MAPPING THE SEQUENCE-FUNCTION LANDSCAPE FOR ANTIBIOTIC RESISTANCE IN THE DHFR FAMILY

by

CARMEN RESNICK

A THESIS

Presented to the Department of Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

March 2023

An Abstract of the Thesis of

Carmen Resnick for the degree of Bachelor of Science
in the Department of Biochemistry to be taken June 2023

Title: Mapping the Sequence-Function Landscape for Antibiotic Resistance in the DHFR Family

Approved: *Calin Plesa, Ph.D.*
Primary Thesis Advisor

Dihydrofolate reductase (DHFR) is an essential enzyme in the folic acid synthesis pathway and has been the subject of intense study in recent decades[1,2]. Despite the wide diversity of homologs, research attention has primarily focused on DHFR proteins from a narrow group of organisms and their mutants[3,4]. In this study we focus on the ability of DHFR to both rescue metabolic function in a knock-out strain and to tolerate treatment against the antibiotic trimethoprim, which will allow us to understand how antibiotic resistance emerges given many evolutionarily divergent starting points. Changes in the mutational landscape of DHFR allows for varying survival rates in the presence of antibiotic inhibitors. We carry out a broad mutational scan using a library of nearly 1,000 DHFR homologs and 22,000 mutants synthesized using DropSynth gene synthesis[5]. Variant fitness is determined in a multiplex survival assay in an *E. coli* Δ*FolA*Δ*ThyA* knockout strain which allows for conditional selection dependent on external supplementation.

We have collected quantitative fitness data on 996 homologs and 22,483 mutants of the DHFR gene based on activity both in the presence and absence of inhibitors, in order to reveal sequence-function relationships and understand how correlations between the fitness landscapes vary as a function of evolutionary distance between homologs. This data can be applied towards

the development of narrow-spectrum and targeted antibiotics and mitigation of resistance through understanding the sequence-function relationships which drive antibiotic resistance.

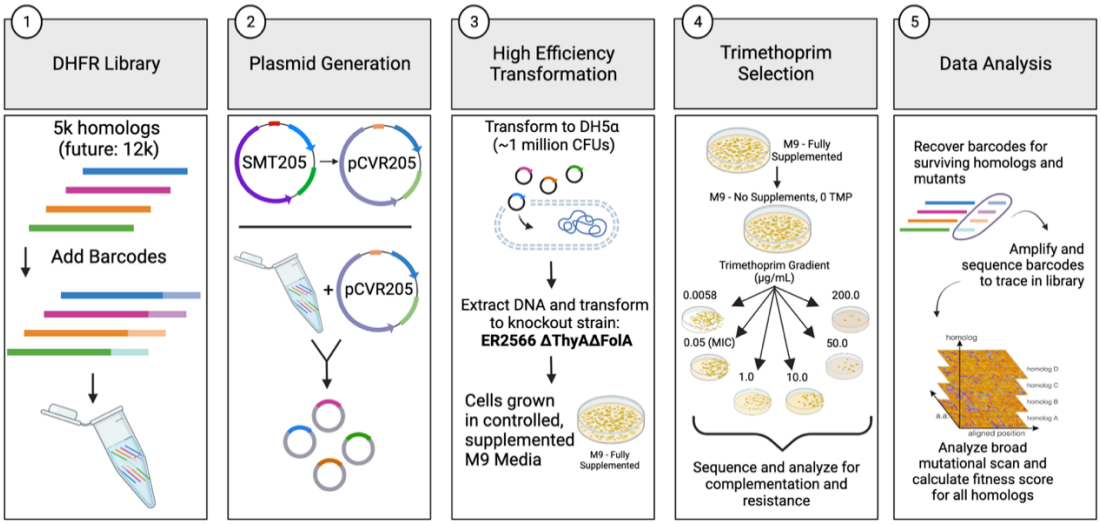


Figure 1: Methods Overview

Overview for the methods conducted in this paper to achieve a broad-mutational scan for DHFR homologs. (1) Library generation was conducted with DropSynth technologies, (2) plasmid SMT205 was adapted from the Reynolds lab to pCVR205, addgene ID #198715, (3) library was expressed in two successive strains to increase transformation efficiency and minimize diversity bottlenecks, (4) selection conditions listed were used to determine both complementation and antibiotic resistance in the library, (5) barcodes for DHFR found in surviving cells was recovered and sequenced to elucidate sequence-function relationship to cell fitness.

Acknowledgements

I would like to begin by thanking the Plesa lab and Dr. Calin Plesa for welcoming me into one of the most supportive and inspiring environments I could imagine. This lab space and the people in it have taught me a love for research and synthetic biology and have welcomed me into the world with open arms. I would like to especially acknowledge Sam Hinton for teaching me all the lab techniques and skills that I know as well as guiding me in how to put the science that I now know so well into words on a paper or in a presentation.

I would also like to recognize Dr. Plesa whose support has provided me the space to travel across the world to present my own work and form connections with professionals in the field, including the opportunity to lead a team of other students to an international competition to present our work in forming a genetically engineered machine (iGEM). This environment has been the best launch pad for a career in bioengineering that I could have asked for.

A huge thank you to the KCUS for giving me a community of other driven students with whom I can share my work, practice public speaking, and share joy in our respective studies. Thank you to my phenomenal Honors College Advisor and representative for my thesis, Brian McWhorter, who helped encourage me throughout my time at the University of Oregon.

Finally, I would like to address the friends and family who have helped me along the way. Thank you to my parents, Shanti and Scott, for encouraging me to continue fighting and studying, even in the toughest times. Thank you to my classmates and friends for the continued support and camaraderie in our studies. And lastly, thank you to my roommates who have stood by my side through late lab hours and study sessions.

Table of Contents

List of Figures

List of Tables

List of Equations

Introduction

As the population increases and temperatures rise, our world becomes increasingly susceptible to disease. One of these culprits which we confront regularly are bacterial infections. Bacteria thrive and flourish in warm and nutrient rich regions, making places like bodies, water, and soil the perfect habitat for propagation. Furthermore, an increase in population density and overcrowding, is seen to correlate with increased instances of infections[6]. These diseases are being approached with tenacity by modern medicine with anti-bacterial drugs (antibiotics), however, the overuse of antibiotics has forced bacteria to evolve and mutate to evade these treatments. When bacteria develop the ability to tolerate inhibition by Western medicine, treatments become less effective, leading to what is known as antibiotic resistance. The Center for Disease Control (CDC) has called antibiotic resistance "an urgent global public health threat"[7], emphasizing the importance of understanding how bacteria are able to develop resistance to the medicine we use against them.

Here, we present a study which explores the mutational landscape for antibiotic resistance in the enzyme dihydrofolate reductase which can be used to inform and assess development of new drugs against target organisms. Mutational landscapes map out the connection between sequence and function in the genetic code of enzymes[8]. We seek to continuously build upon a large body of knowledge to inform a deeper and more rounded understanding of the mechanisms which lead to antibiotic resistance in the enzyme dihydrofolate reductase (DHFR). DHFR is vital in rapidly dividing cells, making it a key target for antibacterial drugs.

Antibiotic Resistance

Bacterial infections are able to propagate rapidly through successive cellular division which allows the infection to exponentially multiply and spread throughout the host[9]. Many

antibiotics are designed to interrupt the division of bacterial cells in order to stop the infection. Antibiotics have become popular both in the treatment of human diseases, but are also ubiquitous in use with animals, livestock, and water treatment[10]. Antibiotics were first identified for their medical use in 1928 with the discovery of penicillin and have since increased in the popularity of use. Discovery of these drugs, however, has since declined[11].

The rate of drug discovery slows as the evolution of resistant bacteria escalates, leading to an influx of trouble caused by an inability to defend against the resistant bacteria with existing tools. Bacterial evolution leads to resistance when changes to the DNA encoding certain proteins produce a change to the function or structure of the protein that allows the bacteria to resist the treatment. Mutations to these genes can be either detrimental, have no effect, or have positive growth effects. The latter of these results is the most concerning. Strong selection pressures will reveal fully resistant mutants, whereas weak pressures allow for resistant mutations to develop over time. We are no longer able to treat an increasing number of bacterial infections using existing medicines because the bacteria have evolved to evade these treatments[11].

Dihydrofolate Reductase

Current methods suggested for limiting the spread of antibiotic resistance include minimizing use in healthcare, thorough education on antibiotics and AR, and minimizing the spread of infections as a whole[10,12]. Clinicians should also be encouraged to limit use to clinical scenarios where antibiotics are strictly necessary and to test for resistance to streamline antibiotic choice. Here, we have constructed a method to assay hundreds of thousands of mutations to the genetic sequence for an enzyme vital for DNA synthesis in cells, Dihydrofolate Reductase (DHFR). DHFR functions by catalyzing the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) through the cofactor NADPH[9] (figure 2). THF is vital for the synthesis of thymidine, one

of the four DNA nucleotides. The role of DHFR in the folate synthesis pathway in DNA replication, shown in figure 2, makes it a promising target for anti-folate drugs[9].

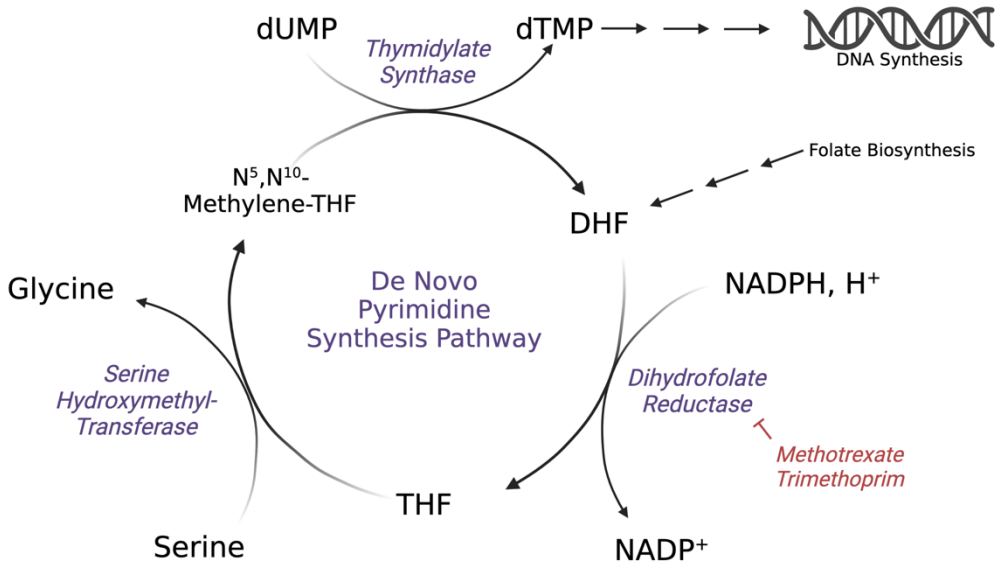


Figure 2: The De Novo Pyrimidine Synthesis Pathway

Dihydrofolate reductase catalyzes the reduction of dihydrofolate to tetrahydrofolate with the cofactor NADPH. DHFR function is tightly coupled with thymidylate synthase. This cycle is an intermediate step in the synthesis of thymidine, one of the four canonical DNA bases. DHFR is inhibited by folate inhibitors such as methotrexate and trimethoprim.

DHFR is targeted by two major drugs – trimethoprim and methotrexate. Methotrexate is designed to specifically target human DHFR making it an appropriate anti-cancer drug, whereas trimethoprim is designed to specifically target bacterial DHFR[13].

DHFR functions both independently and as a bifunctional enzyme in conjugation with the enzyme thymidylate synthase (TYMS)[14]. Both are required together to regulate folate metabolism and respond similarly in response to inhibition[14]. TYMS serves to oxidize the reduced form of DHF back to THF. The coordination between these two enzymes is scrutinized in a comparative genomics map observing 16 proximal genes to the DHFR gene (folA) in an assay to assess the evolutionarily adaptive coupling between DHFR and TYMS, along with other folate-involved

genes[14]. Another important consideration is the intermediate molecules controlled by these two enzymes. Overexpression of DHFR causes a toxic buildup of the intermediary product and metabolic imbalance[15].

The vector described in this study incorporates controlled expression of both DHFR and TYMS. The plasmid pCVR205 used here is adapted from SMT205, a plasmid generated for the controlled expression of both genes alongside the protease Lon. The study conducted a deep-mutational scan to elucidate mutant populations which demonstrated flexibility in response to negative inhibition by enzyme degradation. They measured the rate of DHF conversion to THF as well as selection coefficients on growth rates of the mutants in order to directly compare the fitness of each mutant in the presence and absence of an inhibitor. They found that the mutations and their respective tolerance towards the Lon was grouped by the location of the mutation across the gene[16].

Current literature is limited to observing mutations along a single homolog or small subset of mutations to the gene. Many previous studies observed a few specific mutations to the gene, one of which is mutation to the aspartic acid at residue 27 (D27). A study in 2019 observed conversion rates of DHF to THF upon mutation for D27. They found that mutations at this particular region of the gene conferred decreased growth rates as confirmed by the catalytic activity measured for each of these mutants[3]. This finding is further understood by the correspondence between D27 and the Met20 loop region in DHFR which has been found to engage with the conformational change from binding with NADPH[17].

In our study, DHFR was chosen because it has been widely studied making it a good target molecule for comparative literature. We are able to compare the data assembled in this study with existing data to verify the results, and its short length of only 150 amino acid residues makes it an appealing candidate for the application of high-throughput gene synthesis with DropSynth[5].

Broad-Mutational Scanning

In order to address the mutational scope of DHFR, we have utilized a proprietary technique called DropSynth which allows us to synthesize large and diverse genetic libraries for a low cost. DropSynth uses barcoded magnetic beads in conjunction with a computationally designed collection of 300-bp oligos to generate libraries of genes in a single tube. Genetic sequences are bioinformatically fragmented into 300-mer oligos. These oligos are then matched with complementary barcodes on streptavidin-coated beads and designated to a single oil droplet for each unique gene. The gene fragments can then be cleaved from the beads and assembled within these droplets to form a complete library with any collection of genetic sequences[5].

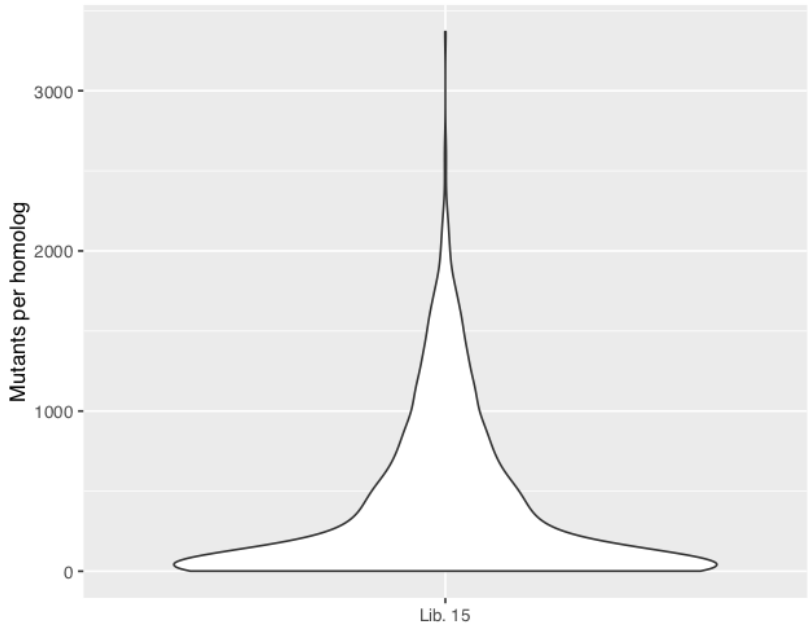


Figure 3: Distribution of mutants within library 15

In concurrence with the homologs which have been synthesized in the library, any infidelity of polymerase activity generates mutants within the library as well. Figure 3 displays a violin plot of the full library used in this study (library 15) and the number of mutants present per homolog. The mean value in this chart is 187 mutants on average per homolog. Figure 3 shows

how the mutants are dispersed within their respective family, ranging from 0 to 150 amino acid sequence mutations away from the parent sequence. The majority of mutants fall within 25 to 100 mutations, though some extremes exist as well, due to deletion errors in the oligos which introduce frameshifts.

For this study, we have taken a previously synthesized *DropSynth*[5] library of 1,536 homologs, each with their respective mutants, and transformed them to a knock-out strain which lacks both DHFR and its confidant TYMS.

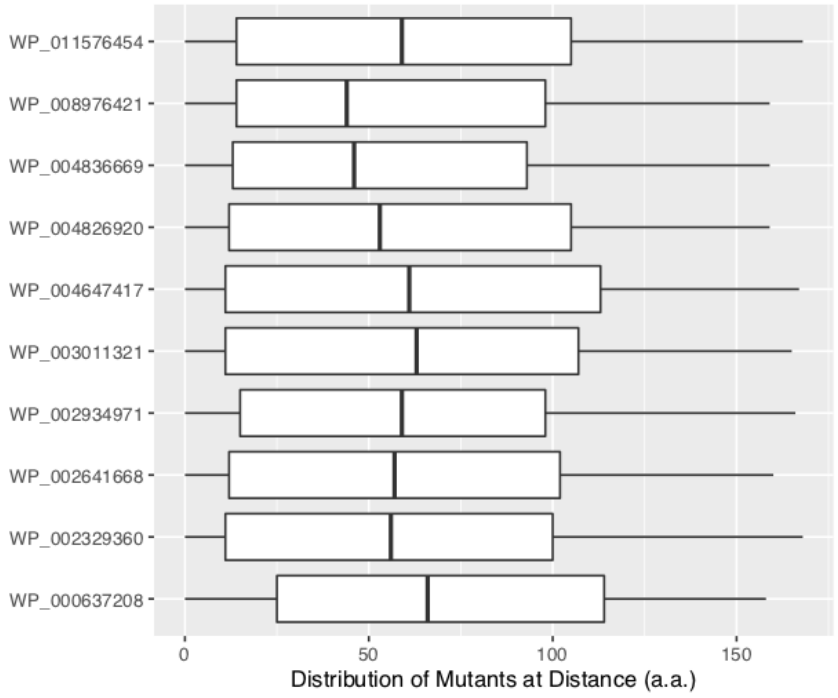


Figure 4: Distribution of mutants from parent sequence for top 10 homologs

Deep-mutational scanning (DMS)[8,18] is a modern technology which looks at large libraries of gene variants to understand how mutations to a gene influence its function. Deep-mutational scans take one gene and make alterations across the DNA sequence, encoding the protein or enzyme through saturation mutagenesis where mutations are systematically introduced across the entire gene[8]. The libraries are often tested in a model organism and exposed to evolutionary pressures or assays to determine the function. Function is mapped against sequence to develop a

mutational landscape that corresponds to the overall fitness. These dataset are known as multiplexed assay of variant effects (MAVES)[19], and are continually being built upon to generate a full database of these mutational assays for many genes of interest.

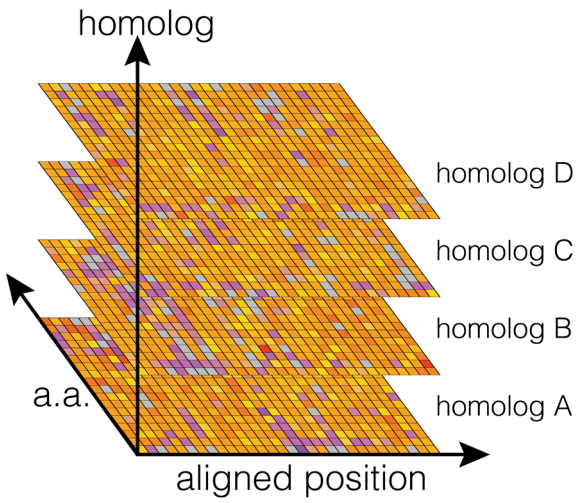


Figure 5: Generic Broad-Mutational Scan

Moving through the z-axis in this diagram assigns each unique homolog to a respective, deep-mutational scan (DMS) for an arbitrary gene. The DMS assigns a specific fitness score for the gene given each unique combination of amino acid along an aligned sequence. The fitness score here is given by color on the heat map, red indicates a positive fitness and blue indicates a negative fitness (Equation 1).

Our DropSynth libraries incorporate an additional level into a deep-mutational scan. We have assayed deep-mutational scans for a library of homologs for the DHFR enzyme, producing a broad-mutational scan (Figure 5), which diversifies the dataset to across divergent species. This format of analysis expands the capabilities of standard DMS to include genes which have common evolutionary starting points. This is particularly accommodating with the DropSynth library which allows for the synthesis of DHFR from divergent species. The broad-mutational scan allows for most unique homologs to be mapped against one another despite the nonconformity of their sequence-space relationship.

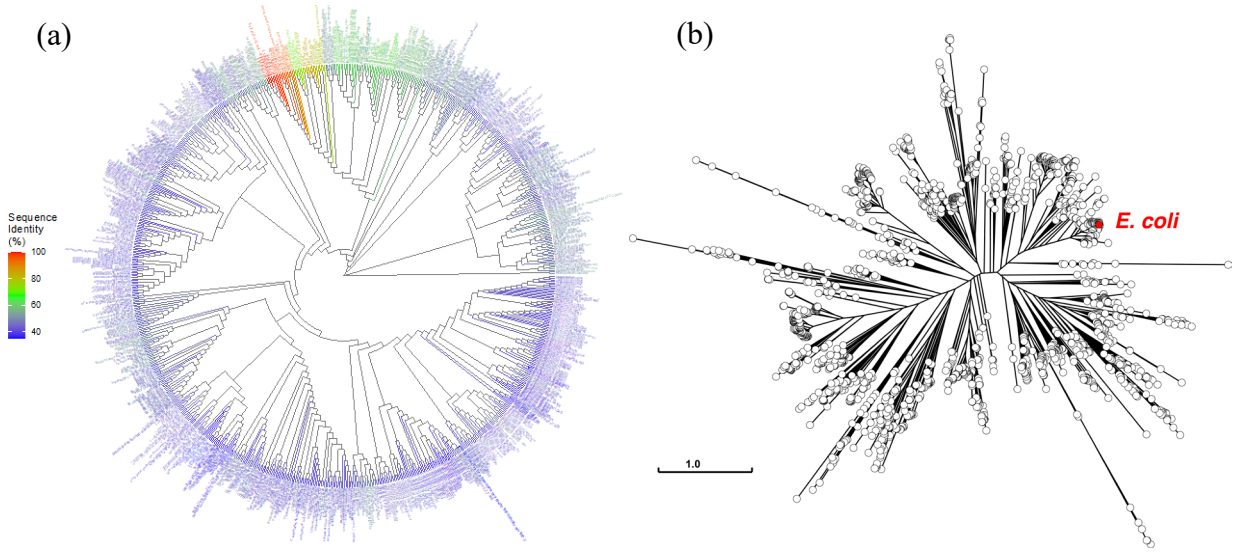


Figure 6: Library distribution

(a) This chart shows the percent of similarity in amino acid sequence between the wild type DHFR in *E. coli* and the other homologs in the library. The colored labels are species names, and the color corresponds with the percent identity seen in the legend on the left. The region in red shows a 100% identity similarity, red is wild type *E. coli* DHFR.

(b) Phylogenetic tree for DHFR library, branch length corresponds to the evolutionary distance of each homolog

Here, we express a library of 1536 divergent homologs in a strain that has both DHFR (FolA gene) and thymidylate synthesis (ThyA gene) recombinantly removed from the strain's genome, forcing the cells to rely solely on the inserted DHFR gene[16,20]. The distribution for this library is shown in Figure 6. Figure 6 (a) illustrates the identity of the library sequences relative to the native structure of *E. coli* DHFR, and the depth of evolutionary divergence for the library can be seen by the branches in figure 6 (b). Each unique genetic sequence in the library contains a random 20 base pair barcode at the end of the gene which allows the gene to be traced back to its exact sequence. In total, this study compiles the largest assay for a single enzymatic family conducted to date. We are in the process of expanding this collection to a single library with 12,000 homologs and their respective mutants.

Methods

The general strategy employed in this study is detailed below in figure 1 in the abstract. A plasmid designed to mediate DHFR expression was adapted to comply with restriction enzyme sites in the DropSynth library, and standard cloning techniques were used to combine the two systems. The library was then transformed to NEB DH5α cells, collected, and replated on to minimal media and a gradient of trimethoprim concentrations. Surviving cells were recovered and the barcodes were extracted from each condition. Barcodes were sequenced and assayed against initial conditions to determine fitness of homologs and the respective mutants in the library. All steps in figure 1 are described in detail below.

Plasmid Generation

The delivery vector for DHFR expression in *E. coli* was derived from the plasmid SMT205 (supplied by the Kortemme/Reynolds Lab, Addgene #134817). SMT205 is designed to employ leaky expression by lac-o repression without IPTG induction in order to mitigate concentration of DHFR in the cells, which can be toxic when overexpressed due to the build-up of the intermediate [14,16]. The plasmid was modified to remove all excess restriction enzyme sites which were required for cloning the library. Three restriction sites were then reintegrated to be synonymous with the FolA gene library (figure 7).

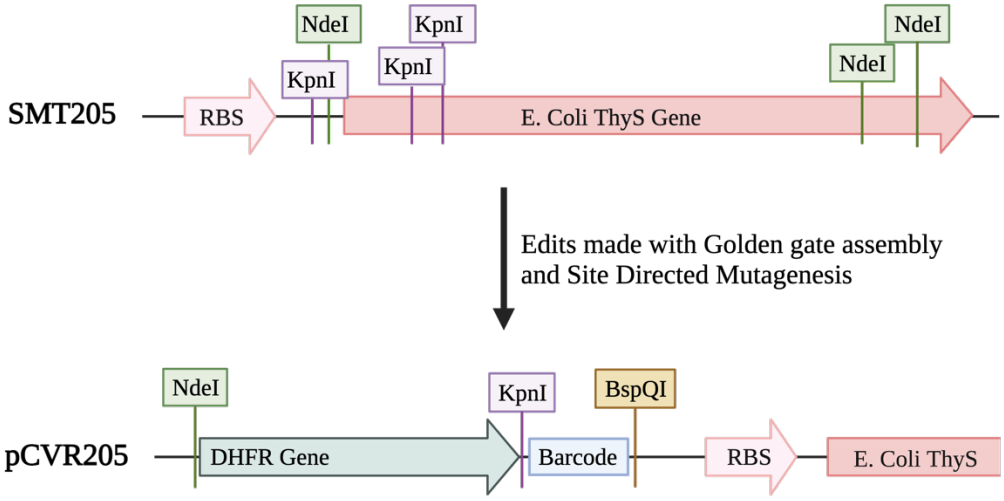


Figure 7: Edits to generate final plasmid, pCVR205

NdeI and KpnI cut sites were removed from the wild-type ThyA gene sequence by utilizing the degeneracy of the genetic code. A gBlock from IDT was designed to remove the restriction enzyme sites and insert one KpnI site and one BspQI site respectively at the 3' end of the C-terminus of the DHFR gene. The sequence for the gBlock is listed in supplementary materials. Golden gate assembly was used to integrate the gBlock into the SMT205 backbone using primers SMT205_bb (supplementals).

The edited plasmid, pCVR1, was transformed to 25µL of electrocompetent DH5α cells and incubated at 37˚C for 16 hours. A combination of internal and external primers to the ThyA region were used to validate the gBlock insertion with sanger sequencing. One colony was selected for Site Directed Mutagenesis (SDM) to make a three-nucleotide substitution to add NdeI to the 5' end of the FolA gene. The primers used for SDM are named pCVR205_SDM and are listed in supplementals.

The new construct, pCVR205, was transformed to electrocompetent DH5α cells (NEB). Single colonies were selected to verify the construct through Sanger Sequencing. This final plasmid can be found at addgene address #198715.

Preparation of Knock-out strain

ER2566Δ*ThyA*Δ*FolA*[20], folate-deficient mutant of ER2566 strain, was donated by the Reynolds lab. Colonies with the correct knockout were made compatible for electroporation with the protocol listed in supplementals for library transformation.

Optimizing transformation efficiency for library diversity

In order to represent the full library and decrease the diversity bottle neck, the protocol was enriched to maximize ligation efficiency and increase total yield in the initial transformation. Library 1 synthesized in DropSynth2.0[5] was amplified using library-specific primers, mi3_R1 and EV4_BspQI. The library was digested and ligated to the plasmid pCVR205 with T4 ligase. A detailed protocol to increase the ligation efficiency can be found in the supplementary materials.

170 ng of ligated product was transformed to 25 µL DH5a cells. Cells were allowed to recover in SOB media (supplementals) then transferred to LB plates along with a gradient of 10X dilutions to quantify the approximate number of undiluted cells plated.

The library was recovered by scraping the cells from the plate using sterilized glass rods. Cells were pelleted and miniprepped with NEB's Monarch Plasmid DNA Miniprep Kit to extract the plasmids containing the library transformants.

Selection and complementation

Media

The experiments were conducted in minimal M9 media made with 1X M9 salts, 0.4% (w/v) glucose, 2 mM magnesium sulfate ($MgSO_4$) and 35 µg/mL chloramphenicol (CAM). Supplemented media was made by adding 80 µM adenosine, 0.5 mM glycine, 4.5 mM inosine, 4.2

μM calcium pantothenate, 0.5 mM methionine, and 0.21 mM thymidine to the M9 media. All components were sterile filtered prior to use.

A trimethoprim gradient from 0 to 400 times the minimum inhibitory concentration (MIC) of 0.5 μg/mL[21] was added to the minimal media with the following concentrations for both liquid media and agar plates: 0 μg/mL TMP, 0.058 μg/mL TMP, 0.5 μg/mL TMP, 1.0 μg/mL TMP, 10 μg/mL TMP, 50 μg/mL TMP, and 200 μg/mL TMP (all components of the media were purchased from SigmaAldrich).

Preparing controls and library

Three control plasmids were constructed using the *E. coli* wild-type DHFR, D27N[3] (a low-function DHFR mutant), and mCherry[22], a non-catalytic Red Fluorescent Protein (RFP). RFP was chosen as a negative control during complementation selection. Plasmids were constructed using standard cloning techniques and electroporated to DH5a cells. Barcoding of controls was done by integrating 20 randomized nucleotides into primer tails and ligated using In-Fusion HD cloning. 15 colonies for each control were selected and sequenced. Controls were uniformly combined and spiked into the library distribution in a 0.1% molar ratio.

Initial Transformation

50 ng of the combined library and controls were transformed in quadruplicate to 80 μL of ER2566Δ*Thya*Δ*Fola*, via electroporation. The electroporated cells were recovered in SOB media, described above. All four transformations were combined and plated on LB and agar plates with chloramphenicol and thymidine supplementation. An additional 6, 10X serial dilutions were plated to quantify transformation efficiency. Cells were grown at 37˚C for 24 hours.

Selection

All cells were scraped and recovered from the plates in liquid LB media. Scrape was washed to remove any supplements with three iterations of spins at 2,500 rcf for 5 minutes, discarding the supernatant each time and resuspending pellet with minimal media. Stocks with 1:1 ratio of cells to 50% Glycerol were saved. Cells were recovered in supplemented M9 media and diluted to an OD600 of 3 units per mL (about $2.4x10^9$ cells per mL). Stocks were saved in 25% glycerol solutions and stored at -80˚C. 5 mL of cells were divided between fully supplemented M9 media plates (described in supplemental materials) and grown at 37˚C for 22 hours. The remaining culture was pelleted and miniprepped with NEB's Monarch Plasmid DNA Miniprep Kit extract the plasmids for sequencing.

Plates were scraped and pellets were washed and recovered as described above in non-supplemented media and diluted to an OD600 of 3. Diluted cells were divided between plates with minimal media and minimal media supplemented with a trimethoprim gradient[4,14,23] in supplementals table S1. Plates were grown at 37˚C for 22 hours. Cells were scraped, recovered in non-supplemented media and miniprepped to extract plasmids for sequencing.

Sequence Screening

PCR was conducted using a pool of 5 pairs of primers designed for compatibility with Illumina Sequencing (table S2). Primers were mixed to a final concentration of 10 ng/µL. Samples from each timepoint were then independently amplified a set of unique primers for Illumina sequencing for a total of 11 timepoints throughout the study. The timepoints are as follows: DNA before transformation (miniprepped from DH5a) immediately post-transformation before growth on plates, after growth on LB media with supplements, after growth on M9 fully supplemented,

M9 without supplement, 0.0058ng/μL TMP, 0.5 μg/mL TMP, 1μg/mL TMP, 10μg/mL TMP, 50μg/mL TMP, and 200μg/mL TMP.

Results

From a library 1,536 unique DHFR homologs, we have gathered data for nearly 1,000 of these homologs and their mutants. Each homolog has an additional average of 187 mutants (Figure 3), arising from either polymerase or oligo synthesis errors in the initial library generation. A barcoding strategy was employed to recover data corresponding directly to the gene sequence and family for all genes present throughout each step described in the methods section of this paper. Through sequencing of the recovered plasmids from both complementation and selection of the DHFR genes, we recovered 7,736,350 usable reads which correspond to 1,621,584 mapped barcodes in the library. Within these reads, 512,729 are unique and mapped variants of a combination of the designed homologs and their respective mutants generated with DropSynth[5].

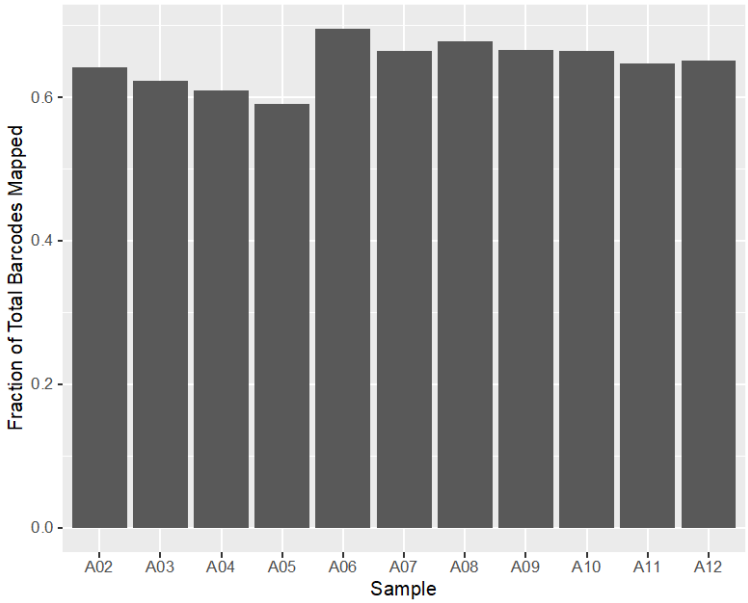


Figure 8: Fraction of Total Barcodes Mapped

This figure displays the total fraction of barcodes in sequencing results that have been previously mapped in the library at increasing concentrations of trimethoprim. A02 corresponds to 0 µg/mL of TMP, A12 is 200 µg/mL of TMP.

Of the barcodes recovered in this assay, between 60% to 70% of the reads have been previously mapped and constitute usable data, seen in Figure 8. This is a preliminary study which uses the aforementioned strategies to generate a body of knowledge which compares the genetic sequence for a large set of homologs and mutants with their ability to resist antibiotic inhibition.

Study Overview

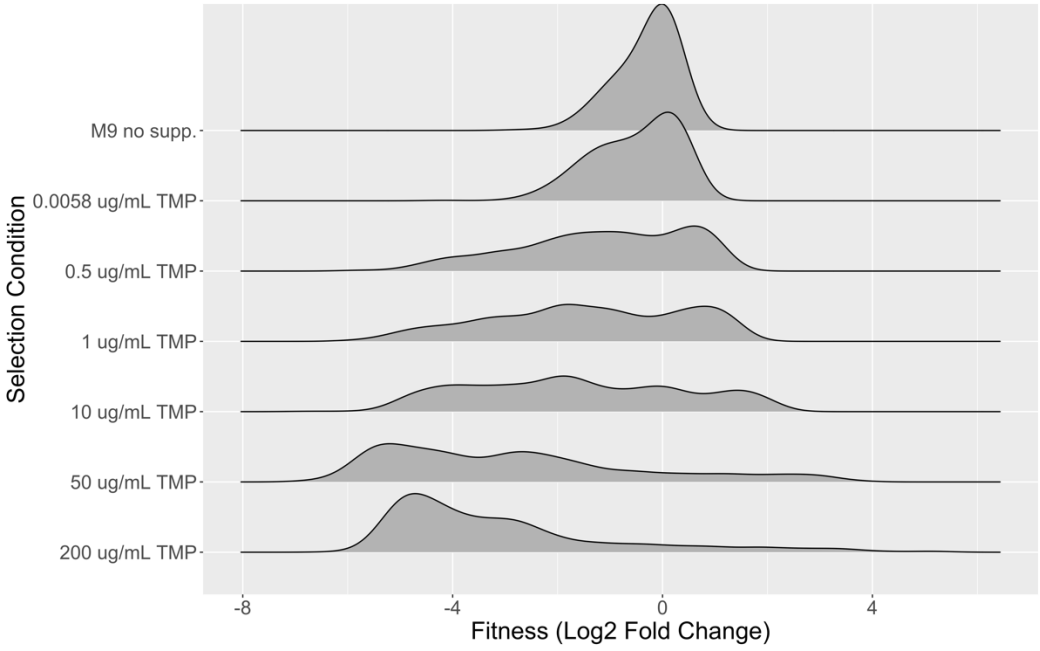


Figure 9: Overall barcode counts across a trimethoprim gradient

This figure displays the effect of selection pressures on cell survival across a gradient of 7 different trimethoprim concentrations ranging from 0 µg/mL to 200 µg/mL. The volume under each peak corresponds to the average fitness for the entire library at each selection condition as calculated with Equation 1. A negative fitness indicates an overall depletion of survival rates in the assay.

The expansive dataset recovered in this study lends itself to abundant methods of data analysis. We will begin by detailing the general populations present at each step of selection. As expected, the number of barcodes present in each selection condition decreases as the pressures are increased. Figure 9 validates the expected results, consistent with a decrease in survival as trimethoprim increases. This validates the methods employed here and supports future studies to

expand the library. The minimum inhibitory concentration of trimethoprim (TMP) against DHFR has been previously identified as 0.5 μg/mL[23].

Figure 9 displays a clear shift from neutral growth to depletion as trimethoprim selection concentrations increase. The areas of interest in the figure are the regions of positive growth indicated with a positive fitness score. These regions, especially seen in the thin tail on the right-hand side of the curve at 200 μg/mL, are comprised of genetic sequences which have antibiotic resistant properties. We also see a large number of sequences which have mild resistance, those with a positive fitness at low concentrations of trimethoprim. Further analysis for each of these variables may be used to inform future drug designs, predict the likelihood of arising resistance in certain species, and build models to predict resistance based on sequence.

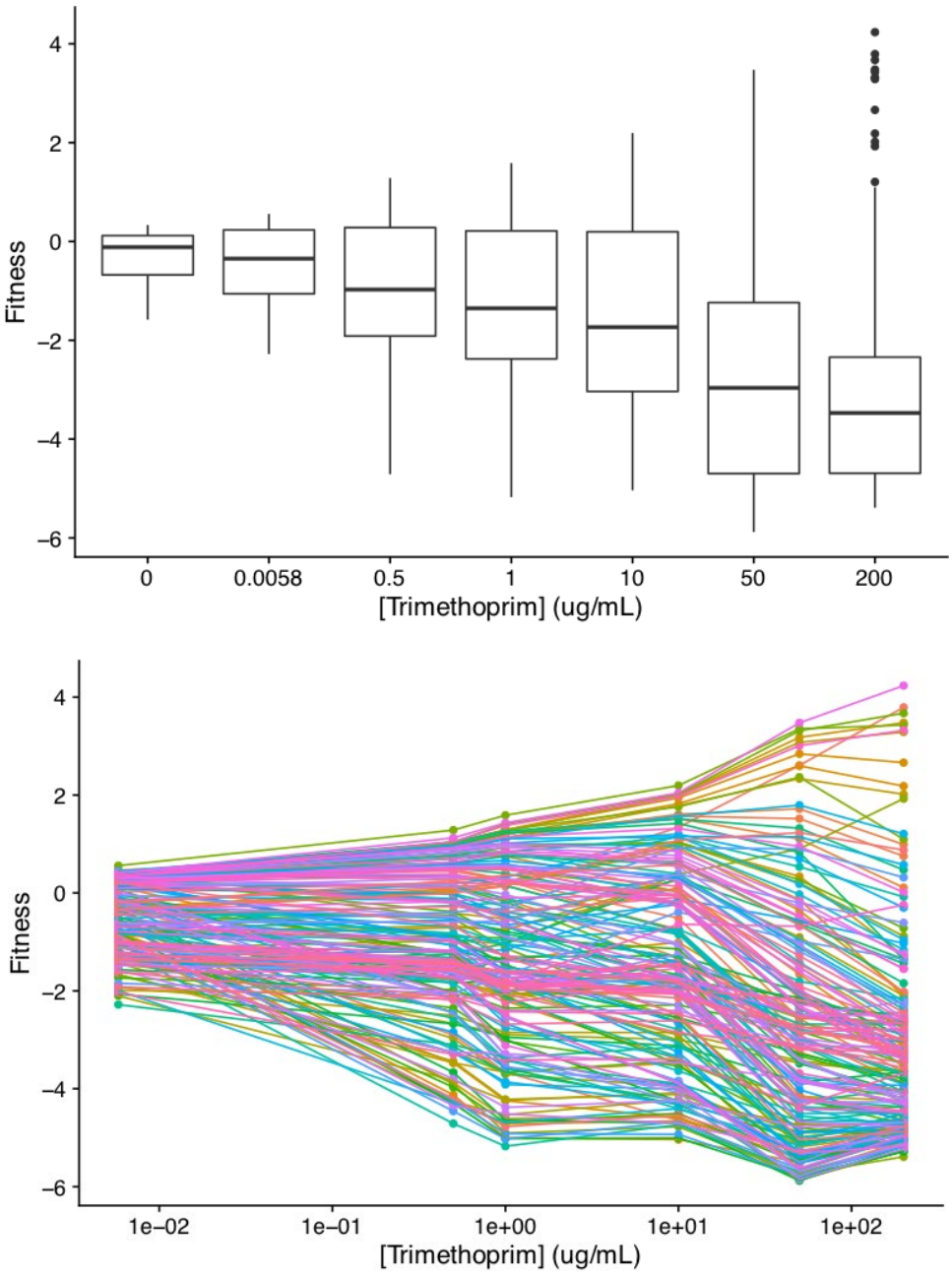


Figure 10: Dose response from selection of barcodes with highest abundance, relative to conditions

(a) Fitness for all 852 homologs with at least 5 barcodes

(b) Fitness vs trimethoprim concentration for the top 100 homologs with most barcodes (>299 barcodes)

Complementation

The library was initially narrowed through selection for the ability of the enzyme to complement function in *E. coli*. This was assayed by expressing the library in the knockout strain and subjecting the cells to minimally supplemented media without thymidine. These conditions tested the cell's ability to survive in stringent conditions with reductase activity reliant on a single DHFR gene from our library. A large percentage of the library was able to complement function. This was determined by examining the change in barcodes present in the assay between the initial transformation and selection in minimal media. The lack of supplementation in the media forced the bacteria to rely on the DHFR from the library, so we are able to assume that any barcodes present in the culture after growth without thymidine supplementation relates to complementing function.

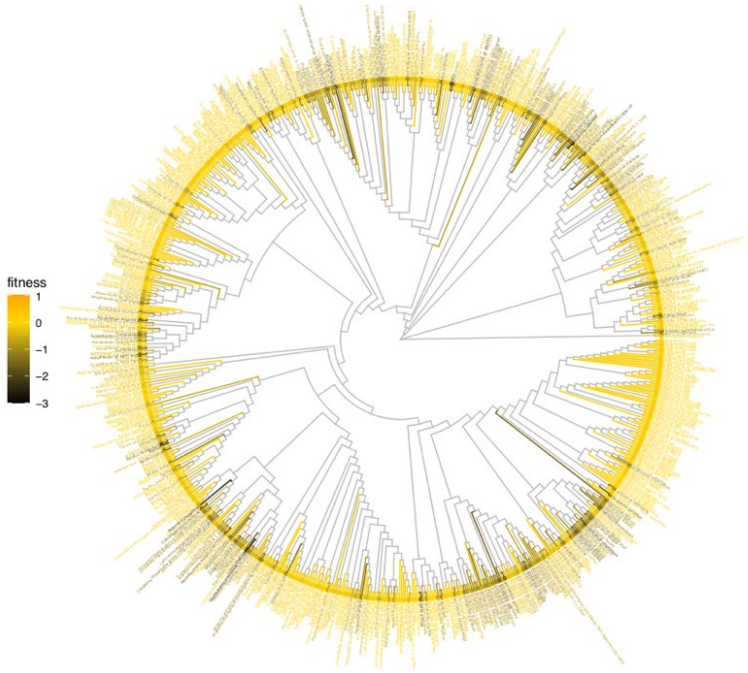
This assay is important because it provides a baseline to compare barcode counts between non-supplemented conditions and trimethoprim inhibited conditions. Figure 11 displays the average fitness of each homolog in the library. Fitness is calculated with Equation 1 below. Yellow corresponds to a neutral fitness score, orange is an enrichment of growth, and black is depletion of the homolog when comparing the initial growth to the species present after selection for complementation.

$$Fitness = \log_2(\mathrm{S2n} + 1) - \log_2(\mathrm{S1n} + 1)$$

Equation 1: Fitness

Complementation: S1n = population at full supplementation, S2n = population with no supplementation; Antibiotic Resistance: S1n = population with no supplementation, S2n = population under trimethoprim inhibition

(a)



(b)

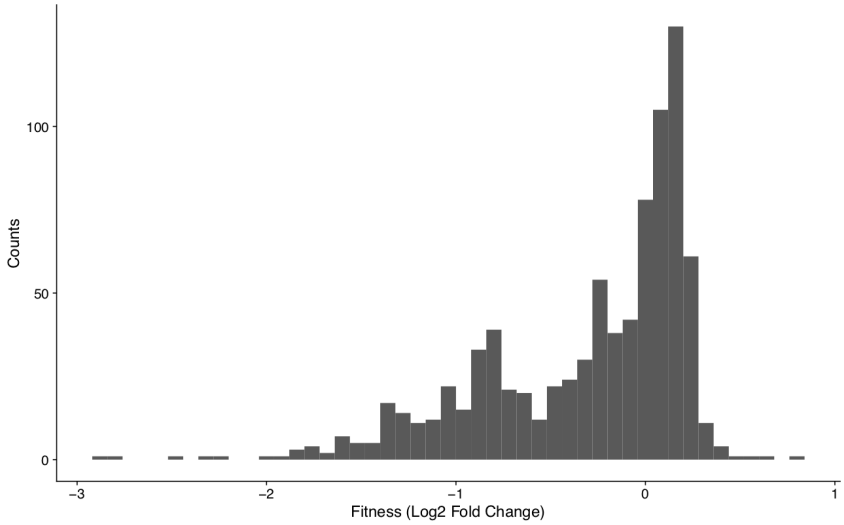


Figure 11: DHFR complementation in *E. coli*

Fitness is scored based on growth compared to the initial population. A fitness score of zero equates to maintenance of a constant level in the population, negative fitness indicates depletion, and a positive fitness score indicates enrichment of the gene in our study.

(a) Complementation of DHFR homologs to *E. coli* knockout strain, grown in non-supplemented minimal media conditions. Color corresponds to the fitness as denoted on the legend to the left, and each branch is grouped by evolutionary distance.

(b) Graphical representation of the data presented in (a). This graph displays the number of barcodes present at each fitness level.

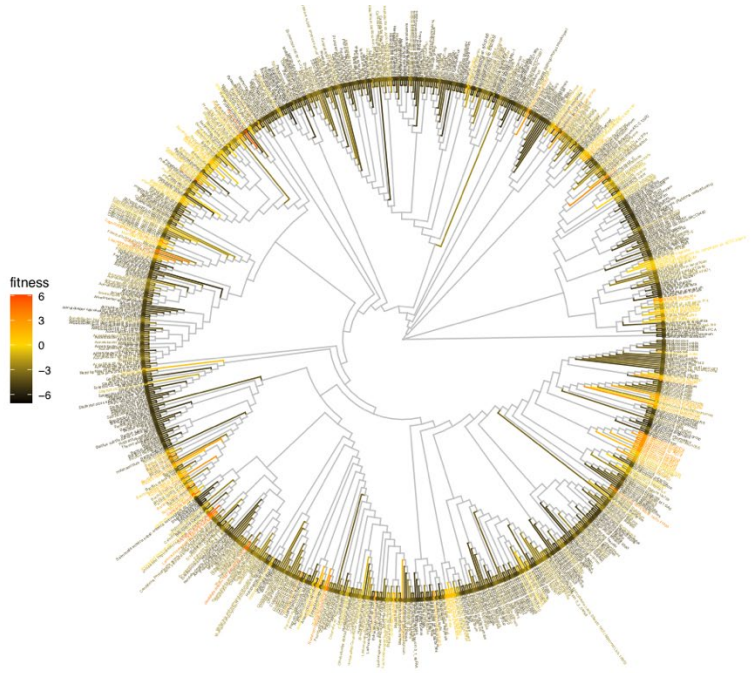
A large fraction of the species present neutral or increased growth rates, indicating complementation with *E. coli*. The population which complements with the host organism can then be tested for antibiotic resistance. There appears to be a group which sees a depletion and can be considered non-complementing, meaning the specific gene sequence does not produce a correct amino acid sequence for successful protein function and therefore is not conducive to survival in *E. coli*. The ratio of surviving species is later used as a baseline of comparison to resistant species in Equation 1.

Antibiotic Resistance

From the selection of homologs and mutants which showed to allow *E. coli* to rescue function, we then assayed the ability to rescue function when exposed to trimethoprim. The surviving population was subjected to a gradient of trimethoprim ranging from 0.058 μg/mL TMP to 200 μg/mL based on the minimum inhibitory concentration of 0.5 μg/mL identified in previous studies[21]. Figure 9 shows the relative homolog fitness in each condition. There is a clear depletion of a majority of the homologs (seen in the dark grey regions on Figure 12), however a number of homologs are seen to be enriched. These homologs are further assessed for sequence-function correlation in Figure 15 below.

While a majority of the homologs are depleted as presented by a large population shift to negative fitness (figure 12b), a few counts are present at a positive fitness. This figure presents a higher average fitness (>3) than observed in the complementing data, likely due to the outgrowth of certain homologs in the absence of the non-resistant species.

(a)



(b)

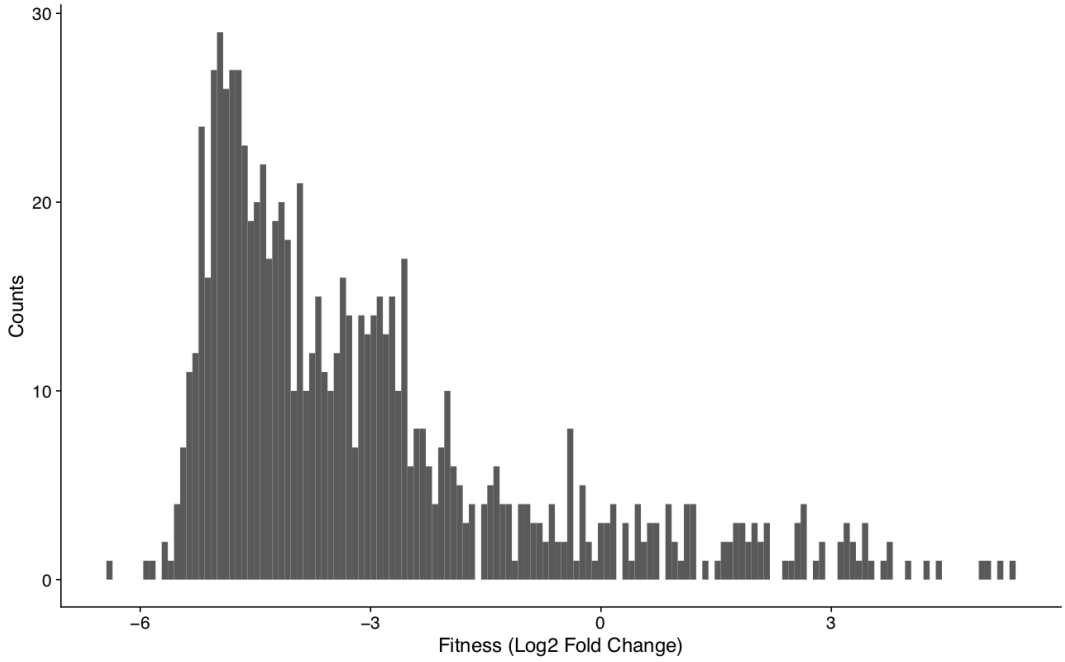


Figure 12: Antibiotic Resistance of DHFR homologs at high trimethoprim concentration (200 µg/mL TMP)

(a) Relative fitness of DHFR homologs under high trimethoprim supplementation of 200 µg/mL.

(b) Histogram of overall fitness for barcode counts present at 200 µg/mL of trimethoprim.

The top 10 resistant homologs with the most barcodes present in the conditions are listed in the table below with their fitness scores. It is likely that the lack of an outgrowth time and the simultaneous depletion of certain homologs during the enrichment of others allowed some to outcompete other homologs.

Table 1: Top 19 homologs in 200µg/mL TMP selection and fitness

NCBI Reference Sequence	Fitness Score	NCBI Reference Sequence	Fitness Score
WP_007654866	5.33309684	WP_009746425	3.50680983
WP_010075211	5.1799017	WP_000637214	3.4780391
NP_775043	5.06145021	WP_003686654	3.4329372
WP_008578924	4.96505343	WP_002897636	3.42965338
WP_007155760	4.40486482	WP_009778768	3.32188267
WP_008990832	4.2357611	WP_002205327	3.2872193
WP_008979999	4.00494272	WP_009660318	3.24142346
NP_267306	3.79203852	WP_008667771	3.23273236
WP_007631135	3.755598	WP_009248943	3.22916732
WP_002930343	3.67266035		

Mutational Landscape of DHFR

The mutational landscape for the DHFR gene was plotted against the wild-type *E. coli* DHFR sequence. The heatmap shown in Figure 13 uses a color scale to indicate relative fitness, which corresponds to the color scheme seen on the 3D models in Figure 14 and 16. These maps provide a landscape to visualize the sequence-space that is inherited by the gene. The correlation between sequence, identity and function is vital to understanding the mechanisms in which DHFR acts. Many factors on this map are important for an understanding of the ideal amino acid sequence. The map displays the median fitness for all homologs with a specific amino acid at a location along the protein sequence.

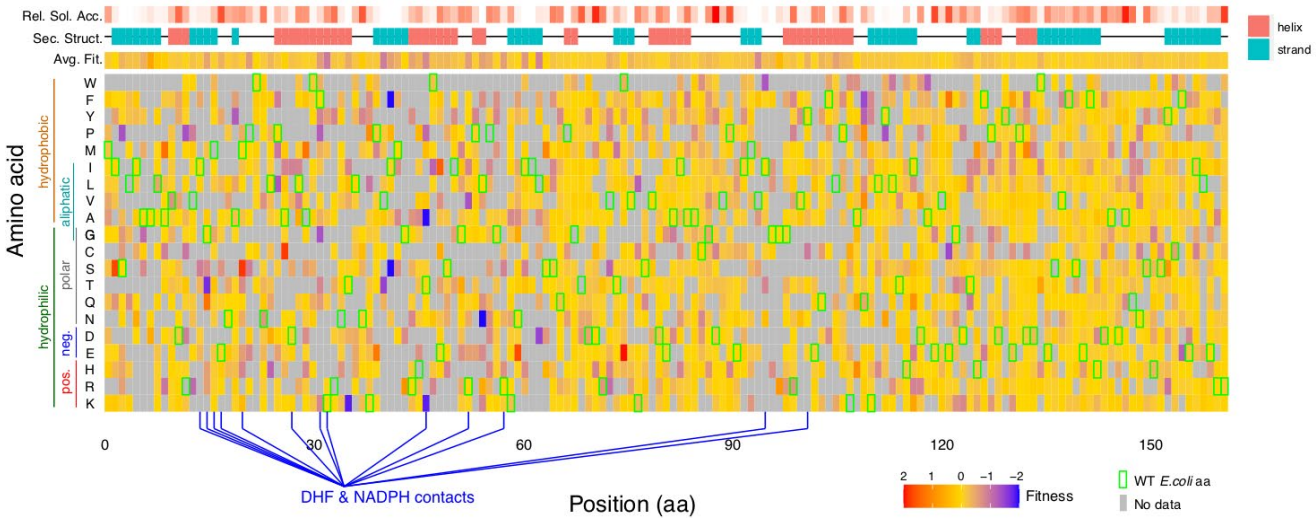


Figure 13: Complementation Heatmap (no antibiotic selection)

This map uses data from 996 homologs and 22,483 mutants all with up to 5 mutations away from the parent homolog, collapsed onto the WT *E. coli* sequence (identified with a green highlight). The map compares the barcodes present between fully supplemented and non-supplemented conditions prior to the addition of trimethoprim.

Figure 13 shows the complementation of homologs to the *E. coli* host. The plot connects directly to the data presented in Figure 11. The sequences for many of the complementing homologs identified in Figure 11 are aligned with the wild-type *E. coli* DHFR sequence position by using an MSA (multiple sequence alignment)[24]. The median fitness for each point on the map was then determined by finding the median fitness for all homologs and mutants with that amino acid at that position. Confidence is improved by only selecting barcodes which appear at least five times in the assay.

The “average fitness” line in Figure 13 shows that there are a few key positions in DHFR which cause higher depletion than others as indicated by the purple bars. This tells us that mutations on those regions are less tolerated by the enzyme and generate lower survival rates in *E. coli*. Contrary to this, there are a few locations which display an enrichment in survival rates with mutations around the positions 7 and 100 on the amino acid sequence, mutations in these

regions appear to facilitate fitness enrichment. This indicates that certain mutations to the sequence in these regions show to increase the ability of DHFR to complement function in *E. coli* under the conditions presented in this study.

Figure 14 illustrates the average fitness for each amino acid position seen in figure 13. This plot was generated by overlaying the average fitness over the molecule with a blue-red color palette using ChimeraX-1.5[25]. Here we see a high tolerance for mutation on the surface exposed regions, and less on the internal residues.

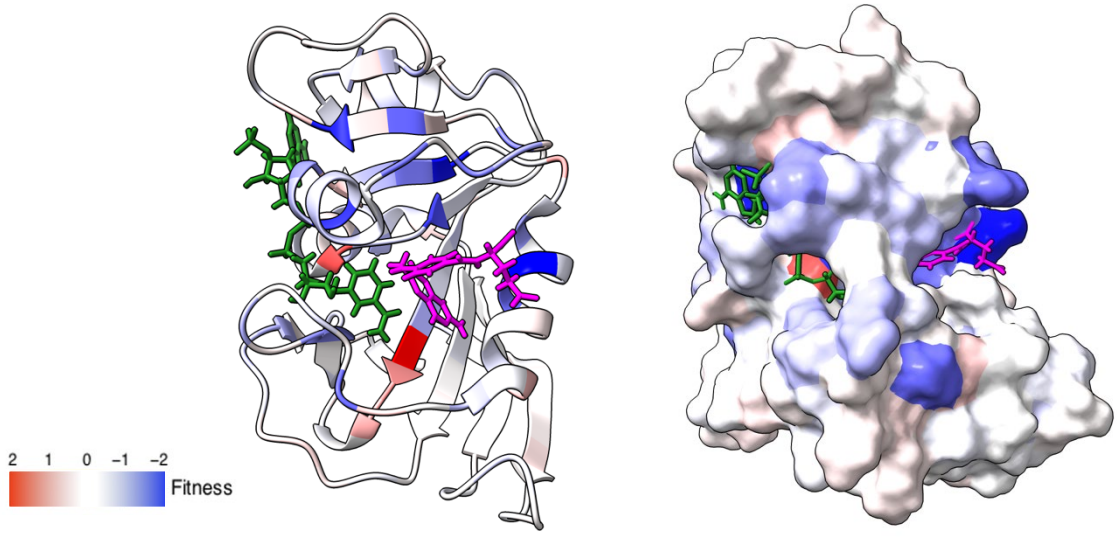


Figure 14: 3D representation of DHFR AA positioning for complementation in *E. coli*. The molecule in green is the cofactor NADPH, and the pink

In Figure 15 we present a map which employs a similar strategy as explained for Figure 13 and assesses the average fitness for homologs that are seen to resist trimethoprim at a concentration of 50 µg/mL, in which surviving homologs are predicted to be resistant to antibiotics.

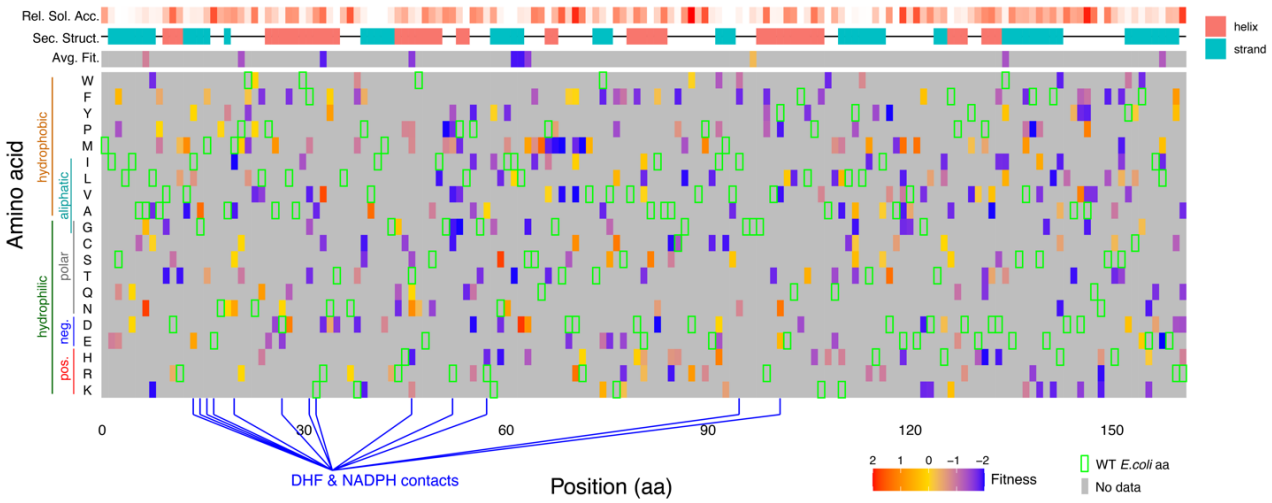


Figure 15: Resistance Heatmap at 50 µg/mL Trimethoprim Inhibition

This map assays sequences present at 50 µg/mL TMP. Note – there is no data for the wild-type sequence because it is inhibited at this concentration.

As seen in Figure 15, trimethoprim inhibition has eradicated a large majority of the homologs present in non-supplemented media. There appears to be a correlation between non-tolerated mutations and contact regions. There are a few mutations of interest, seen as red or orange in the map, which allow the enzyme to completely resist trimethoprim inhibition. Many of these are seen to have a large shift in the type of amino acid, for example, near position 30, a shift from a polar amino acid (arginine and cysteine) to valine, a hydrophobic amino acid, causes an increase in fitness. While this doesn't tell us the specific function of that position, it does indicate that some hydrophobicity is important in that position for the enzyme to resist antibiotic inhibition.

Another application for this data is to overlay the data presented in the heatmaps over the structure of DHFR. Figure 13 is mapped onto DHFR in Figure 14 above, and similarly, figure 15 is plotted on the wildtype *E. coli* DHFR in figure 16. This shows us that regions in blue, indicating a negative fitness or depletion of the homolog during selection, are focused on the external regions of the enzyme.

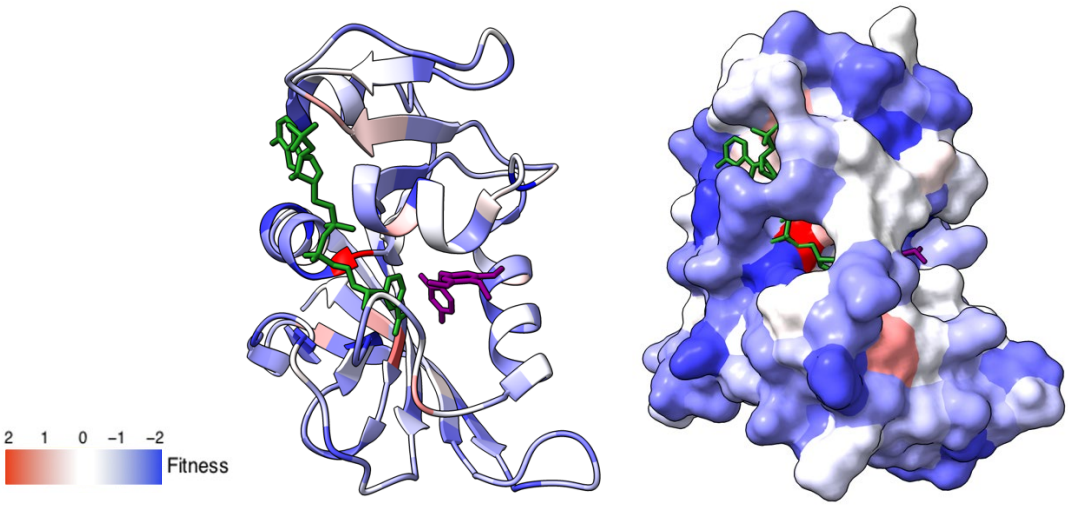


Figure 16: 3D Representation of DHFR AA Positioning for Antibiotic Resistance in *E. coli*. The cofactor NADPH is in green, and the antibiotic trimethoprim is colored purple.

The map is too sparsely populated to draw any positive conclusions, however, a few points suggest some potential areas of interest. Most prominent are a few positions within the active site of the protein to where DHF binds, which are seen to increase fitness against trimethoprim inhibition by the red coloration. This data needs to be compared to the complementation plot to observe positions which show an increase in fitness when compared to the enzymes ability to confer survival without supplementation. Further data and comparisons are required to derive specific resistance-conferring mutations.

Conclusion

Through this study we have designed a multiplexed platform to study large and multiplexed libraries of Dihydrofolate Reductase through a single high-throughput process. The data collected here represents the largest study on the DHFR enzyme to date, totaling data for more than half a million unique sequences. The data is collected in a fashion which allows us to assay the correlation between sequence and function, first for the ability of a homolog to complement function in *E. coli*, and second to survive against trimethoprim inhibition.

This study includes fitness data for sequences that are able to entirely resist antibiotic selection, as well as ones which are only partly inhibited but still manage to survive seen in the selection at low levels of trimethoprim repression. Future studies may observe in more depth the full extent to which some of these selected mutants and homologs are resistant to antibiotic treatments.

We are currently in the process of conducting a replication of this study in parallel with a second library to demonstrate the reliability of the process described in this work. The second library contains a replica of the library assayed here, but with different codon structure. The assay is conducted across multiple timepoints to expand the full scope of inhibition and survival which was limited by the single time point used described above. Future work will aim to expand this study to a larger library with 12,000 unique homologs, which is currently in production.

Additional goals include expansion to another species such as yeast, which will increase the breadth of complementation for some of the homologs which are unable to function in *E. coli*. Additionally, it may be useful to apply this process against different inhibitors or expand the study as a method of testing resistance to new antibiotics in the market.

Supplementary Materials

gBlock sequence from IDT

TGACCAGGTCTCGCTGGAGCGGCGGTAAGGTACCTAAGTGTGGCTGCGGAAC
GCACGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGTCATGAAGAGCCAGG
CGTCGACAAGCTTGCGGCCGCATAATGCTTAAGTCGAACAGAAAGTAATCGT
ATTGTACATCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGAT
ACTGAGCACATCAGCAGGACGCACTGACCGAATTCATTAAAGAGGAGAAAG
GGACCACATGGCAGATCTCATGAAACAGTACCTGGAGCTGATGCAAAAAGTT
CTGGATGAGGGGACCCAGAAAAACGACCGCACGGGGACCGGAACGCTGAGC
ATTTTTGGCCATCAGATGCGCTTTAACCTGCAGGACGGATTCCCGCTGGTTAC
GACCAAACGCTGCCACCTGCGTAGCATTATTCATGAGCTGCTGTGGTTTCTGC
AAGGTGACACTAACATCGCGTATCTGCACGAAAACAATGTGACGATCTGGGA
TGAATGGGCCGATGAAAACGGCGATCTGGGCCCAGTGTATGGTAAACAGTGG
CGCGCCTGGCCAACGCCAGATGGCCGCCACATCGATCAGATCACGACCGTGC
TGAATCAACTGAAAAACGACCCGGACAGCCGCCGCATTATTGTTTCCGCGTG
GAATGTGGGTGAACTGGATAAAATGGCGCTGGCGCCGTGCCATGCATTCTTC
CAGTTTTACGTTGCGGACGGTAAGCTGAGCTGTCAACTTTATCAGCGCAGCTG
CGATGTTTTCCTCGGCCTGCCGTTCAATATCGCCAGCTACGCGTTACTGGTGC
ACATGATGGCGCAGCAGTGCGACCTGGAGGTGGGTGATTTTGTCTGGACCGG
CGGTGATACCCACCTGTACAGCAACCACATGGACCAAACGCATCTGCAACGA
GACCTCGTCA

Primers

Table S1: Primer for plasmid preparation

Primer Pair Name	Forward Sequence (5'→3')	Reverse Sequence (5'→3')
pCVR205_SDM	tatgatcagtctgattgcggcgttagcggtagatcgcgttaTCG GCATGGAAAACGCCA	tgatatctcattattaaagttaaacaaaattatttctacaggG GAATTGTTATCCGCTCACAATTC
SMT205_bb	TGACCAGGTCTCGGCAATTGAGCCGTG AGCCG	TGACCAGGTCTCGCCAGAATCTCA AAGCAATAGCTGTGA
mi3_R1	GTGGAATTGTGAGCGGATAACAATTTC ACACAGGAAACAGCTCATATG	
EV4_BspQI		cagtctaGCTCTTCatgaCGCACATTTCCC CGAAAAGTGCCACCTGACGTCg

Table S2: iTag Illumina Prep Primers

Primer Pair Name	Forward Sequence (5'→3')	Reverse Sequence (5'→3')
CVR205stubBC1_FWD	GCTCTTCCGATCTNGGTACCtaaGTGTGGCTGCGGAAC	GCTCTTCCGATCTNGTGCCACCTGACGTCgtgc
CVR205stubBC2_FWD	GCTCTTCCGATCTNNGGTACCtaaGTGTGGCTGCGGAAC	GCTCTTCCGATCTNNGTGCCACCTGACGTCgtgc
CVR205stubBC3_FWD	GCTCTTCCGATCTNNNGGTACCtaaGTGTGGCTGCGGAAC	GCTCTTCCGATCTNNNGTGCCACCTGACGTCgtgc
CVR205stubBC4_FWD	GCTCTTCCGATCTNNNNGGTACCtaaGTGTGGCTGCGGAAC	GCTCTTCCGATCTNNNNGTGCCACCTGACGTCgtgc
CVR205stubBC5_FWD	GCTCTTCCGATCTNNNNNGGTACCtaaGTGTGGCTGCGGAAC	GCTCTTCCGATCTNNNNNGTGCCACCTGACGTCgtgc

Protocols

PCR Protocol:

Q5 2X master mix from NEB was used in a 50 uL reaction with 0.5uM of primers and 1 ng of DNA for all reactions described. Applied biosystems ProFlex PCR System was set to cycle the temperature to 98˚C for 30 seconds, the following cycle was repeated 30 times unless otherwise stated- 98˚C for 10 seconds, X˚C anneal for 20 seconds, 72˚C for X seconds - followed by a 72˚C final extension time. All reactions were held between 4-12˚ C until removed from the machine. The reaction is followed by a DpnI digest of 0.5 uL enzyme

Electroporation competency protocol:

Overnight cultures from the two colonies were started from 5 mL of LB and 5 µL of thymidine. 50 mL of LB broth was inoculated with 500 µL of the overnight (1:100 dilution) cultures and grown shaking at 37˚C and 250 rpm to an OD600 of 0.6. Cells were chilled on ice for 20 minutes in an ice-water bath. Cells were transferred to 50 mL falcon tubes and centrifuged at 4000Xg for 15 minutes at 4˚C. Supernatant was removed and cells and pellet was resuspended in

50 mL of cold 10% glycerol and centrifuged at 4000Xg for 15 minutes at 4˚C. Supernatant was discarded and pellet was resuspended in 25 mL of cold 10% glycerol. This process was repeated again with 10 mL of glycerol and finally with 2 mL of 10% glycerol. Cells were aliquoted in 100 µL samples and stored at -80˚C.

Optimizing library efficiency and diversity

1. We aimed to have 1,000,000 CFUs represented from our transformations after ligating the backbone (pCVR2) with the library.
2. Start with approximately 10,000 ng for backbone and library product, amplified with PCR.
3. Used biotinylated primers and rSAP addition on the backbone to isolate product from digest reactions and prevent self-ligation.
4. T4 ligase buffer was supplemented with 1:1 ratio of concentrated ATP.
5. Drop dialysis was used clean DNA between the digest and ligation steps and before transformation.
6. 4 transformations were conducted with about 160 ng of DNA each and combined before plating.
7. Cells were recovered by scraping the plates and extracting the plasmid.

Bibliography

1. CDC. Innovative Projects to Slow Antibiotic Resistance. *Centers for Disease Control and Prevention* https://www.cdc.gov/drugresistance/solutions-initiative/innovations-to-slow-ar/projects.html (2022).
2. Recent NIAID Research Initiatives on Antimicrobial Resistance | NIH: National Institute of Allergy and Infectious Diseases. https://www.niaid.nih.gov/research/recent-initiatives-antimicrobial-resistance.
3. Rodrigues, J. V. & Shakhnovich, E. I. Adaptation to mutational inactivation of an essential gene converges to an accessible suboptimal fitness peak. *eLife* **8**, e50509 (2019).
4. High-order epistasis in catalytic power of dihydrofolate reductase gives rise to a rugged fitness landscape in the presence of trimethoprim selection | bioRxiv. https://www.biorxiv.org/content/10.1101/398065v1.full.
5. Sidore, A. M., Plesa, C., Samson, J. A., Lubock, N. B. & Kosuri, S. DropSynth 2.0: High-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res.* **48**, E95 (2020).
6. Bruinsma, N. Influence of population density on antibiotic resistance. *J. Antimicrob. Chemother.* **51**, 385–390 (2003).
7. CDC. Antibiotic Resistance Threatens Everyone. *Centers for Disease Control and Prevention* https://www.cdc.gov/drugresistance/index.html (2022).
8. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
9. Wróbel, A., Arciszewska, K., Maliszewski, D. & Drozdowska, D. Trimethoprim and other nonclassical antifolates an excellent template for searching modifications of dihydrofolate reductase enzyme inhibitors. *J. Antibiot. (Tokyo)* **73**, 5–27 (2020).
10. Aga, D. *et al. Initiatives for Addressing Antimicrobial Resistance in the Environment: Current Situation and Challenges*. (Wellcome Trust, 2018).
11. O'Neill, J. *TACKLING DRUG-RESISTANT INFECTIONS GLOBALLY: FINAL REPORT AND RECOMMENDATIONS*. https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf (2016).
12. CDC. Actions to Fight Antibiotic Resistance. *Centers for Disease Control and Prevention* https://www.cdc.gov/drugresistance/actions-to-fight.html (2022).

13. Zhang, Y., Chowdhury, S., Rodrigues, J. V. & Shakhnovich, E. Development of antibacterial compounds that constrain evolutionary pathways to resistance. *eLife* **10**, (2021).
14. Schober, A. F. *et al.* A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism. *Cell Rep.* **27**, 3359-3370.e7 (2019).
15. Bhattacharyya, S. *et al.* Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity. *eLife* **5**, e20309.
16. Thompson, S., Zhang, Y., Ingle, C., Reynolds, K. A. & Kortemme, T. Altered expression of a quality control protease in e. Coli reshapes the in vivo mutational landscape of a model enzyme. *eLife* **9**, 1–47 (2020).
17. Osborne, M. J., Schnell, J., Benkovic, S. J., Dyson, H. J. & Wright, P. E. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry* **40**, 9846–9859 (2001).
18. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
19. Geck, R. C., Boyle, G., Amorosi, C. J., Fowler, D. M. & Dunham, M. J. Measuring Pharmacogene Variant Function at Scale Using Multiplexed Assays. *Annu. Rev. Pharmacol. Toxicol.* **62**, 531–550 (2022).
20. Lee, J. *et al.* Surface Sites for Engineering Allosteric Control in Proteins. *Science* **322**, 438–442 (2008).
21. Baym, M. *et al.* Spatiotemporal microbial evolution on antibiotic landscapes. *Science* **353**, 1147–1151 (2016).
22. Lambert, T. mCherry at FPbase. *FPbase* https://www.fpbase.org/protein/mcherry/.
23. Bhosle, A. *et al.* A Strategic Target Rescues Trimethoprim Sensitivity in Escherichia coli. *iScience* **23**, 100986 (2020).
24. Clustal Omega < Multiple Sequence Alignment < EMBL-EBI. https://www.ebi.ac.uk/Tools/msa/clustalo/.
25. *Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.*