1B). The Ref protein truncations purified included φW39 ΔN21, ΔN47, ΔN59, ΔN66, and ΔN74 (Figure 1C). The P1 ΔN76 Ref previously purified was included in many assays for comparison purposes as it lacks DNA binding capability (3). Protease inhibitors were used to prevent any further spontaneous degradation during purification of the Ref variants. The molecular weight of each variant was verified by mass spectrometry following purification.

Removing charge motifs in the N-terminal domain decrease DNA binding affinity – The P1 ΔN76 Ref truncation protein is incapable of DNA binding, which clearly indicates that the N-terminal 76 amino acids are necessary for this function (3). To compare N-terminal domain function in DNA binding, fluorescence polarization (FP) assays with fluorescently labeled ssDNA or dsDNA substrates were performed to determine apparent dissociation constants ($K_{d,\ app}$) of each of the N-terminal domain truncations. The DNA binding site size of Ref is needed to determine the true K_d values, however this is unknown and likely changes with each truncation variant. Thus, our data analysis wass limited to reporting apparent K_d values that provide a measure of DNA binding affinity for comparison between Ref variants.

When assayed with a labeled 71mer ssDNA (AJM25), the P1 and φW39 Ref proteins display similar $K_{d,\ app}$ values (14.4 nM and 27.5 nM, respectively) (Figure 2A, 2C). Upon removal of the first of the three charge motifs in the φW39 ΔN21 and ΔN47 Ref proteins, there was a six-fold reduction in binding (119.3 nM and 112.9 nM, respectively) (Figure 2A, 2C).

When the second charge motif was removed (φW39 ΔN59), a 40-fold reduction in binding affinity was observed. The φW39 ΔN66, which has a partial deletion of the third motif, showed a marked reduction in ssDNA binding and DNA saturation was not possible. When all charge motifs were removed, φW39 ΔN74 and previously in P1 ΔN76 Ref proteins, DNA binding was no longer observed (Figure 2C). Our results suggest that each charge motif is important for ssDNA binding affinity for the Ref protein.

Ref N-terminal deletion variants were similarly assayed with 71mer dsDNA substrate. The dsDNA substrate was made by annealing the labeled 71mer oligonucleotide (AJM25) with the complementary 71mer oligonucleotide (AJG52). The P1 Ref and φW39 Ref proteins displayed very similar binding constants (13.5 nM and 23.8 nM respectively) (Figure 2B, 2C). Similar to results acquired in the ssDNA FP assay, binding affinity for dsDNA diminished as each charge motif was removed. The φW39 ΔN21 Ref (56.1 nM) and φW39 ΔN47 Ref (142.5 nM) did not have similar $K_{d,\,app}$ values as seen with the ssDNA. When the second motif and part of the third motif were removed in the φW39 ΔN59 and ΔN66, dsDNA binding was predictably reduced, and φW39 ΔN74 and P1 ΔN76 Ref had no observable dsDNA binding. Overall, we provide evidence that the N-terminal region of Ref is necessary for DNA binding.

Partial N-terminal Ref truncation enhances targeted double-strand break formation – We measured the efficiency of targeted DSB formation with each Ref N-terminal deletion variant using a previously established assay (4). In this reaction, supercoiled M13mp18 cdsDNA is incubated with a 150mer complementary oligonucleotide and the RecA protein, which pairs the oligonucleotide to the cdsDNA creating a D-loop (Figure 3A) (4). Within this D-loop, Ref protein cleaves the strand paired to the oligonucleotide first, which is then followed by a slower cleavage event of the displaced strand to create a 7.25-kb linear dsDNA (ldsDNA) product

(Figure 3A) (4). A representative agarose gel of the products formed in the presence of φW39 Ref variants is shown in Figure 3B. The nicked cdsDNA and ldsDNA were measured as a percentage of the total DNA in each lane and followed over time (Figure 3C, 3D respectively).

Quantification of ldsDNA product revealed that after three hours the P1, φW39, and φW39 ΔN59 Ref proteins were identical in reaction kinetics with about 63% of initial supercoiled DNA converted to linear product (Figure 3D). However, the φW39 ΔN21 and ΔN47, and TMO hybrid Ref improved ldsDNA product formation by15-20% (Figure 3D). At 30 minutes the φW39 ΔN21 and ΔN47, and TMO hybrid Ref delineate kinetically from P1 and φW39 Refs by exhibiting a greater decrease in nicked DNA, which corresponds to faster second-strand cutting (ldsDNA formation) (Figure 3C). The ldsDNA time course supports these results: at 30 minutes we observed that the φW39 ΔN21 and ΔN47, and TMO hybrid Ref produce about 10% more ldsDNA than P1 and φW39 Refs (Figure 3C, 3D). The apparent increase in ldsDNA formation from these Ref variants indicates that removal of the first charge motif is beneficial for the efficiency of the Ref targeted nuclease activity.

Though removal of the first charge motif increases targeted nuclease activity, subsequent removal of the second and third motif, as in variants φW39 ΔN66 and ΔN74, showed decreased targeted nuclease activity. The φW39 ΔN66 Ref showed a two-fold reduction in targeted nuclease activity, indicating that the removal of the 7 amino acids following the second charge motif reduces Ref nuclease activity likely due to a severe loss of the ability to bind DNA (Figure 2). Removal of the remaining third charge motif, as in φW39 ΔN74 Ref, produced only 4% ldsDNA product. It is important to note that both the φW39 ΔN66 and ΔN74 are capable of producing substantial levels of nicked DNA but are limited in their ability to produce ldsDNA.

In addition to investigating the efficiency of these variants to create targeted DSBs, we

were interested in the capacity of the RecA/Ref system for genome editing, specifically gene deletions. A modification of the above assay was applied in order to demonstrate that more than a single targeted cut could be created in one reaction. Either two oligos were introduced whose complementary regions were 500 or 1000 bases apart, or one oligo was introduced whose complementary regions were 500 or 100 bases apart and were connected by an arbitrary 12 base linker (Figure 4A). The representative gel shows reactions taken at 0 and 3 hours for a targeted nuclease assay performed with wt φW39 Ref with "no Ref" and "no RecA" controls (Figure 4B). The initial conditions before addition of Ref are illustrated in Figure 4A. The success of reactions a and b (separate oligos), as well as b and d (linked oligos) for creating targeted double strand breaks are shown in Figure 4B by the appearance of band sizes 6.75 and 6.25 kb at 3 hours. Band sizes 0.50 and 1 kb are not seen on the gel due to their small size. This demonstrates that Ref is capable of creating two separate DSBs *in vitro* with either two separate oligos or one linked oligo, although the reaction including the two separate oligos is slightly more efficient (quantification not shown).

The RecA/Ref system can be used for *in vivo* genome editing – From our interest in comparing the RecA/Ref to the CRISPR/Cas9 system, the procedure used here was adapted from experiments utilizing CRISPR/Cas9 (6). Using this adapted method and an oligo that similarly confers streptomycin resistance to the cell if properly incorporated, it was possible to measure the efficiency of mutation based on the fraction of streptomycin resistant cells per total colony forming units (cfu). An illustration of this experimental set-up is shown in Figure 5A. The complementary oligo is used to target the sequence of interest for creation of a DBS within the RecA-formed D-loop. If the DBS is repaired using the oligo which has a single base change, the point mutation that gives K42T (Str^R) will be incorporated. The experiment was initially

performed to determine what oligo length optimized efficiency. Figure 5C demonstrates, based on normalized efficiency, that oligo length similar to that used *in vitro* is also best *in vivo* (150mer compared to 140mer) (Normalized efficiency is defined as fraction of streptomycin resistant cells per total cfu for RecA/Ref cells divided by same for RecA only cells).

Using the optimized oligo length, further experiments were done to determine the efficiency of this system *in vivo*. Figure 5B shows that when RecA, Ref and the targeting oligo are present, there is a 10-fold increase in streptomycin resistant cells (P value <0.05). Incorporation of the desired mutation from the oligo was confirmed by DNA sequencing. This indicates that the targeting oligo used to create the DSB can also be used as the editing oligo that incorporates a specific nucleotide change. While this data is preliminary, it establishes that the RecA/Ref system is capable of introducing a desired mutation into a specific site within the genome of *E. coli* MG1655.

DISCUSSION

From our investigation of the DNA binding and nuclease activity of Ref N-terminal truncation variants, several important conclusions can be made. First, the N-terminal region is required for DNA binding and is necessary for appreciable dsDNA cleavage by Ref. Second, DSB formation by Ref in a RecA-created D-loop is optimized when some, but not all, of the N-terminal domain is removed. The most efficient Ref variants lacked either 21 or 47 N-terminal amino acids, whereas additional removal of N-terminal amino acids (φW39 ΔN59 Ref) reverted to wild-type levels of activity or led to a dramatic decrease in Ref targeted nuclease function as in the case of φW39 ΔN66 and ΔN74 Ref. In the N-terminal domain, we have identified a charge motif (++..+.+ - +) that is repeated three times. If this motif correctly indicates functional importance, removal of the first motif improves Ref nuclease function, while removal of the third

motif noticeably decreases nuclease function. It is important to note that an analysis of the N-terminal domain did not reveal any known DNA binding domain. These mechanistic insights represent at least part of an optimizing strategy for this system *in vitro* and may also translate to *in vivo* application of this system. This translation has already been seen in regards to oligo length between *in vitro* and *in vivo* experiments (Figure 5C).

Ref is an HNH endonuclease, which is similar to the YIG-GIY homing endonucleases such as I-BmoI (10). I-BmoI, a two-domain protein with separate DNA binding and nuclease functions, functions mechanistically as a monomer to create DSBs through a similar two-step nicking mechanism (10). The DNA-binding domain functions as a molecular tether attached to the globular domain by a flexible linker region, allowing for rotation of the nuclease domain to first nick one strand, rearrange, and nick the other strand, creating the DSB. In our model of Ref, the N-terminal DNA-binding domain similarly tethers Ref to the DNA (likely DNA exposed in the filament groove of RecA or at filament ends) allowing the nuclease domain to nick the paired strand of the D-loop. A difference in the mechanism between I-BmoI and Ref is that Ref requires RecA-ATP hydrolysis in order to undergo the rearrangement that allows for completion of the second nick. TMO hybrid, φW39 ΔN21 and ΔN47 all exhibit decreased DNA binding affinities and enhanced nuclease activity compared to wt Ref. This indicates that a balance in DNA binding and targeted nuclease activity is at an optimum and competition for free DNA between Ref and RecA is reduced. More specifically, when Ref concentrations are in catalytic amounts to RecA or if Ref DNA-binding is reduced slightly (as in the case of the TMO hybrid, φW39 ΔN21 and ΔN47), targeted nuclease activity is consequently enhanced. However, if DNA affinity is reduced further (as in the case of φW39 ΔN66 and ΔN74), the balance in DNA binding and nuclease activity is lost, which subsequently reduces overall nuclease activity. The enhanced

targeted nuclease activity of Ref truncation variants seems related to faster kinetics of the second cleavage event. This may suggest that a shorter N-terminal domain facilitates the necessary rearrangement when RecA-ATP hydrolysis and filament disassembly exposes the displaced strand to Ref cleavage.

Preliminary data for exploring the RecA/Ref system for *in vivo* use has confirmed that this system is able to introduce a desired mutation utilizing a directing oligo to target a specific site within the genome of *E. coli*. Many variations of this system could be implemented to bring about deletions, insertions, point mutations, and possibly other genomic changes. It has been shown *in vitro* that an oligo either missing 10 bases of homology or containing a 10 base insert still pairs with the complementary region of cdsDNA and Ref is still capable of creating a targeted DSB within the D-loop (data not shown). In addition, the results presented in Figure 4B show that Ref is capable of cutting within two D-loops to create two targeted DSBs. If positioned on either side of a gene, could produce whole gene deletions. To utilize this *in vivo* an additional repair oligo would need to be presented for the cell to be able to exclude the flanked region (the gene). The current procedure already requires electroporating one oligo, so the system may be limited if an additional oligo needs to simultaneously enter the cell. In the original experiments, the targeting oligo is also the repair oligo which reduces the number of components necessary, a potential advantage over other multicomponent editing systems.

For whole gene deletions to be tested successfully *in vivo*, a suitable knock-out selectable marker needs to be established. Attempts at deleting 10 bases in the amino-terminus region of galactokinase (*galK*) were unsuccessful due to high spontaneous mutation rates in the conversion of $galK^{+}$ *to* $galK^{-}$ cells (data not shown). This is presumably due to the inherently toxic and potentially mutagenic selection compound, 2-deoxy-D-glucose (DOG). We are instead trying to

utilize *sacB* that encodes levansucrose that polymerizes sucrose to high molecular weight fructose polymers, or levans (11). Cells expressing *sacB* are sensitive to sucrose due to the build-up of levans, and the cell is lysed. Cells that have a dysfunctional *sacB* or are *sacB*$^-$ would be able to grow on LB-agar enriched with sucrose. This would enable proper testing of the RecA/Ref system for gene deletions or gene dysfunction capabilities. The RecA/Ref experiment targeting the destruction of the *sacB* gene would produce cells capable of growing on LB-agar enriched with sucrose. The *sacB* system would allow counter selection for functionally *sacB*$^-$ cells and we expect the spontaneous mutation rate to be reduced since the selection compound, sucrose, is not inherently toxic to the cell.

As stated previously, the CRISPR/Cas9 system uses a short RNA guide that requires a PAM, which limits potential editing targets in the genome, and the potential for off-target effects has already been established (4). The Ref/RecA system opens the door to the possibility of targeting any region of the genome and the substantially longer DNA guide (100-150 bases) may reduce common off-target effects. Data gathered so far for the RecA/Ref system *in vivo* has not shown the efficiency required to target non-selectable genes. However, with increased mechanistic information and an understanding of the dynamic interaction between RecA and Ref, efficiency for *in vivo* use has a great deal of room for optimization. If efficiency limitations can be overcome, the RecA/Ref system has promise for becoming a useful tool in biotechnology.

FIGURES AND TABLES

Table 1: Sequence of oligonucleotides used.

rlb1 150mer	TTTTGGTTTTTATCGTCGTCTGGTAAACGAGGGTTATGATAGTGTTGCTCTTACTATGCCTCGTAATTCCTTTTGGCGTTATGTATCTGCATTAGTTGAATGTGGTATTCCTAAATCTCAACTGATGAATCTTTCTACCTGTAATAATGT
AJM25 71mer	56FAM/CAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGC
AJG52 71mer	GCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTG
AJG53 140mer	GTAGGAAGTCACTTCGAAACCGTTAGTCAGACGAACACGGCATACTTTACGCAGCGCGGAGTTCGGTTTTGTAGGAGTGGTAGTATATACACGAGTACATACGCCACGTTTTTGCGGGCATGCTTCCAGCGCAGGCACGT
AJG54 100mer	CCGTTAGTCAGACGAACACGGCATACTTTACGCAGCGCGGAGTTCGGTTTTGTAGGAGTGGTAGTATATACACGAGTACATACGCCACGTTTTTGCGGGCATGCT
TMO12 80mer	ACGAACACGGCATACTTTACGCAGCGCGGAGTTCGGTTTTGTAGGAGTGGTAGTATATACACGAGTACATACGCCACGTT
TMO13 60mer	CATACTTTACGCAGCGCGGAGTTCGGTTTTGTAGGAGTGGTAGTATATACACGAGTACAT
TMO14 40mer	GCAGCGCGGAGTTCGGTTTTGTAGGAGTGGTAGTATATAC

Figure 1

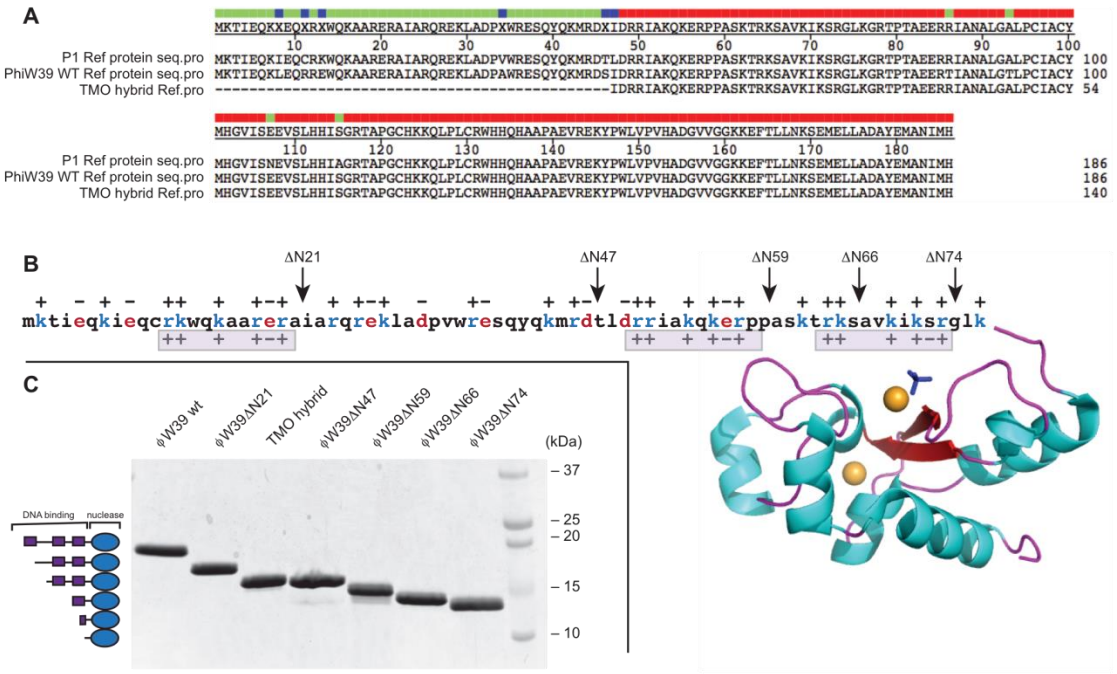
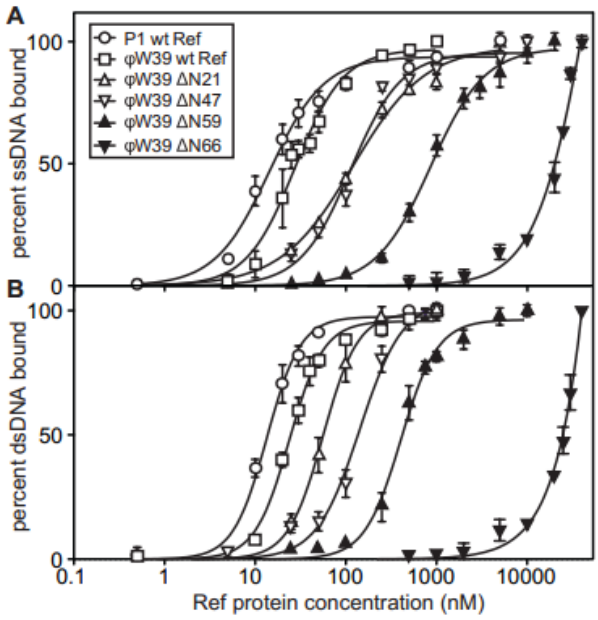


Figure 1: Sequence alignment and overview of Ref truncations based on charge distribution. A) Protein sequence alignment of P1 WT, φW39 WT, and TMO hybrid Ref proteins. There are 10 amino acid differences between P1 Ref and φW39 Ref. TMO hybrid is a truncated version with a mix of P1 Ref and φW39 Ref amino acids at heterologous positions. B) Schematic of protein sequence and charge distribution of unstructured N-terminal 76 amino acids of P1 WT Ref. C-terminal nuclease domain crystal structure is also included. Based on the signature charge distribution (++..+ .+ - +) pattern shown below the sequence, various truncations of φW39 Ref were cloned and purified. C) Schematic of Ref truncations and image of a Coomassie-stained SDS-PAGE containing ~4 µg of each construct. All proteins were soluble and purified in native form.

Figure 2



C

	K_D-ssDNA (nM)	K_D-dsDNA (nM)
wt P1 Ref	14 ± 1	14 ± 1
wt φW39 Ref	28 ± 2	24 ± 1
φW39 RefΔN21	119 ± 9	56 ± 3
φW39 RefΔN47	113 ± 7	143 ± 12
φW39 RefΔN59	856 ± 62	406 ± 22
φW39 RefΔN66	$> 5^4 \pm 3^4$	$> 8^6 \pm 4^6$
φW39 RefΔN74	no binding	no binding
P1 RefΔN76	no binding	no binding

Figure 2: Ref truncations have distinct affinities for ssDNA and dsDNA in comparison to wild type Ref. Equilibrium binding isotherms of WT Ref and truncations in presence of labeled ssDNA or dsDNA as monitored by fluorescence polarization. All data are the average of at least three experiments. Error bars are one standard deviation from the mean. **A)** φW39 ΔN21 Ref and φW39 ΔN47 Ref show reduced affinity for ssDNA in comparison to WT Ref. **B)** φW39 ΔN21 Ref and φW39 ΔN47 Ref also show a decreasing affinity for dsDNA as more of the N-terminus is removed in comparison to WT Ref. **C)** Table of apparent K_D values for each Ref protein analyzed from at least three replicates. No DNA binding was observed with the φW39 ΔN74 Ref or P1 ΔN76 Ref.

Figure 3

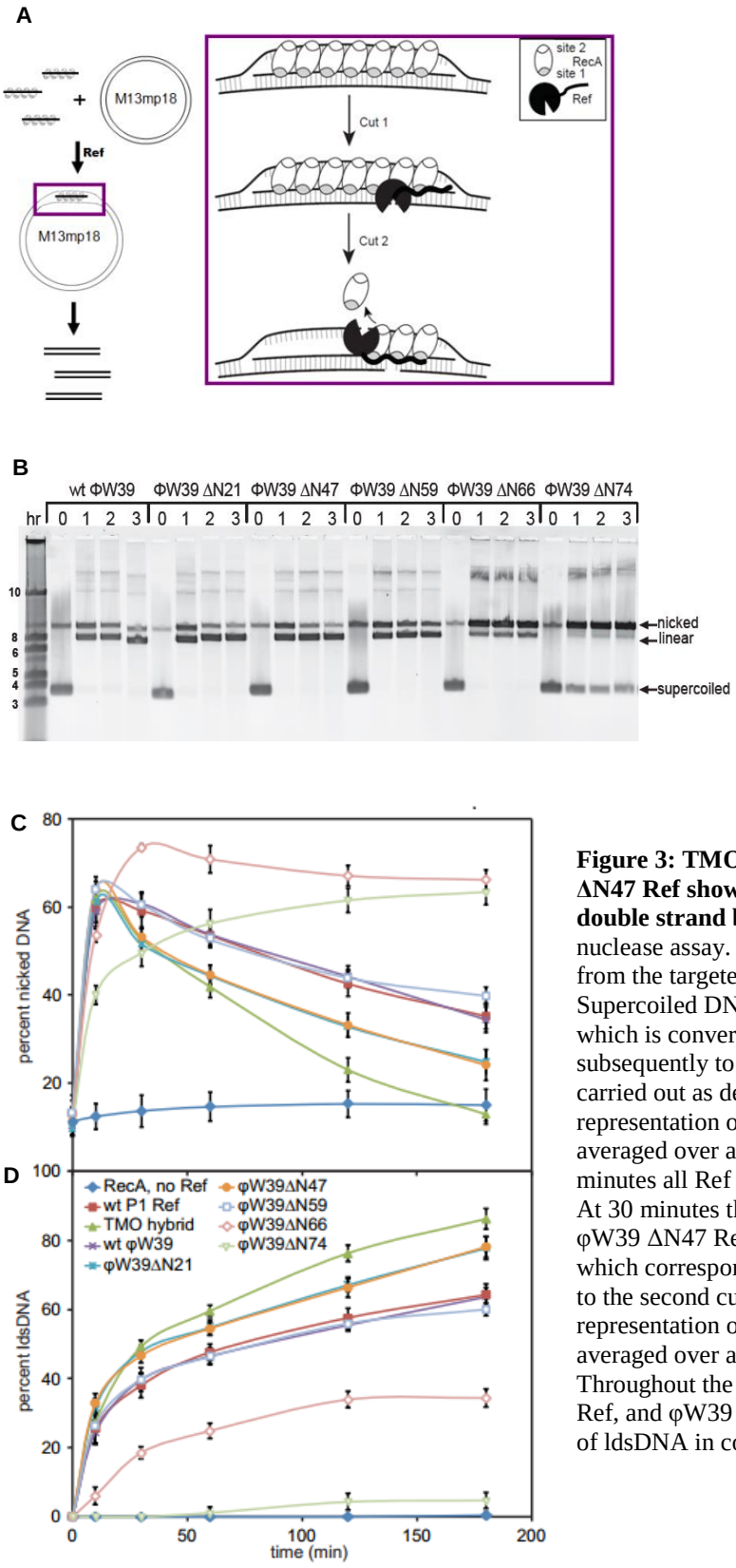


Figure 3: TMO hybrid Ref, φW39 ΔN21 Ref, and φW39 ΔN47 Ref show increased efficiency in creating targeted double strand breaks. A) Reaction schematic of targeting nuclease assay. **B)** Agarose gel representing time points from the targeted nuclease assay with φW39 Ref. Supercoiled DNA is the starting material (bottom band), which is converted first to nicked DNA (top band) and subsequently to linear DNA (middle band). Reactions were carried out as described in Methods. **C)** Graphical representation of nicked product formation over time, averaged over at least three independent experiments. At 10 minutes all Ref proteins exhibit similar amounts of nicking. At 30 minutes the TMO hybrid Ref, φW39 ΔN21 Ref, and φW39 ΔN47 Ref show a greater decrease in nicked product, which corresponds to an increase in ldsDNA formation due to the second cut on the displaced strand. **D)** Graphical representation of ldsDNA product formation over time, averaged over at least three independent experiments. Throughout the time course TMO hybrid Ref, φW39 ΔN21 Ref, and φW39 ΔN47 Ref all exhibit enhanced production of ldsDNA in comparison to WT Refs.

Figure 4

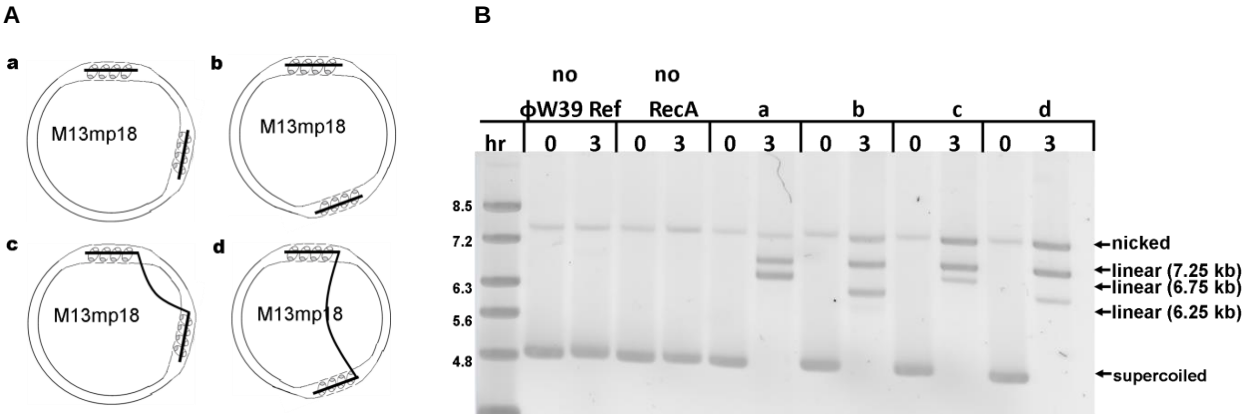
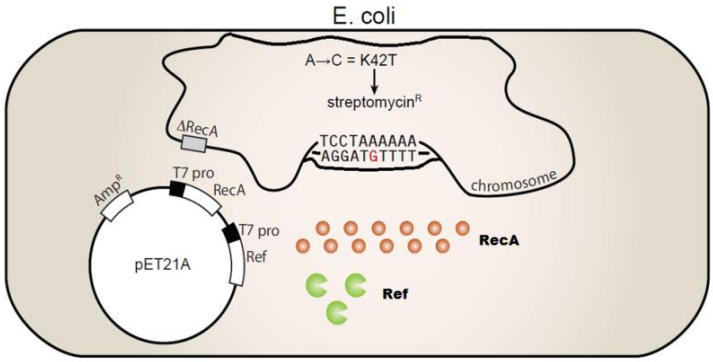


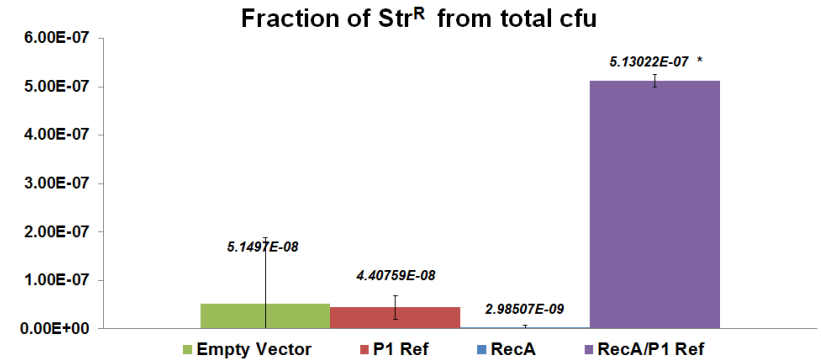
Figure 4: RecA/Ref system can create "deletions" *in vitro* by utilizing two separate D-loops. A) Illustration of each reaction condition before addition of Ref. **a** represents two separate oligos 500 bases apart, **b** represents two oligos 1000 bases apart, **c** represents one oligo whose homologous regions are 500 bases apart connected by 12 base linker, and **d** represents one oligo whose homologous regions are 1000 bases apart connected by 12 base linker. **B)** Representative agarose gel of targeted nuclease assay. Reactions a, b, c, and d corresponds with indicated setup in **A**. Note: size of M13mp18 is 7.25 kb.

Figure 5

A



B



C

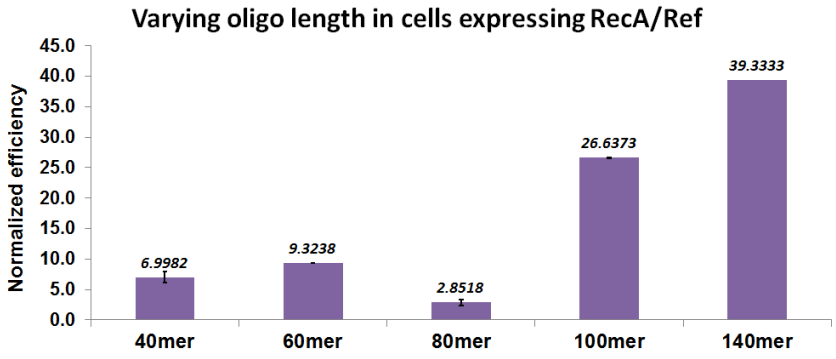


Figure 5: The RecA/Ref system can be used to edit genomes *in vivo*. A) Schematic of RecA and Ref for *in vivo* targeting for genome editing in *E. coli* MG1655. The Ref and RecA are expressed from pET21A and the editing/targeting oligo, introduced by electroporation, incorporates a point mutation (K42T) that confers streptomycin resistance to the cell. **B)** Data for *in vivo* targeting show that when Ref and RecA are present, there is a 10-fold increase in streptomycin resistance cells (asterisk indicates P value < 0.05). Experiment conducted AJG54 as targeting oligo. **C)** Normalized efficiency is optimized when the directing oligo is 140 bases long. (Normalized efficiency defined as fraction of streptomycin resistant cells per total cfu for RecA/Ref cells divided by same for RecA only cells).

AWKNOWLEDGMENTS

This work was supported by the National Institutes of Health Grant GM32335. I would like to thank Elizabeth Wood for constructing many of the strains and plasmids used in this study, and Angela Gruber for her mentorship and assistance throughout this project.

REFERENCES

1. Lu, S. D. and Gottesman, M. (1989). *Journal of Bacteriology* 171, 3427-3432.

2. Windle, B. E. and Hays, J. B. (1986). *Proceedings of the National Academy of Sciences* 83, 3885-3889.

3. Gruenig M. C. *et al.* (2011). *Journal of Biological Chemistry* 286, 8240-8251.

4. Ronayne, E. A. and Cox, M. M. (2013). *Nucleic Acid Research*. doi: 10.1093/nar/gkt1342.

5. Lebowitz, J. *et al.* (2002). *Protein Science* 11, 2067-2079.

6. Jiang, W. *et al*. (2013). *Nature biotechnology* 31(3): 233-239.

7. Messing, J. (1983). *Methods Enzymol*. 101:20-78.

8. Neuendorf, S. K., and Cox, M. M. (1986). *J. Biol. Chem.* 261:8276-8282.

9. Haruta, N. *et al*. (2003). *Journal of Biological Chemistry* 278:52710-52723.

10. Kleinstiver, B. P. *et al*. (2013). *Nucleic Acid Research* 41(10): 5413-27

11. Pelicic, V. *et al*. (1996). *J. Bacteriol.* 178(4):1197-99.