Chapter IV

CRISPR/Cas9 Recombineering-mediated targeted deep mutational scanning of essential genes in *Escherichia coli*

As discussed in chapter III, using the CREATE setup for trackable deep-mutational scan could be challenging due to the differences in editing behavior and unintended outcomes such as off-target activity, incomplete recombination, unintended mutations and repression as opposed to editing. In the current chapter, I developed an alternate technology for CRISPR-cas9 recombineering mediated genome engineering to introduce mutations from an error-prone PCR generated template on the genome at a targeted genomic using a single gRNA. This chapter is adapted from the manuscript "CRISPR/Cas9 Recombineering-mediated targeted deep mutational scanning of essential genes in *Escherichia coli"* by Choudhury *et al*, currently in prep.

4.1 Introduction

As discussed in chapter II, the CRISPR-enabled trackable genome engineering (CREATE) technology, that uses several editing cassettes for genome-wide mutagenesis libraries, could be used for expedited directed evolution to identify mutations in essential genes and global regulators that improved *E coli* fitness in diverse stresses. However, applying CREATE for a DMS proved to be challenging. Anywhere between 10-60% of randomly chosen gRNA targeting different genomic loci have been shown not to induce Cas9:gRNA induced cell death. Consequently, due to variable selection, editing efficiency can vary between 0-100% across gRNA (Garst et al. 2017; Zerbini et al. 2017). Cells with gRNA that fail to induce DSB-mediated death can grow significantly faster than cells with active gRNA undergoing DSB and edit. Consequently in chapter II we observed that in high-throughput, the non-DSB-inducing gRNA, with low editing efficiency, take over the population and reduce overall editing efficiency to only ~1-2% (Chapter II). We also observed in Chapter II that several gRNA also cause unintended mutations, off-target activity, and repression as opposed to an edit (Cui and Bikard 2016; Zerbini et al. 2017). Consequently, cells with no edits and unintended mutations can be falsely tracked as beneficial mutations. Finally, each gRNA is coupled to a different synonymous PAM mutation and synonymous mutations can lead to significant fitness effects, especially in essential genes (Agashe et al. 2013; M. J. Lajoie et al. 2013; Lind, Berg, and Andersson 2010).

We posited that in order to target a single genomic locus, we could use a single pre-screened gRNA and SPM. In this study, we discuss the CRISPR/Cas9-mediated genomic Error-Prone Editing (CREPE) technology. As opposed to other Cas9-mediated high-throughput technologies in *E. coli*, in the CREPE protocol we use a single gRNA to integrate an error-prone

PCR library of the target with the SPM on the genome (**Figure 4.1**). Recently, a similar technology, CASPER, was reported in yeast (Jakočiūnas et al. 2018). However, yeast has a significantly higher recombination efficiency than bacteria such as *E. coli*. Recombination efficiency with linear dsDNA templates is very low in *E. coli* (K. C. Murphy, Campellone, and Poteete 2000) and is poorly understood. Therefore, we varied the homology arm length, and the Cas9 recombineering system to improve recombination and our understanding of recombination using a repair template with single nucleotide changes. We successfully developed a platform that efficiently generates unbiased and diverse genomic mutant libraries with >80% editing efficiency for non-essential genes and >55% efficiency for essential genes. Additionally, while CASPER was used for directed evolution, we adapted CREPE for use as a DMS platform to study essential *E. coli* genes in their native genomic context. Using CREPE, we score the fitness of naturally accessible mutations in the RNA polymerase beta subunit that confer resistance to rifampicin.

4.2 Results and Discussions

4.2.1 Workflow in the CREPE protocol

In the CREPE workflow (**Figure 4.1**), we first screen for a gRNA centered around the target that enables over 95% editing efficiency for replacing the NGG PAM with the synonymous PAM mutation (SPM) to be used in the repair template. We also ensure that the SPM does not affect the fitness of the cells and that their growth is comparable to the wild-type *E. coli*. We amplify and clone the target region with the SPM and sufficient unmutated end-homology (which is used for recombination) into a plasmid and develop error-prone PCR libraries (**Figure 4.1**). We amplify and co-transform the donor error-prone PCR library with the gRNA in cells with active

Cas9 and lambda red recombination proteins, encoded by the pCREPE plasmid, to integrate the mutated donor onto the genome. The pCREPE plasmid encodes *cas9* expressed under the

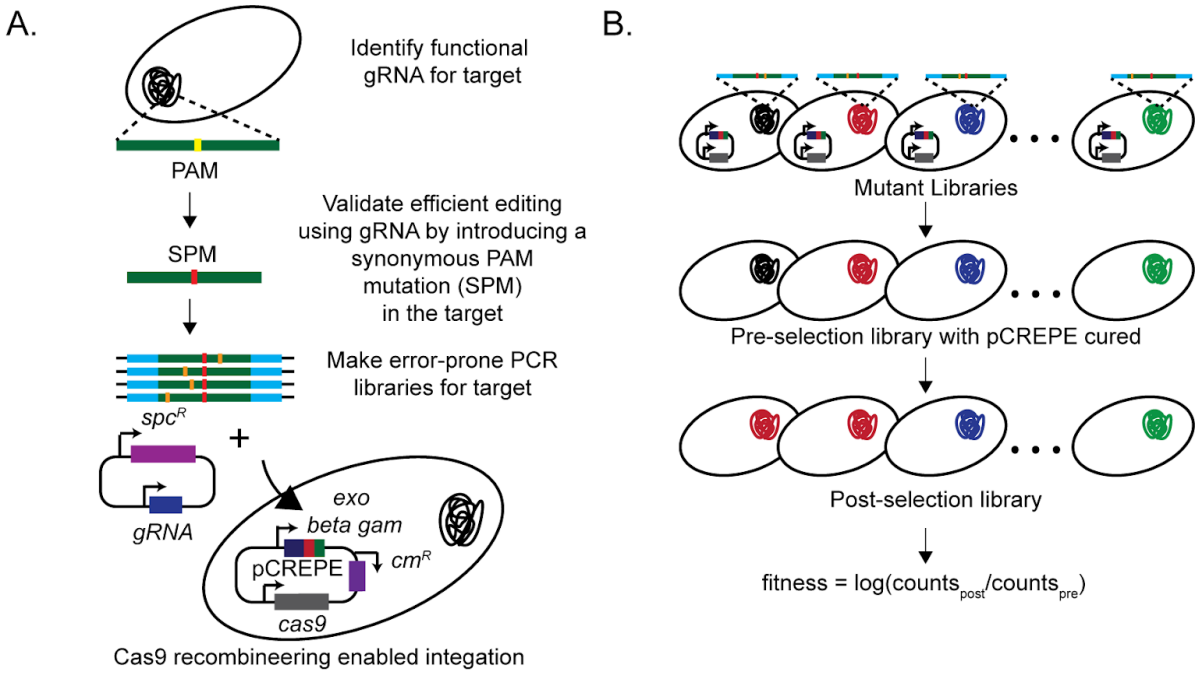

Figure 4.1: (A-B) General workflow for the CREPE protocol

constitutive Pro1 promoter (Davis, Rubin, and Sauer 2011) and the lambda Red recombination genes *exo*, *beta,* and *gam* expressed using the heat-inducible pL promoter, induced by heat shock at 42^{O}C (Yu et al. 2000). The plasmid has the temperature curable pSC101 origin of replication which replicates at 30^{O}C and is cured from the cells at 37^{O}C (Phillips 1999). The plasmid is cured prior to selection to remove the bulky plasmid that could impact cellular fitness and also to avoid fitness changes due to off-target effects of Cas9 (Hsu et al. 2013). We then sequence the target directly from the genome using deep sequencing before and after selection to quantify the frequency of mutations and estimate the distribution of fitness for the mutants in the library (**Figure 4.1**). .

4.2.2 There is a strong preference for integration of low-diversity sequences with PAM-proximal mutations

Cas9:gRNA induced DSBs increase editing efficiency primarily by selecting for edited cells with the SPM introduced by recombineering (Jiang et al. 2013; Cong et al. 2013). Initially we assumed that recombineering using a dsDNA substrate with limited mutations may follow the same mechanism proposed for dsDNA-mediated gene replacement (**Figure 4.2A**). The Lambda-Exo protein processes the dsDNA template into a single-stranded intermediate which is annealed to the Okazaki fragment using both ends by lambda-Beta protein, and the gene replacement is completed by the native replication machinery (Mosberg, Lajoie, and Church 2010) (**Figure 4.2A**).

Beta can stably anneal DNA at both ends of the ss repair intermediate with 1-2 kb long non-homologous region using only 50 bp of flanking homology (**Figure 4.2A**) (Mosberg, Lajoie, and Church 2010; Maresca et al. 2010). Therefore, we expected that using the recombination

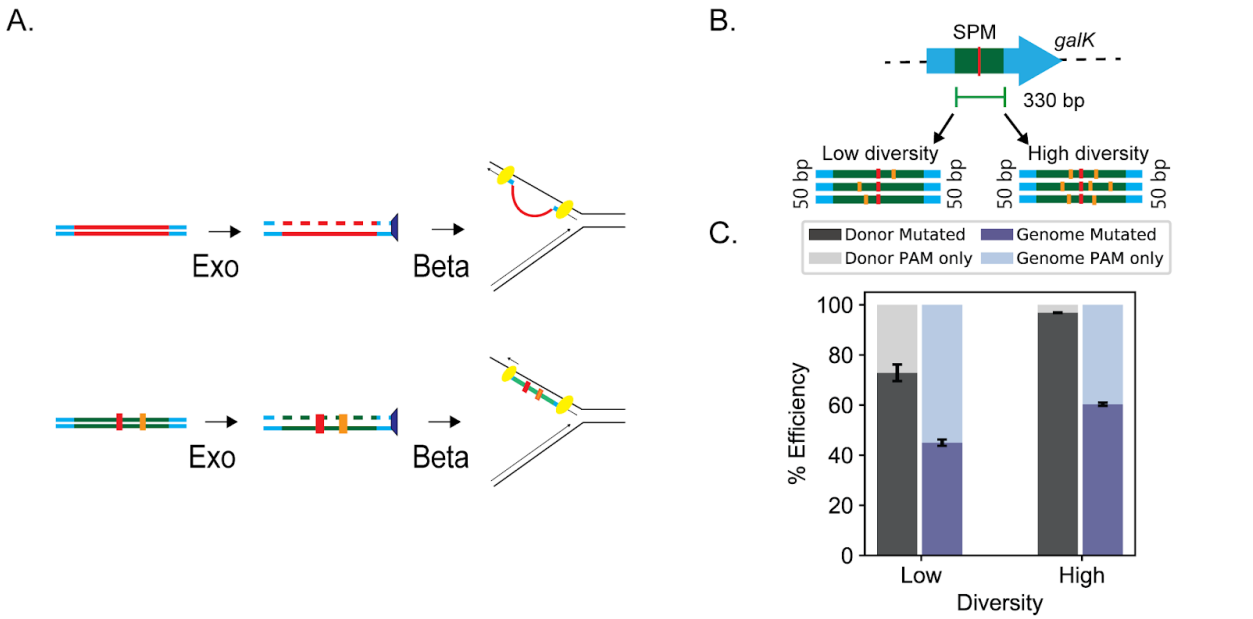

Figure 4.2: Impact of donor diversity on genome mutation efficiency: (A) Proposed mechanism for dsDNA mediated gene replacement using recombineering occurs via a single stranded intermediate (Mosberg, Lajoie, and Church 2010). We assume that lambda-mediated recombination of the CREPE substrate with limited mutations also follos this proposed mechanism. **(B)** Comparison of mutation efficiency between the high and low diversity donors before (black) and after (blue) integration on the genome. The lighter colors represent fraction of sequences with only the SPM. Each value represent mean and error bars represent standard deviation for biological replicates of deep-sequencing data.

template with single nucleotide changes, interactions between the annealed flanking homology may not have a significant impact on recombination and the efficiency of recombination would be similar regardless of the number of mutations in the donor sequence. We mutated a 330 bp region in the *galK* gene as the target using high- and low-diversity donor libraries with 66% and 92.42% sequences with one or more mutations in addition to the synonymous PAM mutation (SPM) (**Figure 4.2B**). Hereon, we refer to the percentage of sequences with mutations in addition to the SPM as the mutation efficiency. As expected, the genome mutation efficiency with the high-diversity donor was higher than the low diversity donor (**Figure 4.2C**).

In the high-diversity donor, while the sequences with 1-5 mutations were uniformly distributed, we observed a significant bias towards sequences with 1 (only SPM) and 2 (1 mutation in addition to the SPM) on the genome (**Figure 4.3A**). Biased preference for sequences

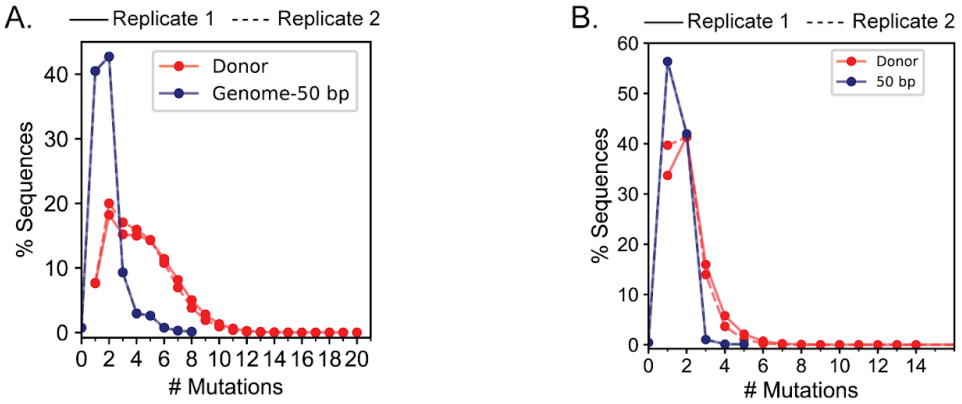

Figure 4.3: Impact of donor sequence diversity on genome mutaiton distribution (A) A comparison of percentage sequence variants categorized by the number of mutations (x-axis) between the high-diversity donor before (red) and after (blue) integration on the genome. #Mutations = 1 corresponds to sequences with only the SPM. The experiments were performed in biological replicates. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively. **(B)** A comparison of percentage sequence variants categorized by the number of mutations (x-axis) between the low-diversity donor before (red) and after (blue) integration on the genome. #Mutations = 1 corresponds to sequences with only the SPM. The experiments were performed in biological replicates. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively.

with fewer mutations was also observed with the low-diversity library (**Figure 4.3B).** Contrary to

our expectations, the efficiency of recombination decreased with an increasing number of mutations.

Additionally if the end-homology is sufficient for Beta-mediated annealing (**Figure 4.2A**), the position of mutations within the target should not impact recombination efficiency (Li et al. 2013). On the contrary, we observed a decrease in the mutation frequency per residue with increasing distance from PAM (**Figure 4.4A** and **4.4B**). Similar observations have been made before using double-stranded plasmid donors (Garst et al. 2017). With the high-diversity donor, the mutation frequency per residue exceeded the error frequency observed in unmutated regions across the entire target. However, with the low-diversity donor high mutation frequency per residue was observed predominantly within 100 base-pair sequence around the PAM (**Figure 4.4B**). The distribution of mutations was significantly biased by the distance from the PAM. Mutations closer to the PAM had higher chances of being integrated on the genome as opposed to ones further from PAM.

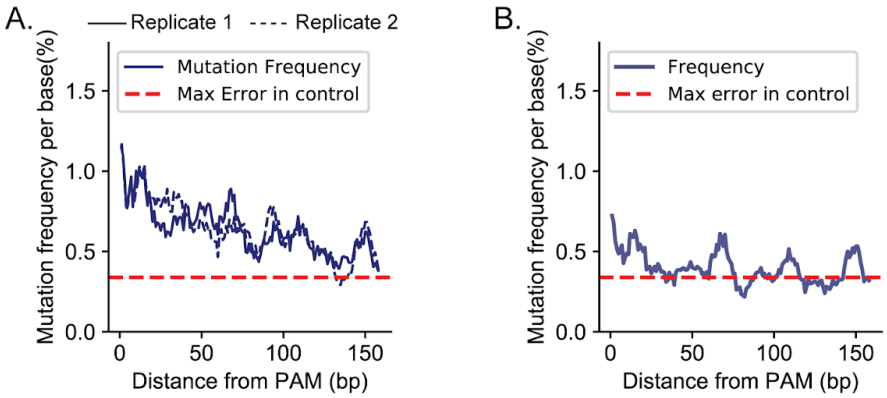

Figure 4.4: Impact of distance from PAM on distribution of mutations (A) Change in mutation frequency per base (%), percentage of sequences with a mutation at the position, using the high-diversity donor represented as rolling mean over 10 bases versus the distance from the PAM measured in base pairs. The experiments were performed in biological replicates. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively. **(B)** Change in mutation frequency per base (%), percentage of sequences with a mutation at the position, using the low-diversity donor represented as rolling mean over 10 bases versus the distance from the PAM measured in base pairs.

4.2.3 Increasing end-homology length improves mutation efficiency only for the high-diversity donor

We posited that the PAM-proximal mutation bias was likely due to better annealing of sequences with mutations closer to the PAM with longer uninterrupted end-homology than sequences with mutations further away from the PAM. Therefore, we evaluated if the PAM-proximal bias could be alleviated by increasing end-homology length. With the high-mutation rate donor, we observed a 11.4 +/- 0.2 % increase in secondary mutation efficiency by increasing the end homology from 50 bp to 250 bp (**Figure 4.5A**). To test if the PAM-proximity dependent bias reduced with an increase in end homology length, we quantified the decrease in mutation frequency with increasing distance from PAM as an exponential decay;

mutation_frequency, $\mu = \mu o * \exp(-\lambda * x)$

where μo is the maximum mutation frequency and λ is the decay rate by position, and x is the distance from PAM. For the high-diversity donor with an increase in end homology length, there was a significant increase in μo ($p<0.01$, t-test) and no significant change in λ ($p>0.05$, t-test) (**Figure 4.5B**). We observed a significant increase in the percentage of sequences with higher diversity (number of mutations in addition to the SPM> 2) on the genome (**Figure 4.5C**). Increasing the HA length did not reduce the PAM-proximal mutation bias but improved mutation efficiency by improving recombination with more divergent sequences on the genome. This was corroborated by the observation that for the low-diversity library, which lacked high diversity sequences in the donor to begin with, the editing efficiency and per-base mutation frequency did not change with the increase in the HA length (**Figure 4.5D, 4.5E and 4.5F**).

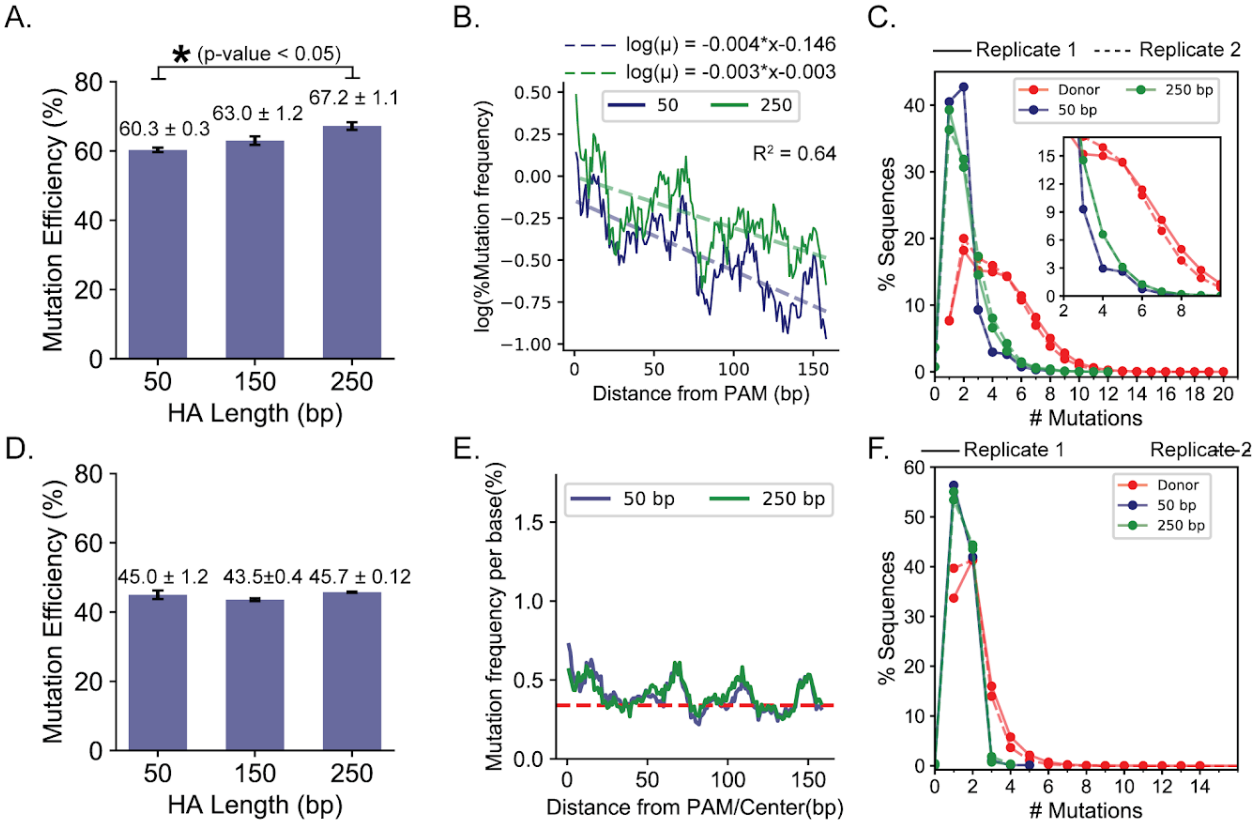

Figure 4.5: Impact of HA-length on mutation efficiency (A) Comparison of % mutation efficiency determined by deep sequencing the genome after integrating the low-diversity donor using end homology of lengths of 50 bp, 150 bp and 250 bp. Significant changes, determined as p-value <0.05 for 1-tailed student's t-test, are demonstrated using the *. Each value represent mean and error bars represent standard deviation for biological replicates. **(B)** Comparison of change in mutation frequency per base (%), percentage of sequences with a mutation at each position, using the high-diversity donor represented as rolling mean over 10 bases versus the distance from the PAM (base pairs) for recombination of the low-diversity donor using 50 bp and 250 bp end homology. **(C)** A comparison of percentage sequence variants categorized by the number of mutations (x-axis) for the low-diversity donor before (red) and after genome integration using 50 bp (green) and 250 bp (blue) long end homology. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively. **(D)** Comparison of % mutation efficiency determined by deep sequencing the genome after integrating the low-diversity donor using end homology of lengths of 50 bp, 150 bp and 250 bp. Significant changes, determined as p-value <0.05 for 1-tailed student's t-test, are demonstrated using the *. Each value represent mean and error bars represent standard deviation for biological replicates. **(E)** Comparison of change in mutation frequency per base (%), percentage of sequences with a mutation at each position, using the high-diversity donor represented as rolling mean over 10 bases versus the distance from the PAM (base pairs) for recombination of the low-diversity donor using 50 bp and 250 bp end homology. **(F)** A comparison of percentage sequence variants categorized by the number of mutations (x-axis) for the low-diversity donor before (red) and after genome integration using 50 bp (green) and 250 bp (blue) long end homology. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively.

4.2.4 Inhibiting *mutL* improved non-PAM editing efficiency by improving the integration of diverse sequences

Replication forks with beta-annealed ssDNA are usually resolved by native DNA polymerases and ligases (X.-T. Li et al. 2013; Sawitzke et al. 2011) (**Figure 4.2A**). Mismatches between the wild-type sequences and the recombination substrates are corrected by methyl-directed mismatch repair (MMR) to reduce recombination efficiency (Sawitzke et al. 2011; Costantino and Court 2003). Therefore, deleting *mutL* or *mutS* genes improves recombineering efficiency with mutagenic single-stranded oligos (Sawitzke et al. 2011; Costantino and Court 2003). However, the background mutation rates can increase significantly in *ΔmutS* and *ΔmutL* strains (Isaacs et al. 2011; Nyerges et al. 2014). Recently, the background mutation rate was significantly reduced by temporarily co-expressing a dominant negative allele of MutL, MutL-E32K, with the lambda Red recombination proteins by heat shock at 42^{O}C using the pL promoter (Nyerges et al. 2016). We cloned the *mutL-E32K* gene similarly in the pCREPE plasmid and compared the mutation efficiency between pCREPE (**Figure 4.2A**) and pCREPE+MutL-E32K plasmids (**Figure 4.6A**) using the high-diversity library with 250 bp HA. In the presence of MutL-E32K the non-PAM editing efficiency improved by 24.2+/-0.8%, and we also observed an increase in mutation frequency per position across the target (**Figure 4.6B**). Expression of MutL-E32K significantly increased the maximum mutation frequency (μ_o ,$p<0.01$, t-test) and decreased the PAM-proximal positional bias of mutations (reduction in λ, $p<0.01$, t-test) (**Figure 4.6C**). Therefore, blocking MMR improved the mutagenesis efficiency using dsDNA template with limited mutations as well.

The specificity of MMR system varies with the various mismatch substrates (G-T, A-C, A-A, G-G > T-T , T-C, A-G >> C-C) (Modrich 1991; Lahue and Modrich 1988), which result in a significant bias in the recombination efficiency for different mutations using single stranded oligos

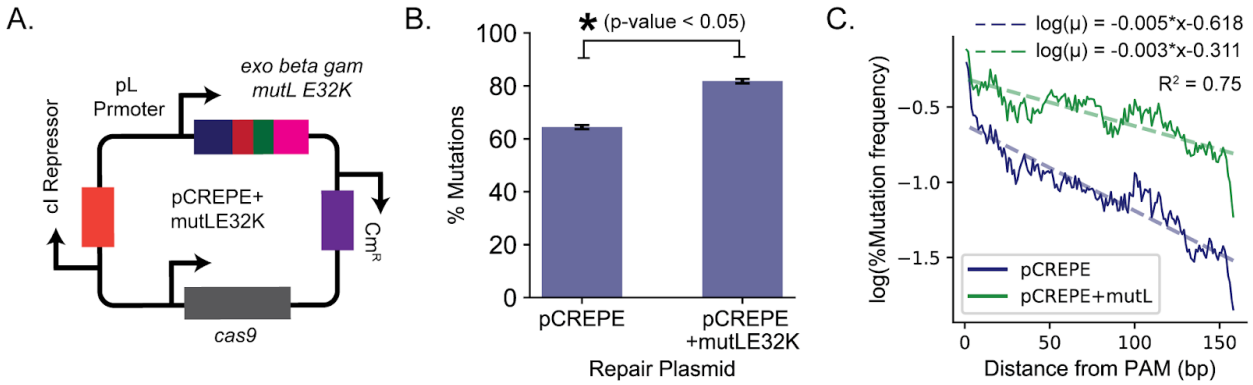

Figure 4.6: Impact of MMR on mutation efficiency (A) Plasmid map for pCREPE+mutLE32K where the mutL-E32K gene was placed under the temperature inducible pL promoter in the lambda recombination operon as described by Nyerges et al (Nyerges et al. 2016). **(B)** Comparison of % mutation efficiency determined by deep sequencing the genome after integrating high-diversity donor with the original pCREPE plasmid and the pCREPE+mutLE32K plasmid. Significant changes, determined as p-value <0.05 for 1-tailed student's t-test, are demonstrated using the *. Each value represents the mean and error bars represent the standard deviation for biological replicates. **(C)** Comparison of change in mutation frequency per base (%), percentage of sequences with a mutation at the position,using the high-diversity donor represented as rolling mean over 10 bases versus the distance from the PAM (base pairs) for recombination between pCREPE plasmid (green) and pCREPE+mutLE32K (blue). The dashed lines represent an exponential decay model fitted to quantify the decrease in mutation frequency with distance from PAM. The equations on top represent the fitted equation, the R-squared value is the lower R-squared value amongst the 2 fits.

(Costantino and Court 2003). These biases which are also observed with Cas9:gRNA DSB mediated recombineering in bacteria can be reduced by inhibiting MMR (Y. Li et al. 2015). Since inhibiting MMR significantly improved mutation efficiency using CREPE, we compared the frequencies for 12 possible base changes with and without MutL-E32K to evaluate if such biases were eliminated. To our surprise, the frequency for different mutations was comparable in the presence and absence of MutLE32K (**Figure 4.7A**). Moreover, the bias in base changes observed on the genome was comparable to the bias observed in the donor sequence (**Figure 4.7A**). This suggested that inhibiting MutL followed an alternate mechanism to improve mutation efficiency.

Using pCREPE+MutL-E32K, we observed a significant increase in the occurrence of sequences with high-diversity on the genome (**Figure 4.7B**). Therefore, the presence of MutL-E32K likely improved the mutation efficiency by improving the recombination of sequences with higher

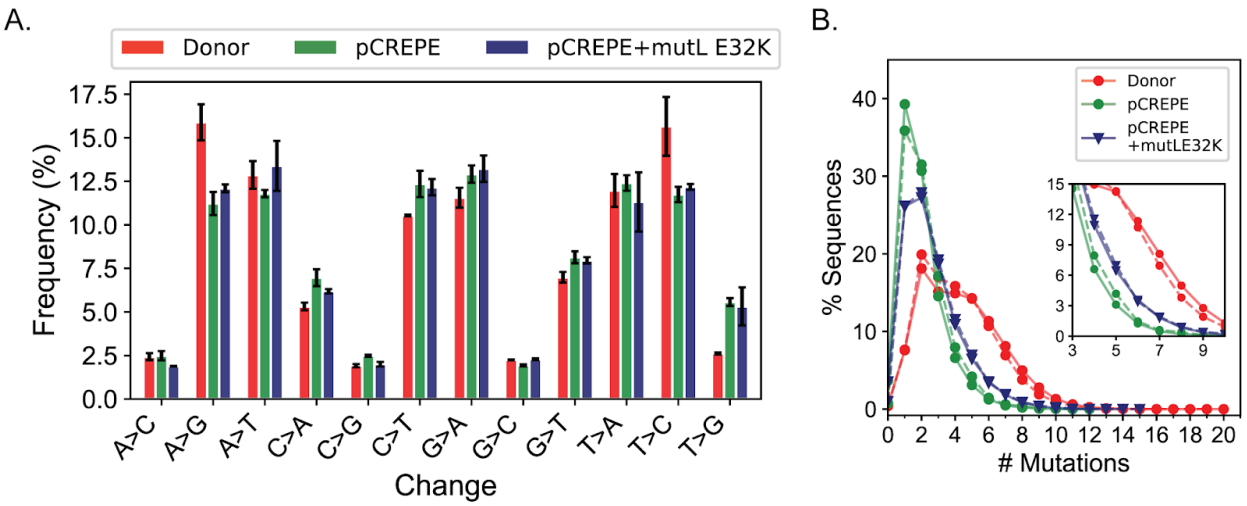

Figure 4.7: Impact of MMR on incorportation of specific mutants (A) Comparison of percentage occurrence of individual base change combinations between the high-diversity donor (red) and the genome after recombination using pCREPE (green) and pCREPE+mutLE32K (blue). **(B)** A comparison of percentage sequence variants categorized by the number of mutations (x-axis) between the high-diversity donor before (red) and after genome integration using pCREPE (green) and pCREPE+mutLE32K (blue). The inset highlights percent of sequences with 3-9 mutations in addition to PAM. #Mutations = 1 corresponds to sequences with only the SPM. The experiments were performed in biological replicates. The trends for the replicates are represented by solid (-) and dashed (--) lines respectively.

diversity.

4.2.5 Cas9-mediated recombineering using template with limited mutations may follow an alternate dominant recombination mechanism

An alternate template switch model proposed in a recent review by Murphy (Kenan C. Murphy 2016) may explain these observations for recombineering using a template with limited mutations (**Figure 4.7**). According to the template switch model, the single stranded intermediate created by Exo is initially annealed by Beta to the lagging strand. Eventually, Beta captures and switches the template for the elongating leading strand and the redirected polymerase use the

incoming strand as a template to finish the replication. However, after replication is completed, it leads to a broken chromosome. This broken chromosome is then resolved using strand invasion

A.

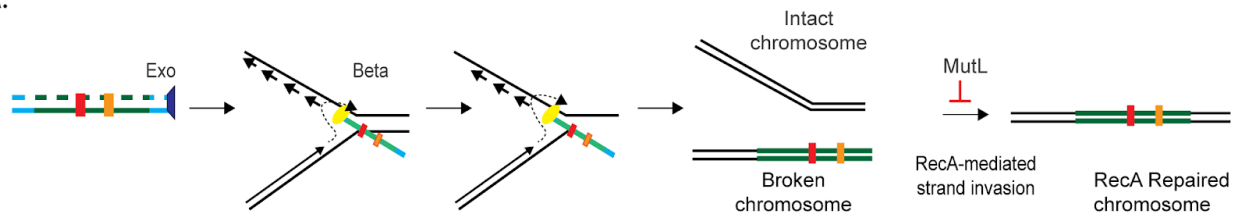

Figure 4.8: Alternate model for recombineering using template with point mutations (A) A RecA dependent alternate template switch model adapted from the review by Murphy (Kenan C. Murphy 2016).

with native RecA recombination. RecA-mediated recombination can form crossover products between homologous sequences over as low as 8 base-pair long sequences (Hsieh, Camerini-Otero, and Camerini-Otero 1992). So, as the distance between the SPM and targeted mutation increases, the chances of individual mutations getting incorporated independently increases. Since the SPM prevents Cas9:gRNA-induced cell death, the bias is controlled by PAM proximity (**Figure 4.4**). If the polymerases replicate using the recombination donor as template, the polymerase will not recognize the base changes as mismatches, which explains why the mutation specific biases due to MMR are not observed after recombination. MutS and MutL block branch migration during RecA mediated strand invasion (Worth et al. 1994; Tham et al. 2013) and the inhibition is stronger for sequences with higher diversity. This explains the low efficiency of recombination with more diverse sequences (**Figure 4.3**). RecA mediated recombination is more sensitive to HA lengths over 50 bp than lambda Red mediated recombination, which explains the improvement in recombination of divergent sequences with HA length (**Figure 4.5**). Finally, the presence of MutL-E32K blocks inhibition of strand invasion by MutL, which is stronger for more diverse sequences, to further improve recombination with diverse sequences (**Figure 4.6**). We

need to further investigate this model, which may help us identify additional targets and strategies to further improve CREPE.

4.2.6 The CREPE technology allows efficient mutagenesis of regulatory and essential genes in *E. coli*

We next measured the mutation efficiency using the optimized CREPE strategy for targeting essential genes and genes involved in global regulation. We targeted 270 bp long regions in the genes: *crp*: a global metabolism regulator, *rpoB*: the beta subunit of RNA polymerase, and *mreB*:

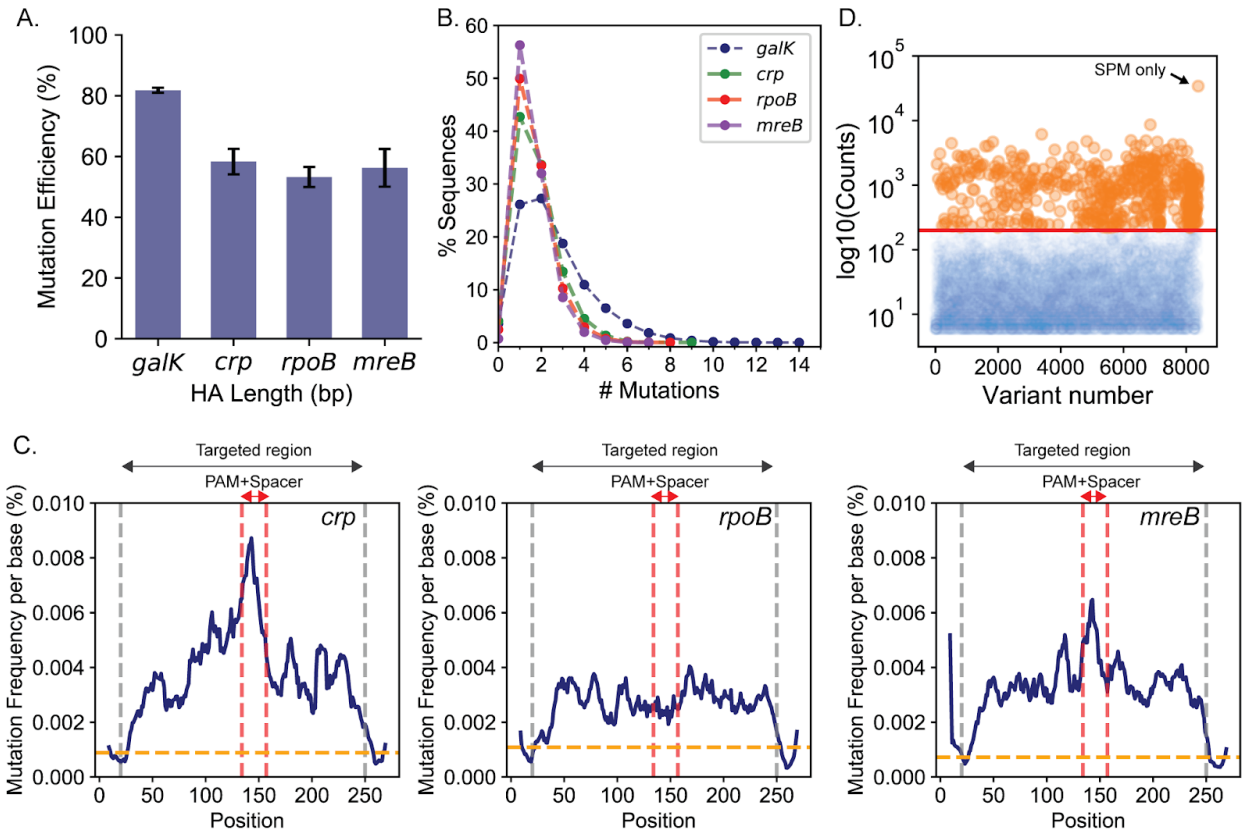

Figure 4.9: CREPE-mediated targeting of different genomic loci (A) Comparison of % mutation efficiency determined by deep sequencing by integrating the high-diversity donor using CREPE for the non-essential galK and the global regulator *crp*, and essential genes *rpoB* (RNA polymerase beta subunit), and *mreB* (cyto-skeletal protein). **(B)** A comparison of percentage sequence variants categorized by the number of mutations (x-axis) between the high-diversity donor between the non-essential *galK* (blue) and the global regulator *crp* (green), and essential genes *rpoB* (RNA polymerase beta subunit) (red), and *mreB* (cytoskeletal protein) (blue). **(C)** The change in mutation frequency per base (%), percentage of sequences with a mutation at the position,using the high-diversity donor represented as rolling mean over 10 bases along the length of the targeted region (within grey vertical lines, excluding the SPM) for each targeted locus *crp* (left), *rpoB* (middle) and *mreB* (right). The red vertical dashed line highlight the PAM+spacer region within the target. **(D)** The log10(counts) (y-axis) associated with a particular variant (variant number x-axis) chosen at random from the rpoB library. The counts for the SPM only highlighted by the arrow.

an essential cytoskeletal protein (**Figure 4.9A**). We observed between 50-60% editing efficiency for the essential genes on the genome (**Figure 4A**). The mutation efficiency of these genes was lower than *galK* (**Figure 4.9A**), probably because of the loss of function mutations in these genes would be lethal. This was corroborated by the occurrence of fewer variants with a high number of mutations per variant, as a higher number of mutations is more likely to lead to loss of function (**Figure 4.9B**). In each of the genes, we observed a significant increase in mutation frequency per base across the targeted region (**Figure 4.9C**). We also observed an unbiased distribution of variants with mutations in addition to the SPM (**Figure 4.9D**). Therefore, CREPE allowed successful mutagenesis of several essential genes in *E. coli*.

4.2.7 CREPE allowed fitness measurement of naturally accessible mutations for resistance against rifampicin

The high editing efficiency and diversity with CREPE can enable precise fitness estimates for the library. In order to demonstrate high-throughput fitness scoring, we studied the targeted *rpoB* window for resistance against rifampicin. rifampicin is an essential drug used for short-course chemotherapy against tuberculosis (Conde and Lapa E Silva 2011), which is currently the most prevalent and fatal infectious disease ("Global Tuberculosis Report - World Health Organization," n.d.). rifampicin acts by binding to the catalytic core of RpoB to inhibit the elongation step in transcription. Resistance to rifampicin arises due to mutations predominantly in three regions of the *rpoB* gene termed as rifampicin resistance-determining regions, RRDRs I (residues 507-533), II (residues 563-572) and III (residue 687) (**Figure 4.10A**) (Sandgren et al. 2009). The region we targeted covered a 90 amino acid region in RpoB which included half of

RRDR I and the entire RRDR II (**Figure 4.10A**). We designed the recombination template such that

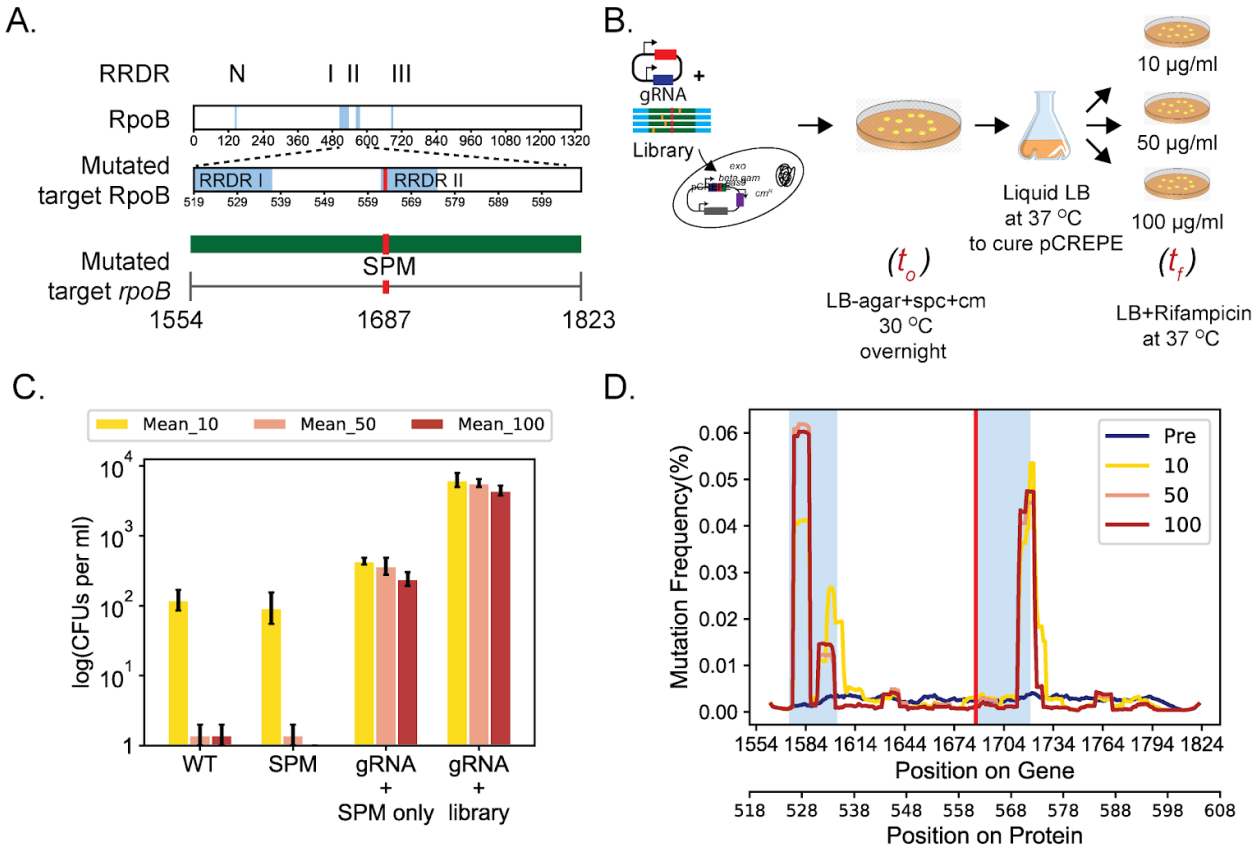

Figure 4.10: CREPE-mediated mutagenesis of *rpoB* for resistance to Rifampicin (A) Four distinct regions within rpoB (N,I,II and III) are rifampicin resistance determining regions (RRDRs). We used to CREPE to make a library covering a 270 bp long region in *rpoB* covering half of RRDR I and entire RRDR II. **(B)** Experimental setup for studying resistance to 3 different concentrations of Rifampicin (10 µg/mL (yellow) , 50 µg/mL (pink), and 100 µg/mL(red)) (Methods). **(C)** (Left to right) Cell growth determined as log(number of colony forming units or CFUs per mL of culture) after growth on LB Agar with different concentrations of Rifampicin, 10 µg/mL (yellow) , 50 µg/mL (pink), and 100 µg/mL(red), for wild type E coli MG1655, single colony of Ecoli MG1655 + synonymous PAM mutation (SPM), a library of colonies recovered by scraping colonies developed after transforming cells with the *rpoB* targeting gRNA and only the SPM mutation, and with the rpoB targeting gRNA and only the *rpoB* error-prone PCR library (see Methods for details). **(D)** The change in mutation frequency per base (%), percentage of sequences with a mutation at the position, represented as rolling mean over 10 bases along the length of the targeted region for the cells with the *rpoB* genomic error-prone PCR library at different concentrations of Rifampicin, 10 µg/mL (yellow) , 50 µg/mL (pink), and 100 µg/mL(red).

RRDR II was PAM-proximal and RRDR I was PAM-distal, to validate fitness estimates for mutations along the entire target length using CREPE (**Figure 4.10A**).

We compared the change in frequency of mutants in the library right after construction (t_o) and after growth on 3 different concentrations of rifampicin, 10 μg/ml, 50 μg/ml and 100 μg/ml in biological triplicates (t_f) (**Figure 4.10B**). At each of these concentrations, the SPM used in the template for the library did not alter fitness, measured as colony forming units (CFUs) per ml of cell culture, as compared to wild-type cells (**Figure 4.10B**). The number of resistant CFUs was two orders of magnitude higher for the library at 10 μg/ml and 4 orders of magnitude higher at 50 μg/ml and 100 μg/ml (**Figure 4.10B**). A significant increase in mutation frequencies was observed within both targeted RRDRs demonstrating successful and the distribution of mutations was unbiased (**Figure 4.10C**). Interestingly, we also observed a significant increase in the number of resistant CFUs when we plated cells in which the gRNA was co-transformed with a template having only the SPM on different concentrations of rifampicin (**Figure 4.10B**). While over 100 single and double resistant mutants were identified with the *rpoB* library, only 4 mutations were identified with the SPM only template primarily located in the PAM proximal RRDR (**Appendix Figure 2**). The resistance could be an outcome of unintended mutations introduced in the template DNA during PCR, or by error-prone polymerases expressed during the SOS response to DSBs at the target (Mallik et al. 2015). Regardless, we observed greater than an order of magnitude more resistant mutants in our library compared to the cells transformed with the SPM only template.

We next calculated the fitness of each mutant as the log-fold change in the frequency between t_o and t_f normalized to the change in frequency of a wild-type control to identify the mutations that conferred resistance (**Figure 4.11A** and **methods**). The fitness values for the mutations common across all replicates correlated strongly (**Appendix Figure 3**) (Pearson's correlation coefficient between 0.81-0.98). We averaged the fitness scores for common mutations across replicates using the Fisher scoring iterations based maximum likelihood estimates (**methods**) (Rubin et al. 2017). We observed a bimodal distribution of fitness

effects for all mutations at each rifampicin concentration (**Figure 4.11A**). A bimodal distribution of fitness is expected for antibiotic selections because a variant is either resistant or sensitive to rifampicin. We used the

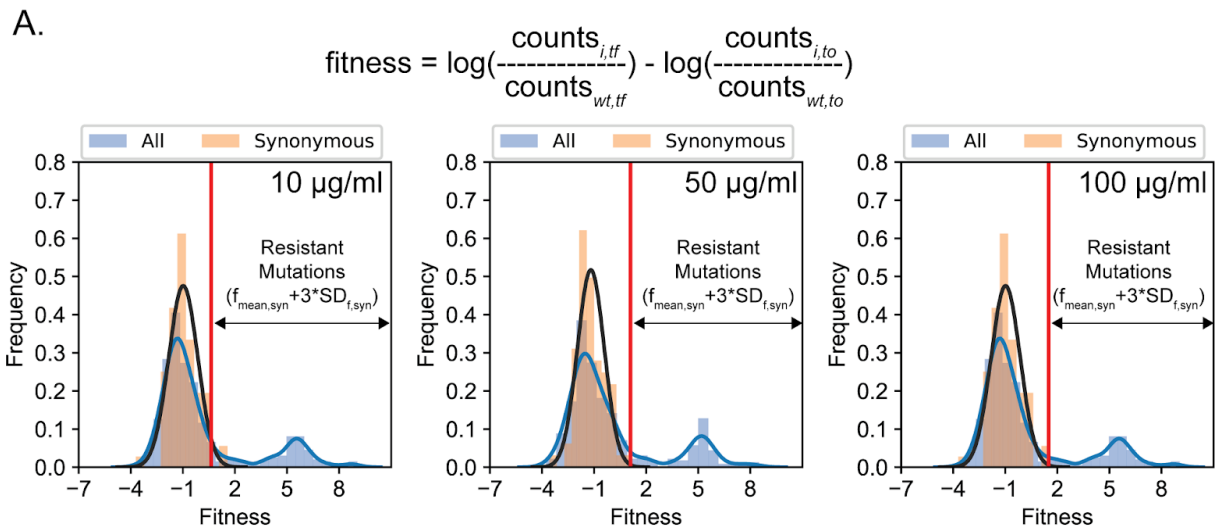

Figure 4.11: Fitness estimated for resistance to Rifampicin (A) Distribution of fitness estimates for all mutations in the rpoB library (blue) and only synonymous mutations in the rpoB library at different concentrations of Rifampicin, 10 µg/mL (left) , 50 µg/mL (center), and 100 µg/mL (right).

synonymous mutations in the population as a control to differentiate between the sensitive and resistant mutations (**Figure 4.11A**). Resistance to rifampicin is primarily caused by mutations that inhibit the interaction of the binding pocket to rifampicin. Therefore, synonymous mutations are unlikely to be resistant. We obtained a single binomial distribution of fitness with a mean negative fitness for synonymous mutations. We defined fitness greater than three standard deviations of the mean fitness of synonymous mutations as resistant (**Figure 4.11A**).

4.2.8 Beneficial mutations corroborate epidemiological findings and biochemical properties

We first studied the fitness effects of single non-synonymous mutations in the population. We estimated fitness score for 355 single non-synonymous mutations across replicates covering 88 of the 90 targeted amino acid residues with mean 3.7 mutations per residue. We found 17 resistant

mutations each for 100 µg/ml and 50 µg/ml, and 39 mutations resistant at 10 µg/ml rifampicin (**Figure 4.12A** and

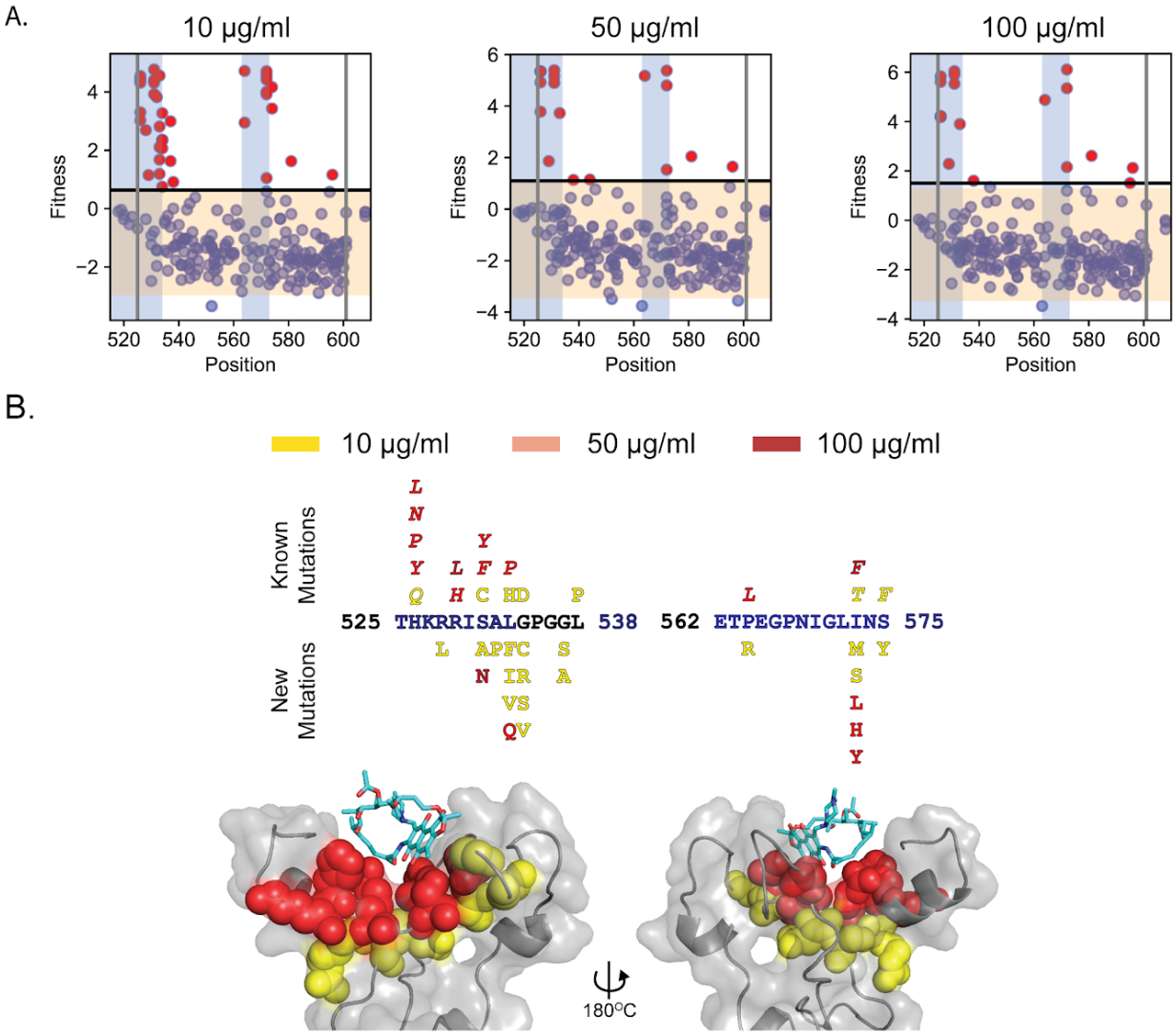

Figure 4.12: Mutations identified to confer resistance to Rifampicin (A) Fitness for mutations (blue) at each position in the RpoB target at 10 µg/mL (left) , 50 µg/mL (center), and 100 µg/mL (right). The grey lines represent the region within which mutagenesis was performed. The blue windows represent the RRDRs. The orange window represents the range of fitness values within 2.96 fitness standard deviations,plus and minus,) the mean fitness of synonymous mutations in the library. Any mutation with fitness greater >= mean fitness of synonymous mutations + 2.96*standard deviation of fitness of synonymous mutations were selected as resistant and represented using red dots. **(B)** Mutations within RRDR I (525-538) and II (562-575) identified at different Rifampicin concentrations 10 µg/mL (yellow) , 50 µg/mL (pink), and 100 µg/mL(red), with the known mutations above the target sequence and new mutations below the target sequence. The structure (cartoon+grey surface) represents Rifampicin (sticks) binding pocket mutations with spheres representing the mutations identified at 10 µg/mL (yellow) , 50 µg/mL (pink), and 100 µg/mL(red).

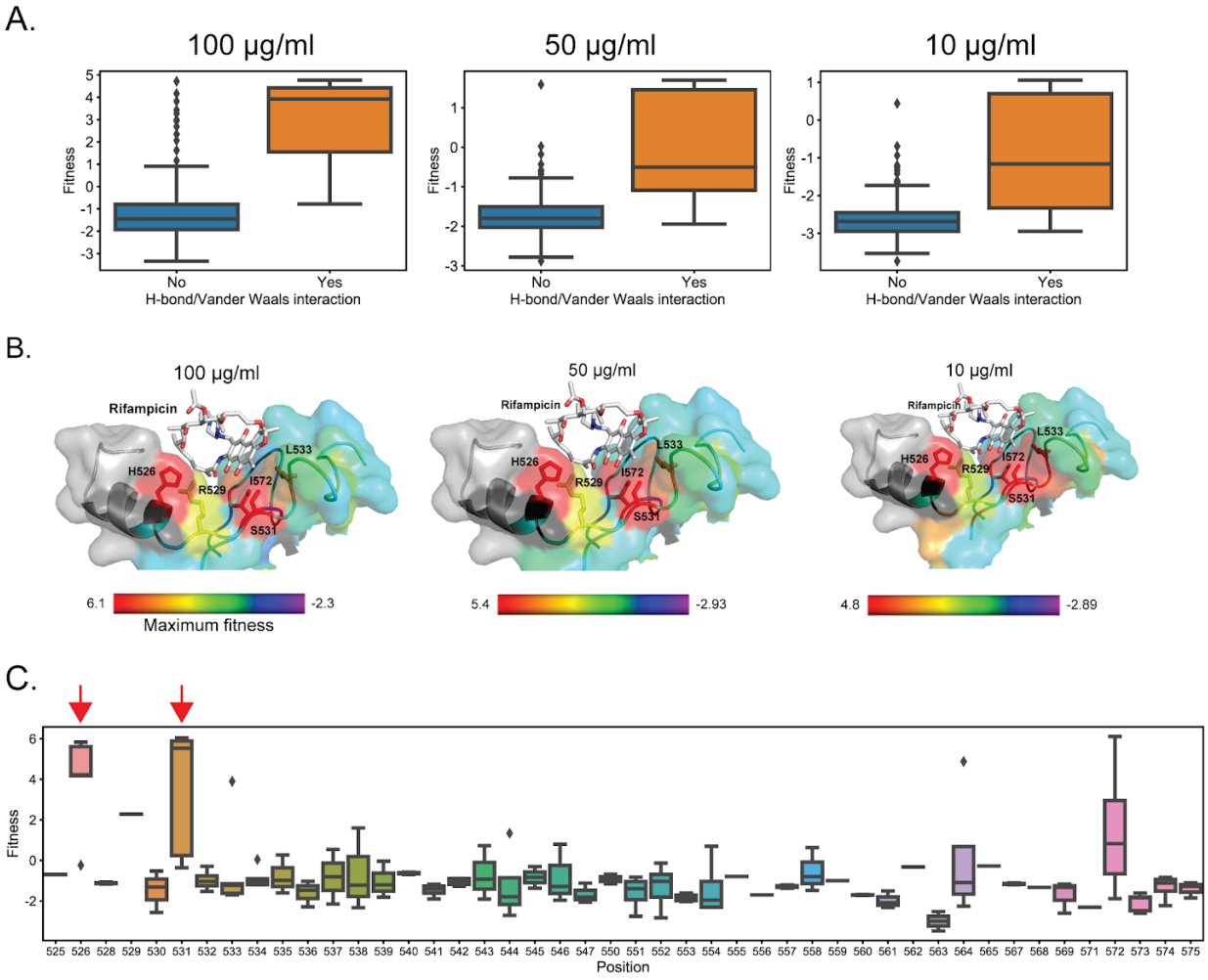

Figure 4.13: Fitness of biochemically interacting and epidemeologically relevant residues (A) Each plot compares the distribution of fitness between residues whose side-chains that make hydrogen bonds and Van der Waals interactions with Rifampicin (Orange) and residues that do not interact with Rifampicin (Blue) at 100 µg/mL (left), 50 µg/mL (middle), and 10 µg/mL (right) of Rifampicin. **(B)** The structure (cartoon + surface) represents Rifampicin (white sticks) binding pocket. The residues are colored using a heat map (range below the structure) demonstrating the maximum fitness at each residue. Residues with sticks are one that form H-bonds or Van der Waals interactions with Rifampicin (Campbell et al. 2001). The fitness scored were evaluated at 100 µg/mL Rifampicin. The fitness scores were evaluated at 10 µg/mL (left), 50 µg/mL (center) and 10 µg/mL (right) Rifampicin . **(C)** Each box-plot represents the distribution of fitness effects for mutations evaluated at each amino acid of the targeted RpoB region in RRDR I (525-538) and II (562-575), and the residues between the RRDRs at 100 µg/mL Rifampicin.

Appendix Table 5). Of the 41 resistant mutations identified at different concentration (**Appendix Table 5**), 42% were already known to confer resistance to rifampicin and 81% of the mutations occurred in sites known to confer resistance to rifampicin (Sandgren et al. 2009; Zhou et al. 2013; Campbell et al. 2001), predominantly at or adjacent to the rifampicin binding pocket (**Figure 4.12B** and **Appendix Table 5**). We identified 22 new rifampicin resistance mutations (**Figure 4.12B**). We reconstructed 9 out of the 22 new mutations and found that 8 out of 9 reconstructed mutants were resistant to rifampicin (**Appendix Figure 5**).

Amongst the 88 targeted residues, positions H526, R529, S531, L533 and I572 each of which has either H-bonding or Van-Der-Waals interaction with rifampicin (Campbell et al. 2001), had significantly higher fitness as compared to other residues within the pocket (p-value < 0.0001 student's t-test for 10, 50 and 100 μg/ml, **Figure 4.13A and 4.13B**). Substitutions at positions S531 and H526 that represent the majority of known mutations in clinical isolates (~41% and ~36% respectively) , had the highest mean fitness compared to all other positions (**Figure 4.13C)**. Therefore, fitness estimates map to known biochemical properties and clinical observations and within the target.

4.2.9 Fitness estimates help understand biochemical bases for selection of different mutations with change in rifampicin concentration

The evolution of resistance depends on treatment strategies *i.e.*, the rifampicin dosage (Wistrand-Yuen et al. 2018; Palmer and Kishony 2013). We observed that more mutations were selected at sub-inhibitory 10 μg/ml of rifampicin (**Figure 4.12A and 4.12B and Appendix table 5**) than at higher antibiotic concentrations (50 and 100 μg/ml). This is expected as the number of

accessible paths to resistance are higher at subinhibitory antibiotic concentrations *i.e.*, lower stress conditions (Lindsey et al. 2013; Wistrand-Yuen et al. 2018). Interestingly, while all mutations resistant to rifampicin at higher antibiotic concentrations (50 and 100 µg/ml) were resistant at sub-inhibitory concentrations (10 µg/ml), a second population was observed that was resistant only to 10 µg/ml of rifampicin (**Figure 4.14A**). Variants identified at 10 µg/ml of rifampicin had lower MIC values than variants identified at 100 µg/ml of rifampicin (**Appendix Figure 4** and **5**).

Rifampicin resistance occurs primarily by blocking the access of rifampicin within the binding pocket (Campbell et al. 2001). While all high-resistance conferring substitutions were in residues directly in contact with rifampicin, one subset of substitutions with resistance to lower rifampicin concentration occurred in residues next to the binding pocket, which may have lower inhibitory effects (**Figure 4.12B**). A second subgroup of the low-resistance mutants in residue S531, which H-bonds with rifampicin, substitutions to amino acids with short side chains such as S531C and S531A that would only prevent the H-bond conferred low-resistance, whereas bulkier substitutions such S531L and S531N allowed resistance to higher concentrations of rifampicin (**Figure 4.14B**). The bulkier substitutions apart, from breaking H-bond, have been demonstrated to push the adjacent fork loop to increase the solvent exposure of rifampicin and to drastically increase the resistance (Molodtsov et al. 2017). Therefore, the less bulky residues without the capability of the conformational change would be expected to have lower resistance. Similarly in Van der Waals interacting residue such as I572, less-bulky substitutions such as I572S and I572T that would only break VDW interactions had low resistance as compared to the more bulky high resistant changes I572L, I572F, I572Y and I572H that may both break VDW interactions and cause

steric hindrance (**Figure 4.14C**). This demonstrates that resistance at certain residues such as S531 and I572 resistance mechanisms are more complex than just steric inhibition, which

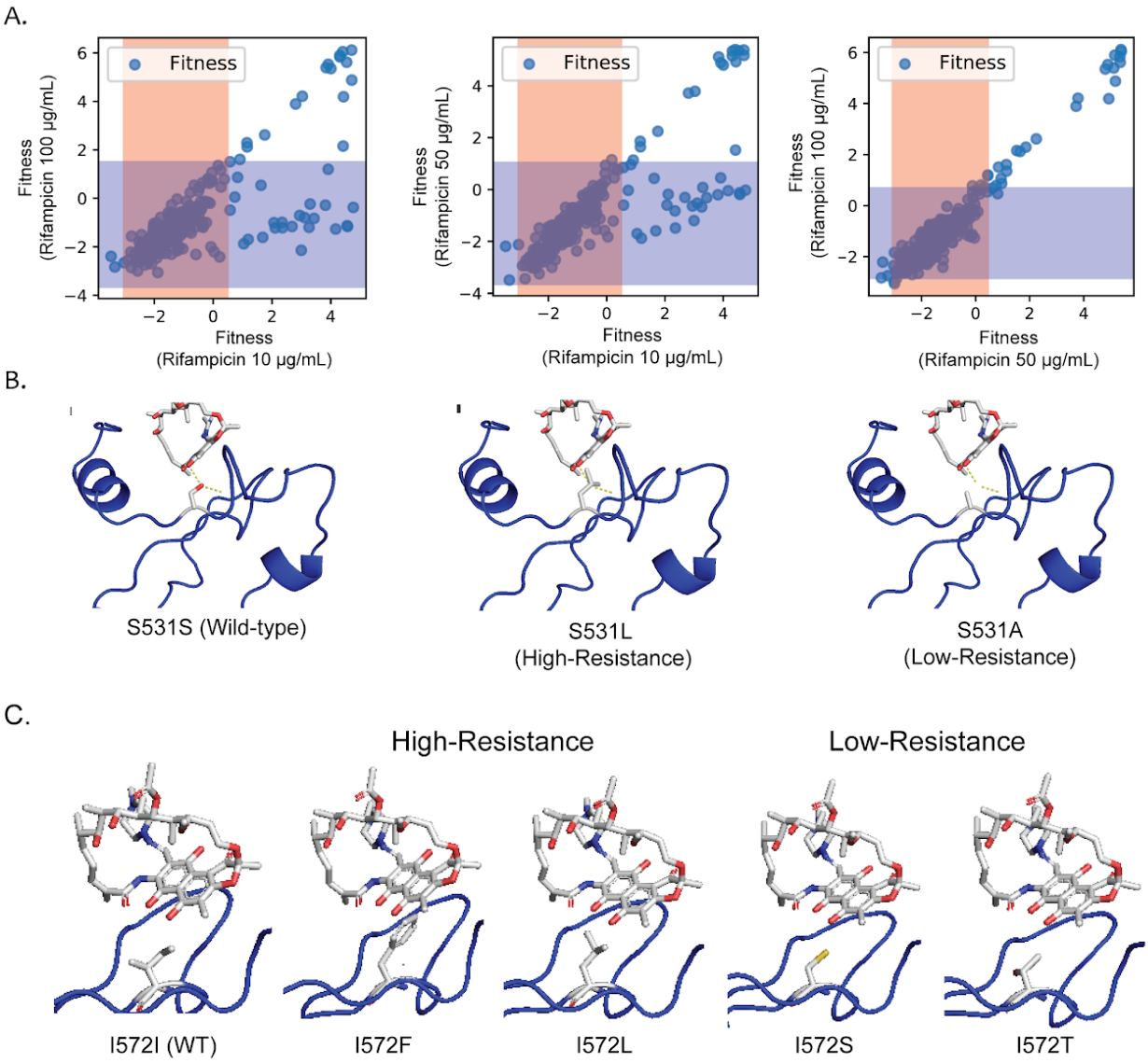

Figure 4.14: Mutational spectrum between low and high Rifampicin concentrations (A) Correlation of fitness estimated for the same mutant at 10 µg/mL Rifampicin (x-axis) and 100 µg/mL Rifampicin (y-axis) (left), 50 µg/mL Rifampicin (x-axis) and 100 µg/mL Rifampicin (y-axis) (center), and 10 µg/mL Rifampicin (x-axis) and 50 µg/mL Rifampicin (y-axis) (right). The orange and blue regions represent the window within 2.96 fitness standard deviations (on each side) around the mean fitness of synonymous mutations at 10 µg/mL and 100 µg/mL respectively. **(B)** In each structure, Rifampicin (white sticks colored by atoms) is bound to rpoB (blue catoon). The amino acid side chain in sticks represent the S531 residue in the wild-type sequence (S531S, left), and when it is mutated to high-resistance residue (S531L, middle) and low-resistance residue (S531A, right). **(C)** In each structure, Rifampicin (white sticks colored by atoms) is bound to rpoB (blue catoon). The amino acid side chain in sticks represent the I572 residue in the wild-type sequence (I572I, left), and when it is mutated to high-resistance residue (I572F and I572L, middle) and low-resistance residue (I572S and I572T, right).

corroborates a recent finding that high resistance is an outcome of several changes within the binding pocket (Molodtsov et al. 2017). Overall, the fitness estimates at different concentration sensitively capture biochemical contribution of different substitutions based on their position and type of change that allow resistance to different rifampicin concentrations. Similar sensitive fitness estimates at different drug concentrations may improve biochemical understanding of resistance.

Variants with lower MIC are preferentially selected at lower rifampicin concentration in the laboratory and in the clinic (Lindsey et al. 2013; van Ingen et al. 2011; Berrada et al. 2016). Interestingly, we observed that despite the weaker resistance and lower MIC (**Appendix Figure 4 and 5)**, the range of fitness values for the mutations resistant only at 10 μg/ml was similar to the range of fitness observed for variants selected at 50 and 100 μg/ml (**Figure 4.14A**). Mutations with high rifampicin resistance occur close to the catalytic site of RNA polymerase (Campbell et al. 2001) and consequently are detrimental to the cell (Brandis and Hughes 2018). The less bulky substitutions or substitutions away from the active site may have less detrimental effects. Therefore, we posited that, while high-resistance mutations have high fitness owing to their strong inhibition, the mutations resistant to only low-concentrations may have the equivalent high fitness despite weaker inhibition due to better growth. This hypothesis was confirmed when we found that

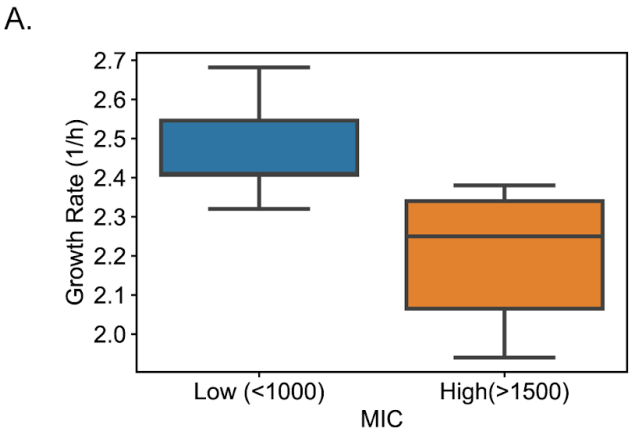

Figure 4.15: Growth rates of selected mutations at low and high Rifampicin concentrations (A) The plot compares the distribution of growth rate for mutations with lower improvement in minimum inhibitory concentration (MIC) (blue) and ones with higher minimum inhibitory concentration (MIC) (red).

the low-resistance mutations grew significantly better than the high-resistance ones in the absence of rifampicin (Student's t-test p-value < 0.01, **Figure 4.15A**). These observations explain why low-resistance mutations are preferred at lower concentrations of rifampicin, an observation made both in *E. coli* (Lindsey et al. 2013) and *M. tuberculosis* (Billington, McHugh, and Gillespie 1999; Sander et al. 2002). Low antibiotic concentrations treatments are clinically relevant for long-term TB treatment using multiple antibiotics (van Ingen et al. 2011). The fitness estimates highlight that mutations with better growth selected at lower concentration may also have better success in environment as compared to the ones with fitness costs, which explain their successful transmission in low-concentration treatments (van Ingen et al. 2011).

4.2.10 Epistasis in double mutants

The deviation of the actual fitness of a double mutant from the predicted fitness based on each of the single mutations is referred to as epistasis. It is quantified as the difference between the actual fitness of the double mutant and the sum of the fitness of the single mutants (Weinreich, Watson, and Chao 2005). Therefore, we plotted a correlation between the sum of fitness for individual mutations in the double mutant versus the actual fitness of the double mutant (since the fitness formula we use is a measure of log fitness). The fitness correlation between the actual and predicted values would lie on a 45^{O} diagonal in the case of no epistasis, below the diagonal in the case of negative epistasis (fitness of double mutant is lower than the sum of individual), and above the diagonal in the case of positive epistasis (fitness of double mutant is higher than the sum of individual). We split the mutations into 3 groups: both resistant, one resistant and one sensitive and both sensitive (**Figure 4.16**). Each group demonstrated a unique epistatic behavior (**Figure 4.16**).

A.

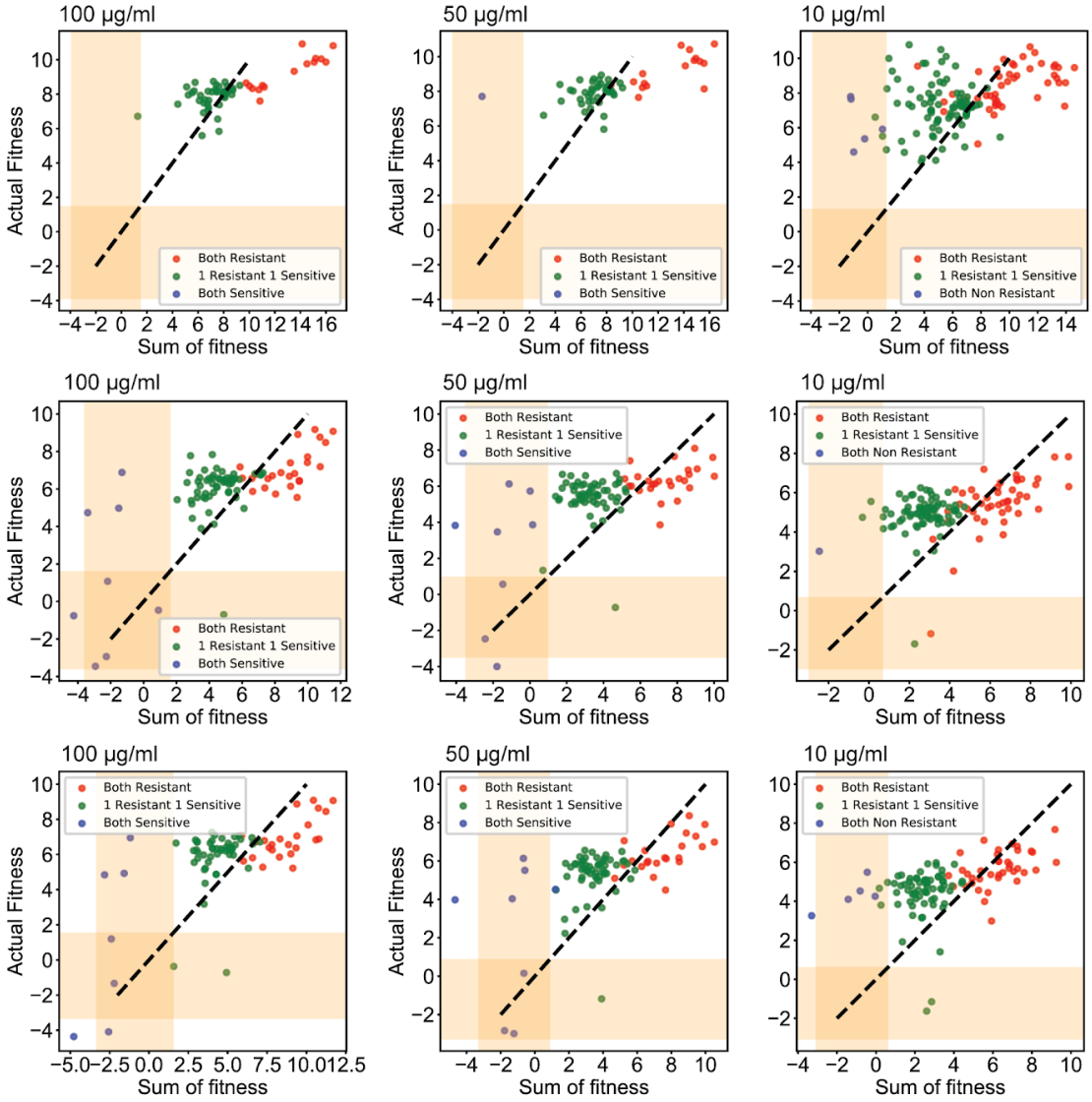

Figure 4.16: Measuring Epistasis (A) Comparison of actual fitness determined by CREPE and predicted fitness determined as sum of fitness for individual mutations in the double mutants for resistance to different concentrations of Rifampicin across the 3 biological replicates. The double mutants were categorized as combination of 2 resistant mutations (red), 1 resistant and 1 sensitive mutant (green) and 2 sensitive mutations (blue).

We observed rare occurrences of reciprocal sign epistasis, where individual mutations were not resistant to rifampicin, but resistance emerged in the combination. In most of the combinations with resistance to 100 µg/ml of rifampicin, like G534S+S574F, S574Y+L533I and G537C+I572M, individual mutations had high fitness at 10 µg/ml of rifampicin. Individually,

A.

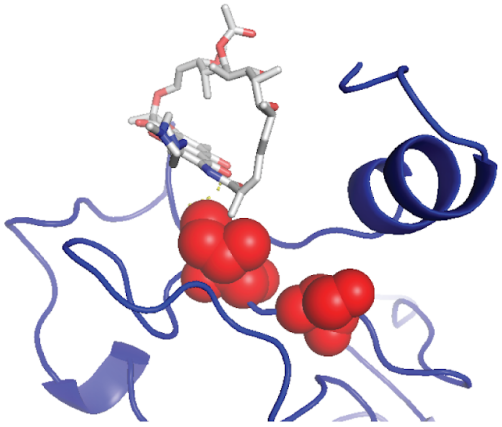

Figure 4.17: Reciprocal Epistasis: (A) Mutant combination at residue I572 and S574, that are individually non-resistant to Rifampicin at all concentrations, demonstrate reciprocal sign epistasis in combination.

reconstructed G534S+S574F, S574Y+L533I had MIC of >1500 µg/ml and 500 µg/ml respectively (**Appendix Figure 6**). In the case of I572N+S574Y, the mutations occurred at positions known to confer rifampicin resistance (**Figure 4.17A**). The mutations were bulky substitutions in residues close to each other near the rifampicin binding pocket and are therefore likely to interact to generate the observed reciprocal sign epistasis. In each case the reciprocal sign epistasis was justified. Similar to our observations with different rifampicin concentrations (**Figure 4.14**) and previous studies (Molodtsov et al. 2017), these findings again highlight that resistance to rifampicin maybe an outcome of several complex changes within the binding pocket and not just steric inhibition.

Over 70% of beneficial mutation combinations had negative (antagonistic/diminishing returns epistasis) epistasis, where the actual fitness was lower than the fitness-sum. Antagonistic epistasis between beneficial mutations is common across different systems (Tokuriki et al. 2012; Khan et al. 2011) (**Figure 4.18A**). However, higher fitness of the double mutants as compared to each individual mutation was unexpected in several cases. Single mutations such as H526L,

P564L, and I572F individually have a MIC >1500 μg/ml (**Appendix Figure 3**), 15-fold higher than

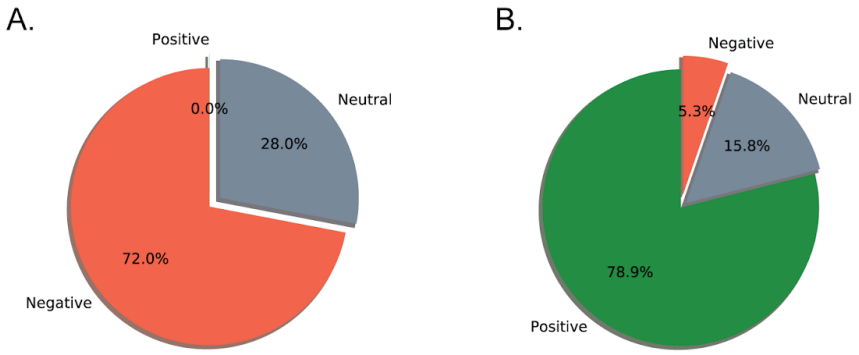

Figure 4.18: Epistasis across mutation combinations: (A) Distribution of type magnitude epistasis (positive (green), Negative (orange) and neutral (grey)) in double mutations, that are combination of individual resistant mutations. **(B)** Distribution of type magnitude epistasis (positive (green), Negative (orange) and neutral (grey)) in double mutations, that are combination of one resistant and one sensitive mutation.

the selection concentration of 100 μg/ml. Moreover, the beneficial mutations were individually detrimental to the host in the absence of rifampicin (**Figure 4.15**). This observed improvement in fitness likely occurred because the beneficial mutations compensated for each other's fitness-cost. This observation, has also been made previously for individual fitness estimates for several double resistant mutation combinations in *Pseudomonas aeruginosa* (Hall and MacLean 2011). Additionally, compensatory mutations in RpoB, which are themselves resistant to rifampicin, have also emerged in lab-evolved *M. tuberculosis*, *E. coli* and *Salmonella enterica* (Reynolds 2000; Brandis et al. 2012).

The Compensatory fitness improvement may also explain the high positive epistasis (in 78% of the combinations) between sensitive and resistant mutation combinations (**Figure 4.18B)**. Due to the fitness cost of resistant mutations, compensatory mutations are known to drive transmission of drug-resistant bacteria across populations (Casali et al. 2014; Zhang et al. 2013; Comas et al. 2011). Our data reveal there may be a high frequency of available compensatory mutations within a 90 bp region of the RpoB. It corroborates the previous observations in

laboratory evolution where compensation more often than not has occurred through mutations within the *rpoB* (Reynolds 2000; Brandis et al. 2012; Casali et al. 2012; Comas et al. 2011). Here, by revealing how large is the target for compensation within *rpoB*, we suggest that prolonged usage of rifampicin may result in the emergence of RpoB alleles with high resistance and limited cost, which would be an explosive cocktail for the propagation of resistance.

In addition to the genetic background, as highlighted above, the evolutionary consequence of a mutation is also determined by the growth environment (Weinreich, Watson, and Chao 2005; Remold and Lenski 2004). At each of the rifampicin concentrations, we quantified epistasis as $\boldsymbol{\varepsilon} = F_{AB} - F_A - F_B$. For mutations common across all concentrations, we correlated the epistasis between different concentration. Again, while the epistasis correlated strongly for higher rifampicin concentrations, we observed 2 distinct populations for correlation between low and high concentrations (**Figure 4.19A)**. While strong positive correlation was observed in **ε** for most mutation combinations, the magnitude of **ε** was higher for some mutations at the 100 µg/ml. In addition to the compensatory evolution, this environmental modulation of epistatic interactions under strong doses of rifampicin may complicate the adaptive path for evolution of higher level of

A.

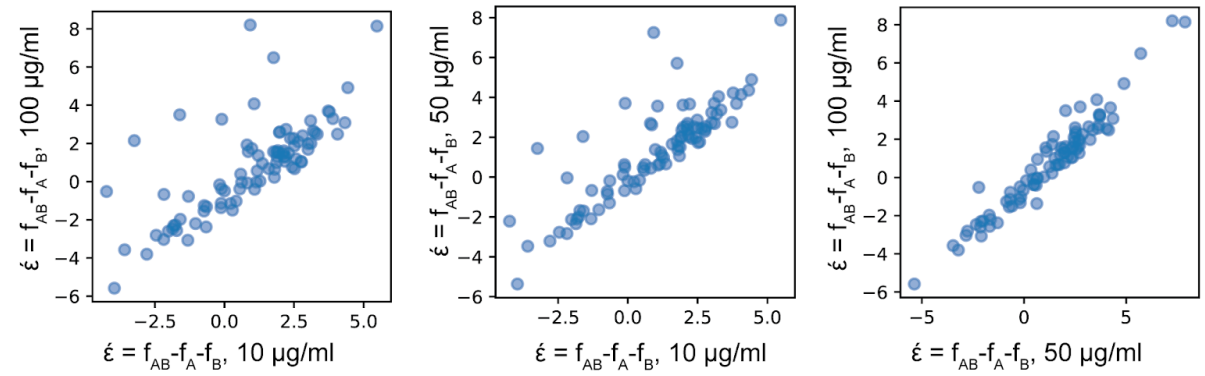

Figure 4.19: Comparing epistasis between environments: (A) Correlation of magnitude of epistasis determined by ε = fAB - fA - fB at 2 different concentrations of Rifampicin.

resistance even further.

4.3 Conclusions

In summary, we presented the CREPE technology for introducing mutations using an error-prone PCR donor and a single gRNA to introduce targeted diversity on the genome (**Figure 4.1**). Recombineering using a linear dsDNA donor with limited mutations is poorly understood. We found that recombination using dsDNA donor with few mutations is limited by sequence diversity. By modifying donor mutation rate, homology arm length, and mismatch repair we significantly increased the library editing efficiency and reduced library biases (**Figure 4.2-4.7**). The improvements we made to the CREPE system suggest that recombination may follow an alternate RecA-mediated template switch model (**Figure 4.8**). Using CREPE, we were able to introduce a library of mutations across different essential and non-essential genomic loci (**Figure 4.9**). The high efficiency and diverse mutant library allowed us to perform DMS for naturally accessible mutations that confer resistance to rifampicin (**Figure 4.10-4.19**).

Beneficial mutations for strong selections such as antibiotics can be identified using low-editing efficiency libraries as well. For example, rifampicin resistance mutations have been identified using the high-throughput CRISPR-enabled trackable genome engineering (CREATE) technology using several gRNAs (Garst et al. 2017). However, in CREATE, despite using a much larger library with saturation mutagenesis of the same *rpoB* window, the study managed to identify 10 resistance conferring single mutations at three residues, as opposed to over 41 beneficial mutations identified at 17 residues identified with CREPE. Additionally, as opposed to any previous study, using the sensitive fitness estimates from CREPE, data collected from one week of experiments corroborated biochemical, evolutionary and epidemiological findings gathered over

two decades of previous research. These included mapping of rifampicin interacting and clinically significant residues (**Figure 4.13**), the biochemical basis of increased resistance paths at lower concentration (**Figure 4.14**), and the effect of rifampicin dosage on favoring different spectrum of mutations (**Figure 4.14-4.15**). Additionally, using the double mutants obtained using the error-prone PCR template, we also managed to understand the trends in epistasis and evaluated the impact of compensatory mutations on improving fitness (**Figure 4.16-4.17**), which plays a significant epidemiological role in transmission of drug resistant *M. tuberculosis*. To the best of our knowledge, no other technology to date has demonstrated the fitness estimation for an essential gene at such high-resolution on the *E. coli* genome. Such expedited DMS of essential genes can be an invaluable tool for rapid understanding of their function in diverse cellular processes.

The higher efficiency, diversity and better fitness estimates with CREPE as opposed to other high-throughput CRISPR mediated technologies is due to the use of a single gRNA. A single gRNA ensures that all cells during the library construction undergo the same amount of DNA DSBs, which can help avoid biases in the library construction. The single gRNA would also correspond to the same synonymous PAM mutation for all library variants further controlling sources of noise in fitness estimates. The diversity can be generated in a single transformation step. CREPE is inexpensive as it can be performed using only eight primers as opposed to pools of thousands of oligos. All plasmids in the CREPE system are curable. Curing Cas9 can further ensure the elimination of a bulky plasmid and off-target activity of Cas9 during selection. By curing the plasmids, the selection can be performed in the wild-type genetic background in the absence of plasmids and antibiotics-markers. Finally, we directly sequenced the genomic loci to

track the variants, which allows direct fitness estimates without the need to use barcodes as a proxy.

4.3 Materials and Methods

4.3.1 Strains and plasmids and cloning Methods

All editing experiments were performed in *Escherichia coli* K-12 MG1655 strain. The E coli was grown in Lysogeny Broth (LB) media for all experiments.

We first switched the promoter of the cas9 from the pBAD promoter to the pro1 promoter (Davis, Rubin, and Sauer 2011) using a previously published plasmid, pX2-cas9 (Bassalo et al. 2016) (**https://www.addgene.org/85811/**). In order to construct the pCREPE plasmid, the pro1 Cas9 cassettes was amplified and cloned into the pSim5 plasmid that expresses lambda Red recombination proteins under the lambda phage pL promoter, controlled by the temperature sensitive cI857 repressor. The pSim5 plasmid was obtained from the court lab (Datta, Costantino, and Court 2006). The pCREPE+mutLE32K was constructed by cloning *mutL-E32K* gene in the lambda red recombination operon of pSim5 using the primers described previously (Nyerges et al. 2016). This allowed expression of *mutL-E32K* using the temperature inducible pL promoter as well, such that the MMR machinery was blocked only during recombineering. The gRNA plasmid under J23119 promoter was purchased from addgene (https://www.addgene.org/71656/). For each of the genome targets, different spacers were cloned in the gRNA for the different targets (**Appendix Table 6**). In order to construct the error-prone PCR libraries, the target with end

homologies (500 bp for *galK* and 250 bp for *crp*, *rpoB* and *mreB*) on each side were amplified from the genome and cloned in the pSAH031 backbone (https://www.addgene.org/90330/), which was recently developed for construction of unbiased plasmid mutagenesis libraries (Higgins, Ouonkap, and Savage 2017). Subsequently, 250 bp long gblocks with the Synonymous PAM mutation (SPM) used to replace the wild-type sequence in the error-prone plasmids using CPEC cloning. All amplifications were performed using Kappa biosystems high fidelity polymerase (catalog #07958897001) following manufacturer defined amplification protocol. The plasmid inserts and backbones were amplified with sufficient overlapping homology (35-40 bp long) and all cloning was performed using circular polymerase extension cloning or CPEC (Quan and Tian 2011) using12.5 µL (with at least 100 ng of backbone) of an equimolar insert: backbone mixture, and 12.5 µL of NEB 2X Phusion Master Mix (Catalog #M0530). PCR reactions were performed using cycling conditions initial denaturation at 98 ^{O}C-30 s,10X(98 ^{O}C-10 s, 55 ^{O}C-10 s, 72 ^{O}C-90 s), and a final extension at 72 ^{O}C for 120 s followed by a hold at 12 ^{O}C). Ten µL of the CPEC reaction was transformed into competent cells by electroporation. The transformed cells were plated on LB and appropriate antibiotics as listed in the table.

4.3.2 Error Prone PCR library construction

The error prone PCR libraries were constructed using the Agilent GenemorphII Random mutagenesis kit (Part #200550). Primers were designed for each targeted region with overlaps to allow cloning into the pSAH031 backbone with (**Appendix Table 6**). The mutation rate in the PCR reaction can be controlled by altering the initial template amounts. For the experiments with high and low diversity libraries of *galK*, we used 10 ng and 400 ng of template respectively. For the

error prone PCR for *crp*, *rpoB* and *mreB* libraries, we developed high-diversity libraries only using 10 ng template/reaction. The following error prone PCR conditions were used: Initial Denaturation:95^{O}C - 2', 30X(95^{O}C - 30s, Tm - 30s, 72^{O}C-1min), Final extension: 72^{O}C-1min. The PCR Tm were determined using NEB Tm calculator. After error-prone PCR, the libraries were gel purified and cloned into their respective backbones with unmutated end homologies using the NEBuilder HiFi DNA assembly kit (catalog #E2621). The libraries were cloned in Lucigen Elite E. cloni electrocompetent cells (catalog #60061) and grown on LB+Kanamycin overnight at 37^{O}C In each error prone PCR experiment, over 50,000-100,000 colonies were scraped in LB and the plasmid error-prone library was extracted using the Qiagen Mini-Prep extraction columns.

4.3.3 Cas9-mediated lambda red recombineering

Cas9-mediated lambda red recombineering was used to introduce single and library of edits on the genome following the heat-shock protocol (Sharan et al. 2009). The homology donor templates with single edits or error-prone libraries were amplified and gel purified with 250 bp end-homology arm length unless described otherwise in the text. Starter cultures of cells with the pCREPE or the pCREPE+mutL plasmid were grown overnight at 30^{O}C. In the morning, fresh cultures were started with a 100-fold dilution of the overnight cultures in fresh media and grown at 30^{O}C up to a mid-log OD measured at 600 nm of 0.4-0.5. At this point, the cells were immediately placed in a shaking water bath set at 42OC and the lambda red recombination operon (and mutl-E32K in case of pCREPE+mutL) was induced for 15 minutes. Then, the cells were immediately placed on ice and rapidly cooled by shaking and kept on ice for 15 minutes. The cells were centrifuged at 7500 x g at 4 ^{O}C for 3 minutes. The pellet was washed with 25 mL 10%

glycerol solution by resuspending and centrifuging at 7500 x g at 4 ^{O}C for 3 minutes thrice. Finally, the cells were resuspended at 100-fold lower volume of 10% glycerol than the starting culture volume. Dialyzed mixtures of the gRNA expressing plasmid and the homology donors with single or library were electroporated into 50 µL of the washed, now electrocompetent cells. The cells were recovered in LB for 3 hours and subsequently plated on LB+agar and appropriate antibiotics.

4.3.4 Selection in rifampicin

As described above, we constructed 2 independent libraries of genomic error-prone PCR mediated rpoB libraries using the CREPE protocol. For each library, we scraped ~10,000-15,000 colonies after construction in LB and stored multiple glycerol stocks of the library. We also scraped ~10,000-15,000 of colonies obtained from cells transformed with the gRNA + dsDNA homology donor with only the SPM. Prior to the selection, we thawed one glycerol stock of library 1, 2 glycerol stocks of library 2 and 3 glycerol stocks for gRNA +SPM only grew them independently in 100 mL LB at 37OC for 4 hours to cure the pCREPE+mutL plasmids. Simultaneously, we diluted overnight cultures of single isogenic cultures of wild-type E coli MG1655 parent and *E coli* MG1655 + SPM only in LB and grew them at 37^{O}C for 4 hours. After growth, we normalized the OD measured at 600 nm for each of our samples and plated 100 µL of several dilutions of the cultures on 3 different concentrations of rifampicin. We used to plates with most well resolved colonies to determine the resistant CFUs/mL of rifampicin. We saved the glycerol stocks used for each time point for the zero time point (to) of the selection experiment and we scraped ~10,000-15,000 at each concentration as the final time point (to) of the selection experiment for next-gen sequencing.

4.3.5 Genomic deep sequencing

In order to perform sample prep DNA was extracted using the boiling protocol. 50 µL of cell sample for each experiment were washed with twice with PBS and finally suspended in 50 **µL** of TE buffer (pH 8.0). The resuspended cells were boiled in at 100OC for 10 minutes. Then we amplified each library using primers with Illumina Nextera adapters (Supplementary table 3) using Kappa polymerase following manufacturer guidelines. Each cassette-experiment PCR product was then amplified with a unique experimental Nextera barcode. All sequencing was performed using the Nextera Next-generation sequencing MiSeq 2X300 kit with Illumina.

4.3.6 Next-gen sequencing Data Analysis

A custom analysis pipeline was built to quantify recombination trends and fitness calculation for our experiments (Supplementary codes 1, 2 and 3). Paired illumina reads were first assembled using the usearch mergepairs algorithm (Edgar 2010). Then, the assembled reads were aligned to the wild-type sequence of the target gene being studied using the usearch global alignment algorithm (Edgar 2010). The alignment generates a text file which summarizes details (id, number of mutations, E-value, mismatches, indels and mismatch/indel positions as a qrowdots alignment output). Variant counts were estimated as value counts for the qrowdots alignment output. Finally, we wrote a custom code to extract information about the nucleotide and amino acid changes corresponding to each qrowdot output (Supplementary Code 1).

4.3.7 Fitness calculations for resistance to rifampicin

Fitness calculations for resistance to rifampicin with each replicate were estimated using two time point enrichment score calculation algorithm described previously (Rubin et al. 2017). The fitness for each variant was estimated as (Supplementary Code 2):

$$fitness,\ f = log\left(\frac{C_{i,sel}+0.5}{C_{wt,sel}+0.5}\right) - log\left(\frac{C_{i,input}+0.5}{C_{wt,input}+0.5}\right)$$

where input signifies a unique variant and wt designates the wild-type variant. Foreach score, we estimated an error using. We use poisson approximation to calculate the standard error

$$Standard\ Error = sqrt(\frac{1}{C_{i,sel}} + \frac{1}{C_{wt,sel}} + \frac{1}{C_{i,input}} + \frac{1}{C_{i,input}})$$

After fitness estimates, we use data filtering to eliminate reads that may be erroneous. Since we targeted an essential gene, occurrence of stop codons would be impossible. Therefore, we eliminated reads with the following filter:

Ci >= C,max-stopcodon + 2.56*sqrt(C,max-stopcodon)

After read filtering, we combined the fitness estimates from each replicate using maximum likelihood estimates for variant score and standard error using Fisher scoring iterations (Rubin et al. 2017). The same protocol was also repeated with only synonymous mutations in our dataset. We classified a mutant as resistant to rifampicin if the fitness was greater than 2.96 standard deviations than the mean fitness of synonymous mutations.

4.3.8 Epistasis Measurement

We first isolated the double mutants for which single mutants agreed with counts filters. We then calculated the sum of fitness scores and the error associated with the sum using standard propagation of error. In order to assign magnitude epistasis, we used the following definition:

FAB > FA+FB: positive

FAB = FA+FB: neutral

FAB < FA+FB: negative

Where FA, FB and FAB are fitness values for mutant A, mutant B and combine mutant AB. In order to statistically determine the magnitude epistasis, we performed a one tailed Student's t-test using the fitness and standard error in estimation of the fitness, where the null hypothesis that the sum of fitness is the same as the fitness of double mutant was rejected only when the p-value was <= 0.01.