THE STRUCTURAL DISTRIBUTION OF EPISTASIS IN A PAIR OF ESSENTIAL METABOLIC ENZYMES

APPROVED BY SUPERVISORY COMMITTEE

Kimberly A. Reynolds, Ph.D. (Advisor)

Ryan E. Hibbs, Ph.D. (Chair)

Luke M. Rice, Ph.D.

Hongtao Yu, Ph.D.

DEDICATION

To my mom, Quan Thi Huynh. And my sisters: Lan, Tam, Ann, and Kathy.

THE STRUCTURAL DISTRIBUTION OF EPISTASIS IN A PAIR OF ESSENTIAL METABOLIC ENZYMES

by

THUY NGOC-THI NGUYEN

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2021

THE STRUCTURAL DISTRIBUTION OF EPISTASIS IN A PAIR OF ESSENTIAL METABOLIC ENZYMES

Thuy Ngoc-Thi Nguyen, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2021

Supervising Professor: Kimberly A. Reynolds, Ph.D.

Interactions between proteins provide the basis for cells to perform metabolism, grow, divide, move, and appropriately respond to external stimuli. Because proteins do not act as independent entities, the genetic background influences the effect of a mutation in unexpected ways. This context-dependence of mutational effects is epistasis. Extensive progress has been made in our ability to identify epistasis between proteins. However, how the epistasis between a pair of proteins is distributed across the amino acid sequence is less clear. Previous work characterized this sequence-level epistasis between proteins that bind to form a physical complex. Until now, the structural pattern and magnitude of epistasis between pairs of mutations spanning interacting metabolic enzymes remained uncharacterized.

In my dissertation work, I deeply examined the context dependence of mutations for two essential enzymes in the bacterial folate metabolic pathway, Dihydrofolate Reductase (DHFR) and Thymidylate Synthase (TYMS). To achieve this goal, I used deep mutational scanning assays on DHFR in the context of varying activities of TYMS. The result is a rigorous dataset with epistasis measurements over the entire amino acid sequence of DHFR. I found that the positions with the greatest magnitude of epistasis within the structure of DHFR lied at the active site. However, the sign of epistasis at the DHFR active site was dependent on whether TYMS was active. Beyond the active site, the distribution of positive epistasis among the positions of DHFR was also context-dependent on the state of TYMS. Therefore, we can think of the active site as a non-physical "interface" between protein pairs that do not form a physical complex but share an intermediate.

The potential consequences of this dataset on the epistasis between DHFR and TYMS are profound. This dataset is fundamental towards our understanding of how epistasis mechanistically emerges in nonlinearities between catalytic activity in enzymes, protein abundance, and cellular growth rate. This experimental dataset is also necessary to credibly validate predictions of epistasis from models of statistical co-evolution.

TABLE OF CONTENTS

List of figures

List of Tables

CHAPTER ONE
Introduction

1.1 An introduction to epistasis

Interactions between proteins are the foundation of cellular systems. These interactions include direct binding between proteins in physical complexes, substrate and product sharing between enzymes in metabolic pathways, and/or proteins that post-translationally modify one another. Together, these interactions provide the basis for cells to perform metabolism, grow, divide, move, and appropriately respond to external stimuli. Because proteins do not act as independent entities, the effect of mutating one protein often depends on the mutational status of other interaction partners. That is, genetic background influences the effect of mutation, sometimes in unanticipated ways. These interactions among proteins limit our ability to rationally predict how perturbations (e.g. mutations or changes in expression) affect cellular phenotypes. Understanding the pattern of interactions between proteins in the cell is thus an essential step towards building quantitative models that can predict the phenotypic effects of mutations.

The context dependence of mutational effects is quantified using a measurement termed epistasis. Epistasis is the unexpected effect of combinations of mutations – in essence, how much the effect of making a mutant changes given a change in genetic background. Epistasis has long been used as a tool in classical genetics to identify the order of genes in a pathway.[1] In these experiments, one typically measures how the effect of a deleting a gene depends on the presence (or absence) of a second knockout. For example, two enzymes which perform parallel reactions leading to an essential metabolite might be "synthetically lethal" – deleting each enzyme individually is well-

tolerated, while deleting both enzymes together is lethal. These types of experiments provide a gene-level picture of interactions between a pair of proteins. But how is epistasis distributed at the level of individual amino acid mutations? Intuition suggests that the pattern of epistasis between specific mutations across a pair of proteins will be heterogeneous, with mutations at some positions being highly epistatic, and others not at all. Yet it remains unclear which positions are most strongly coupled between proteins, and how these epistatic positions are arranged on the protein structure. Previous work has sought to address this question for physical protein complexes (see section 1.2). Until now, the structural pattern and magnitude of epistasis between pairs of mutations spanning interacting metabolic enzymes remains uncharacterized. In my dissertation work, I sought to deeply examine the context dependence of mutations for two essential metabolic enzymes, with the goal of understanding the constraints that metabolic interactions place upon protein sequence.

To compute epistasis between a pair of mutants, it is necessary to first identify what kind of epistasis we are measuring. The quantitative definition of epistasis is dependent on the type of null model we choose. Epistasis between mutations in a single protein is often directly related to thermodynamic free energy between residues of a protein.[2] For example, if the experimenter wishes to study the landscape of thermodynamic free binding energy of a ligand-binding protein, they would measure the change in free energy of binding for two single mutants and the double mutant. In this case — because free energy is a state function — the null model is an additive one. Consider mutants of two amino acid positions, A and B within a protein. The epistasis between A and B is:

$\Delta\Delta G_{AB} = \Delta G_A - \Delta G_{A|B}$ (equation 1.1)

ΔG_A is the change in free energy of the mutant A in the context of a wild-type genetic background. $\Delta G_{A|B}$ is this same measurement, but in the background of mutation B. Conceptually, if these two positions are independent and not epistatic to each other, there should be no impact of the state of B on the effect of a mutation in A. The difference between ΔG_A and $\Delta G_{A|B}$ would be zero. If B is epistatic to A, $\Delta\Delta G_{AB}$ will be non-zero. This definition of epistasis holds true in terms of ΔG_B and $\Delta G_{B|A}$.

$\Delta\Delta G_{AB} = \Delta G_B - \Delta G_{B|A}$ (equation 1.2)

After expanding the definition of epistasis in terms of the individual measurements of free energies of each mutant allele (normalized by the wild-type free energy) the additive model of epistasis is:

$\Delta\Delta G_{AB} = G_A + G_B - G_{AB}$ (equation 1.3)

In this form, the epistasis is equivalent to the difference between the free binding energy of the double mutant (G_{AB}) and the sum of the free binding energies of the single mutants (G_A and G_B). This mutants is epistatic if the double mutant deviates from the additive effects of the single mutants. However, additivity need not be the expectation when one considers epistasis in terms of growth rate, rather than free energy. In my thesis, I measured the effect of single and double

mutations on bacterial growth rate. These are well described by a log-additive or multiplicative model of epistasis in equation 1.4.[3]

$$E = G_{AB} - G_A G_B \qquad \text{(equation 1.4)}$$

Here, E is the epistasis and G now represents bacterial growth rate. G_A and G_B are the growth rate effects of the single mutants, A and B. G_{AB} is the growth rate effect of a double mutant. When A and B do not effect each other, the epistasis is zero and the growth rate of the double mutant can be predicted from knowing the growth rates of the single mutants. When epistasis is positive, the double mutant growth rate exceeds the combined single mutant growth rates. In combination, A and B buffers the growth rate defects of either single mutant. When epistasis is negative, the double mutant growth rate is lower than the combined single mutant growth rates. This means that the presence of both mutations at once aggravates the growth. Non-zero epistasis is typically referred to as sign epistasis, a condition where expectation of the effect of multiple mutations deviates from the cumulative phenotypes of the individual single mutants.

These equations make clear why nonlinear interactions between mutations present a challenge for rationally predicting how mutations affect the fitness of an organism. A mutation might be deleterious in one context, yet beneficial or neutral in another. This has profound consequences for interpreting the effects of disease-associated mutations.

For example, the same sequence of alpha-synuclein, a protein that causes neurodegenerative disease, is non-pathogenic in mice but is pathogenic in humans.[4] This variant of alpha-synuclein is epistatic with another protein in the genetic background, imposing a barrier to studying models of disease in mice. This epistasis imposes a hurdle towards developing effective mouse model system of neurodegenerative disease. Moreover, over 30 human genes have pathogenic mutants that exhibit this kind of species-specific epistasis, called Dobzhansky-Mueller Incompatibilities.[4]

Moreover, epistasis shapes the evolutionary trajectory of proteins.[5] Consider sign epistasis - if two mutations are individually deleterious, but beneficial in combination, this creates an evolutionary "valley" that is difficult to cross under conditions of constant selection pressure.

As a starting point to understanding how epistasis shapes the function and evolution of metabolic enzymes, I characterized epistatic interactions between two well-studied essential metabolic enzymes. In the following sections, I first describe prior work on epistasis for the better-studied case of physical protein complexes (for context), then what we know about epistasis in metabolic enzymes, next my model system (the enzymes dihydrofolate reductase and thymidylate synthase), and finally, my approach (deep mutational scanning, or DMS). I end with a brief summary of my results.

1.2 Epistasis in physical complexes

In order to bind, it is necessary for proteins that make up a physical complex to overcome the thermodynamic barrier of de-solvating the interface and losing conformational energy, and come

together to form a stable interface maintained by non-covalent interactions. So what is the structural pattern of interactions that underpin binding? A foundational alanine scanning mutagenesis study in 1999 between the interface of human growth hormone and the human growth hormone binding protein revealed that the interactions within the interface are organized into "hot spots" of free binding energy.[6] These "hot spots" are only a few residues within the interface but are responsible for the greatest contributions to forming the complex. This study illustrated how large-scale mutagenesis of a single protein pair can provide general insight into how amino acid interactions between proteins are distributed.

Advances in next-generation sequencing and binding-based selections has enabled further studies of mutational effects at unprecedented scale. In 2018, Diss and Lehner used saturation mutagenesis to comprehensively study the interaction between Fos and Jun, two proteins that bind together to form a transcription factor. Here, the ability for Fos and Jun to assemble was measured in all possible single and double mutations in their respective leucine-zipper binding domains.[7] In contrast to Clackson and Wells, the authors were able to consider double mutants throughout the protein, not just localized to the binding interface. The epistasis between Fos and Jun were driven by two mechanisms, one due to specific structural and biochemical interactions at the interface, and the other due to a three-parameter thermodynamic model of binding that was agnostic to the location or identity of the mutations.

While Diss and Lehner focused on a single binding interaction, other work has explored epistasis between cognate and non-cognate binding proteins. Limited DMS assays of the interfaces between

a trypsin inhibitor and three homologs mapped the landscapes of free binding energies between a trypsin inhibitor and three trypsin homologs.[8] Though they were only able to measure half of the double mutants in their assay, they observed that the pattern of epistasis in the landscapes of free binding energies differed with each homolog. When the trypsin inhibitor was paired with either of the two non-cognate trypsins, the binding affinities among the mutants broadly suffer, flattening the peak of the mutational landscape. They observed both positive and negative epistasis in key positions at the interface of the trypsin inhibitor in the background of the two non-optimized trypsin homologs. Taken together, this work on three distinct model systems — hGH/hGHbp, Fos-Jun, and trypsin/trypsin inhibitors — shows how deep mutational scans can reveal organizational principles for physically interacting proteins. For my own thesis work, these studies provided inspiration for characterizing the pattern of epistasis between non-binding (but functionally coupled) enzymes.

Importantly, large scale mutagenesis experiments – though informative – are expensive and laborious. An alternative approach to mapping interactions between proteins is the computational analysis of co-evolution.[9–11] The basic premise of this approach is that interactions between residues should result in their correlated evolution (co-evolution) across homologs. In this case, statistical analysis of co-evolution in large and diverse sequence alignments can be used to predict interactions. This appears to be true for physical complexes: co-evolution studies show that epistasis at physical interfaces is reflected in the protein sequence. In perhaps the most cited example of this, the histidine-kinase and response regulator proteins that make up a two-component signal transduction system in bacteria show co-evolution between structurally

localized positions at the physical interface.[12] Pairwise co-evolution was computed from a multiple sequence alignment (MSA) of each component using Mutual Information. These key residues that were necessary and sufficient to identify and generate mutants to re-wire the specificity of the histidine-kinase to phosphorylate non-cognate response regulator proteins. A similar framework was applied to identifying the key positions that dictate selectivity in synthase and receptor proteins used for Quorum sensing, a cell to cell signaling system in bacterial populations.[13] Like in the two component system, substitutions at these co-evolving positions was necessary and sufficient to engineer the synthase to make a non-cognate small molecule and to engineer the receptor to respond to a non-cognate small molecule.

These studies have revealed how the amino acid sequence of a protein encodes physical interactions. The impact of this is clear: methods of sequence co-evolution have been scaled up to both to predict groups of proteins that form physical complexes and to predict which residues lie at the interface on a proteome-wide scale.[14,15] For instance, Cong et al. integrated a series of existing coevolution methods to develop a pipeline that identified protein-protein interactions among over 4000 proteins in the *E. coli* proteome. This pipeline generated MSAs of the orthologous proteins in *E. coli* from the sequences across over 40,000 genomes. The paired MSAs were analyzed with a series of methods that computed sequence co-evolution with local statistics like Mutual Information for local residue to residue interactions and global statistics like Direct Coupling Analysis and GREMLIN which identify the contacts between residues within a protein from computing covariation. Significantly co-evolving protein pairs are then further filtered after

identifying their physical interfaces using a docking method. This analysis pipeline predicted 804 protein-protein interactions and their interfaces, including existing and novel complexes.

However, these studies reveal a bias in the current literature towards considering co-evolution and epistasis at physical interfaces. The convenient thing about developing a method that predicts the proteins that form a physical complex is that the predictions can be validated with existing structural datasets, and large scale interaction screens (e.g. yeast two-hybrid data, or mass spectrometry studies).[16–18] In contrast, we lack gold-standard experimental data for predicting epistasis between non-binding proteins. My thesis work provides a template for gathering these data, and provides a first picture of epistatic interactions between two sequential metabolic enzymes at the residue level. These experimental datasets are necessary to credibly validate predictions of epistasis from models of statistical co-evolution.

1.3 Epistasis due to non-physical interactions: non-linearities between catalytic activity, protein abundance, and cellular growth rate can generate epistasis.

In my thesis work, I will examine the pattern of epistasis between two metabolic enzymes that share an intermediate. So what processes can lead to epistasis between non-binding but functionally coupled proteins? Here, it can be useful to categorize epistasis between proteins as "specific" and "global".[19] Specific epistasis describes the epistasis between proteins that is due to biophysical and molecular interactions in the structure and function of the proteins. Thus, specific epistasis is typically quantified as non-additive effects of mutations on particular biochemical or biophysical parameters, like the dissociation constant of physical complexes (K_d), free energy of

folding (ΔG), or catalytic power (k_{cat}/K_m). In this case, the mathematical definition of epistasis is when the combined effect of mutations deviate from a linear null model (Fig. 1.1). This linear null model is well described in Figure 1.1A. In a given protein, the effect of the single mutants, A and B and the double mutant AB can be measured *in vitro* additive biophysical traits like catalytic activity and *in vivo* phenotypes like cellular growth rate. The null expectation is that the additive effect of the individual mutants are equivalent to the effect of the double mutant. When a pair of mutants do not have proportional effects on both catalytic activity and growth rate, A and B are identified as epistatic to each other. One cause of specific epistasis are the interactions between residues at the interface in a physical complex that we considered in section 1.2.

However, epistasis is not limited to non-additive interactions at the level of individual molecules (and biochemical parameters). When epistasis is considered at the level of growth rate, it becomes clear that epistasis can also emerge from global, non-linear relationships between genotype and phenotype.[19] For example, consider Figure 1.1B. In this case there is a non-linear function that relates some biochemical or biophysical parameter to growth. To illustrate, consider the relationship between enzyme activity and growth, which we expect to saturate – in this case, we might reasonably expect a plateau where further improvements in activity do not yield growth rate enhancements because the enzyme is already "good enough". Under these conditions even if individual mutations have additive effects on the underlying protein's activity, we might observe epistasis at the level of growth rate (Fig. 1.1B). This is exactly how Diss and Lehner interpreted the epistasis between Fos and Jun. They constructed a non-linear thermodynamic model relating binding affinity to fraction bound: this function was able to describe a majority of the double

mutant effects on the ability for Fos and Jun to bind and form a complex. The minority of mutational effects that were not captured by this non-linear function were due to specific, structural mechanisms at the interface. The major hurdle to defining non-specific epistasis is to identify and fit such a function. This is a non-trivial task as it would necessitate *in vitro* measurements of protein function on top of deep mutational scanning assays to collect high-throughput *in vivo* growth rate measurements. However, my work benefits from a long history of *in vitro* work on my model enzyme. In Chapter 4, I discuss the potential to construct a (mathematical) null model that can separate my growth-based epistasis measurements into "global" and "specific" components. In the absence of such a model, I expect my data reflect epistasis due primarily to the non-linear relationship of catalytic activity to growth rate, but may contain potential specific interactions between the enzyme pair. In any case, my data should reveal the pattern of epistasis, regardless of mechanism.

As for physical interactions, functional interactions between non-binding proteins can also drive co-evolution. For example, prior work from our lab showed that synteny (correlated physical proximity in the chromosome) and co-occurrence (correlated presence and absence among species) could identify some metabolic interactions.[20] Of the gene pairs with known function, most form a physical interaction, the rest are in the same metabolic pathway. A small subset of these gene pairs are coupled by a shared metabolite. A constraint on their intermediate could be driving co-evolution between subsequent enzymes in metabolism. If so, co-evolution might present a powerful strategy to map interactions between functionally linked proteins, not just those that physically bind.

1.4 The model system: A pair of enzymes in bacterial folate metabolism

To study epistasis between metabolic enzymes, I selected a pair of enzymes in folate metabolism – Dihydrofolate Reductase (DHFR) and Thymidylate Synthase (TYMS). These two enzymes catalyze sequential reactions, wherein the product of TYMS is the substrate for DHFR. Folate metabolism, also known as one-carbon metabolism, is vital source of amino acids and nucleotides like methionine, serine, glycine, thymidine, and purine biosynthesis (Fig. 1.2A).[21] The reduced folate species in this pathway have a shared chemical backbone and function, which is to carry and transfer these one-carbon units in the synthesis of these key building blocks in the cell. [21] Thus, we can expect that mutations which disrupt DHFR and TYMS function should effect growth rate in a measurable way. Importantly, prior work from the Reynolds lab indicates that DHFR and TYMS co-evolve strongly with one another, but co-evolve very little with the rest of folate metabolism (and indeed the genome).[20] Analyses of both gene synteny and gene co-occurrence suggest that these two enzymes form a modular unit, and this was corroborated through epistasis measurements for select mutants and forward evolution experiments.[20] Thus, DHFR and TYMS represent a simplified two-enzyme system for my studies, in which I expect epistasis between the two enzymes but less epistasis to the surrounding system.

DHFR is the only enzyme in folate metabolism that catalyzes a reduction of folate (from DHF to THF), while TYMS is the only enzyme that catalyzes the corresponding oxidation (from THF to DHF). As a consequence, point mutations of DHFR that slow down catalytic activity led to: (1) an accumulation of DHF, the shared intermediate between DHFR and TYMS and (2) a depletion

of downstream reduced folate species that drive amino acid and nucleotide synthesis. DHF accumulation likely inhibited the poly-glutamation reaction on reduced folates, and lowered their abundance in the cell.[22] These imbalances in folate metabolites were partially rescued when the DHFR mutant was paired with a catalytically inactive TYMS. The mechanism underlying the epistasis between DHFR and TYMS are constraints on folate metabolite abundances like the toxic effect of accumulation DHF and need to retain the pool of downstream reduced folates in nucleotide and amino acid synthesis. Based on these observations, I expect to see that mutations in DHFR are often buffered by loss of function mutations in TYMS, giving rise to positive epistasis.

Beyond the modular nature of epistasis between the DHFR/TYMS pair, an additional benefit of using both DHFR and TYMS as model system is that they are well-studied enzymes. We know their structures, function, and conformational dynamics. Let us first review basic biochemical information about these two enzymes: their role in the folate metabolic pathway, the basis for biochemical coupling, their structural biology, and effect on growth rate.

1.4.1 The structural biology and biochemistry of DHFR and TYMS

DHFR, per its name, catalyzes the reduction of the folate, dihydrofolate (DHF) to tetrahydrafolate (THF). Its cofactor, NADPH, acts as a hydride donor to DHF. This small, approximately 19 kDA, enzyme functions as a monomer (Fig. 1.2B). The structure of *E. coli* DHFR cycles through a series of distinct and dramatic conformational changes during the redox reaction.[23,24] In brief, key loops at the active site of DHFR facilitate the binding and release of substrate and cofactor throughout

the cycle. This includes the Met20 loop, which covers the opening of the active site and packs against NADPH in the closed conformation. Hydrogen bonds between pairs of residues in the F-G loop and the Met20 loop stabilize the closed conformation. In the closed conformation, the hydride transfer between NADPH and DHF forms $NADP^+$ and THF. After this, DHFR assumes the occluded conformation where the Met20 loop moves to open up the active site to release $NADP^+$. The Met20 loop in the occluded conformation is stabilized by hydrogen bonds with the G-H loop. These conformational changes along the catalytic cycle of DHFR has been a useful model system for understanding how conformational changes physically drive the catalysis of biochemical reactions in enzymes.

TYMS, per its name, synthesizes thymidine by transferring a carbon unit from the reduced folate, 5,10-methylene THF, to dUMP [25] The products of this reaction are dTMP, a nucleotide essential for DNA synthesis, and DHF, the substrate of DHFR. Functional *E. coli* TYMS is a homodimer (Fig. 1.2C). The active site is formed at the interface of the two homodimers.[26] In the active site itself, four arginine residues directly bind the phosphate group of dUMP. Two arginine residues are in one homodimer (R21 and R166) and the other two are in the opposite homodimer (R126 and R127). After TYMS binds substrate and cofactor, catalysis is initiated by a nucleophilic attack on dUMP by the thiol group in position C146.[25] R166 forms hydrogen bonds with the phosphate group of dUMP and the thiol group of C146.[27]

1.4.2 DHFR and TYMS activity is directly linked to growth

As essential enzymes, perturbations in the catalytic activity of DHFR and TYMS can be directly detected from changes in growth rate.[28,20] This is clearly observed in the relationship between *in vitro* enzyme activity and bacterial growth rate. In 2011, Reynolds et al. generated a series of point mutations in DHFR and measured their respective *in vitro* Michaelis-Menten steady-state enzyme kinetic parameters. The catalytic activities (k_{cat}/K_m) of these mutants spanned over 5 orders of magnitude. These changes in enzymatic activity are directly detectable from their effect on bacterial growth rate (Fig. 2.3). This relationship is monotonic, where lower DHFR activity decreased growth. These DHFR point mutants make up what we call the "Calibration Curve". For TYMS, the analogous Calibration Curve is limited to three variants: WT, Q33S, and R166Q. Because my thesis work studies these three variants in detail, below I will further discuss the biochemical roles of Q33S and R166Q.

As mentioned earlier, R166 is one of the four arginine residues that coordinate the phosphate group of dUMP. As such, it is perhaps unsurprising that the R166Q mutation is fully detrimental to TYMS function. Catalytic activity of R166Q TYMS is not detectable. In contrast, mutations at the other three arginine residues that contact dUMP result in reduced TYMS activity, but not a total loss of function. A crystal structure of *E. coli* R166Q TYMS provides a structural basis for why mutations at this position are intolerable for TYMS catalytic function.[27] The structures of R166Q TYMS and WT TYMS are highly similar to each other, with a RMSD = 0.24 Å. The major difference at the active site is the orientation of C146. In the mutant structure, C146 does not form hydrogen bonds with R166Q and the side chain shifted closer into the nucleotide-binding site. This change in the active site likely decreases binding affinity for dUMP and prevents the precise

orientation the thiol group in C146 for catalysis of the methylation reaction. This mutation in TYMS directly affects growth. *E. coli* with the R166Q mutation in TYMS are auxotrophic for thymidine and will not grow unless their growth media is supplemented with thymidine. When supplemented with thymidine, *E. coli* with this mutation grow as well as WT.

Position Q33 in TYMS lies at the interface between the two TYMS homodimers and has no known role in catalysis. A limited saturation mutagenesis assay on 25 positions within and around the active site showed that all mutations to Q33 had growth rate effects that ranged from 50-100% of the WT. Out of the TYMS mutants from this saturation mutagenesis study with "moderate" growth rate effects, we were able to successfully protein purify Q33S TYMS and measure steady-state enzyme kinetics (see section 1.9). Overall, Q33S TYMS was only slightly slower than WT TYMS at synthesizing thymidine (Table 1.1, Table 1.2). The V_{max} for Q33S TYMS is significantly lower than the WT. The Michaelis Constants (K_m) for both dUMP and 5,10-methylene THF (MTHF) were not significantly affected by the mutation. The catalytic activities (kcat/Km) are also statistically similar between WT and Q33S. The growth rate *E. coli* with Q33S mutation in TYMS is similarly unaffected, however we will later see that Q33S is strongly epistatic with many mutations in DHFR. Taken together, DHFR and TYMS are an epistatically linked protein pair, in which growth rate is strongly linked to enzyme function. I am equipped with *in vitro* steady state Michaelis-Menten measurements for a number of DHFR and TYMS point mutants, which is necessary to characterize the behavior of my selection assay (see Chapter 2).

1.5 Deep mutational scanning

Given DHFR and TYMS as a model system, I decided to measure epistasis via a deep mutational scan (DMS). DMS assays provide a very high-throughput way for investigating both *in vitro* and *in vivo* functional properties of a protein of interest, given a well-designed assay. All DMS experiments start with a saturation mutagenesis library which contain all possible single mutations at every single amino acid position in the protein sequence. The library then undergoes a selection assay that targets the function of the protein. This assay must report on the mutational effects of the function of the protein of interest. The mutants that survive the selection step have greater fitness, or are more functional, relative to the remaining mutants in the library. The mutants that struggle to survive the selection step are less functional and less fit. The issue with using DMS as a tool is that developing such an assay is a non-trivial task that requires a significant time investment – often over a year. The assay must report on a physiologically relevant parameter that can be measured at high-throughput, typically this is measurement based on bacterial growth but can vary depending on the protein of interest. For instance, bacterial growth rate can act as a proxy for thermodynamic free binding energy, protein stability, etc.[29] Once such an assay is well-developed, the saturation mutagenesis library can then be generated and go through a selection step. At a minimum, the library is sampled at the before and after selection, but it is also possible to take a series of time points during selection to resolve small changes in growth rate. Each sample is then deep-sequenced using Next-Generation Sequencing (NGS) technology to count the frequencies of each mutant in the population. These frequencies can be used to estimate the relative growth rates of individual mutants, which can be connected back to the parameter of interest (e.g. catalytic activity).

With the resulting data, the experimenter must overcome the challenge of analysis and interpretation of these large-scale measurements. Currently, there is the option of processing these data with software package that processes and analyzes generic NGS data from DMS experiments, *Enrich2*.[30] Note that in our work, we process, analyze, fit, and visualize all NGS data in customized software written in python3. This process of selection and deep sequencing to estimate the population of alleles in the library can resolve differences as small as approximately 2% differences in growth (as computed from F31Y/WT and F31Y/Q33S in the calibration curve, see Chapter 2).

In the past two decades, DMS has been a useful tool to examine the sequence constraints within a single protein. A DMS study on the small protein domain, PDZ, revealed that the pattern of positions in the structure that were crucial to its function were consistent with the pattern of the most strongly co-evolving amino acid positions in a model of sequence co-evolution.[31] DMS experiments on the TEM beta-lactamase, the enzyme that causes ampicillin resistance in bacteria, and Hsp90, a heat-shock chaperone protein, showed that the pattern of mutational effects is strongly dependent on selection pressure by the cellular environment and protein expression levels.[32] In *E. coli,* a key enzyme in protein homeostasis is Lon protease, an enzyme that identifies and degrades damaged proteins. In laboratory strains of *E. coli*, this protease is absent. In Thompson et al. DMS of *E. coli* DHFR in the presence and absence of Lon protease shapes the distribution of fitness effects. In this study, a greater proportion of mutations are advantageous to *E. coli* growth in the absence of Lon protease. When Lon protease is present, the *E. coli* growth rate is more sensitive to mutations in DHFR that may have lowered protein stability but retained

catalytic activity.[33] Under conditions with high selective pressure, these proteins were much more sensitive to mutations, leading to a change in the shape of the distribution of fitness effects (bacterial growth rate) where the number deleterious mutations increased and mutations with neutral and beneficial fitness effects decreased.

These numerous DMS experiments have been successful in studying protein structure and function within individual proteins. Consistently, across different model systems, these works find that proteins are fairly robust and are tolerant to most mutations. These mutations that do affect function in a DMS are rarer, are usually deleterious, and tend to localize at positions that are key to the structure and function of the protein. Across a range of conditions, the distribution of mutational effects is not static, but dynamic according to the selective pressure imposed on the protein.[32–34]

Though DMS has been a useful tool for studying individual proteins, only one study has used this method to study epistasis *across proteins.* In this work, the model system, Fos and Jun, are two alpha-helical proteins interact through a leucine zipper to form a transcription factor. In this study, Diss and Lehner generated all possible single and double mutants in leucine zipper domains of Fos and Jun and measured their ability to form the transcription factor.[7] The basis of their assay is the complementation of a methotrexate resistant DHFR fragmented across Fos and Jun. When Fos and Jun interacted, the DHFR fragments became active, and enabled the yeast cells to grow in the presence of methotrexate. The library was sampled and deep sequenced before and after selection with methotrexate. A protein-protein interaction (PPI) score was computed from the WT-normalized frequencies of each mutant in the population. A thermodynamic model was able

predict a majority of the PPI scores of the double mutants from the effects of the single mutants. The epistasis in the other mutants that were not predicted by the thermodynamic model were due to specific, structural interactions that tended to localize at the interface between Fos and Jun. Overall, this study was the first time epistasis between different proteins was comprehensively measured at the amino acid sequence level. In my dissertation, I use DMS as a tool to study how epistasis to TYMS is distributed in the full sequence and structure of DHFR. I measured the effects of all the mutants in a saturation mutagenesis library of DHFR in the context of a fully functional WT TYMS, the moderately active Q33S TYMS mutant, and the fully inactive R166Q TYMS mutant. This work will be the first time DMS is used to assess epistasis in the amino acid sequences across proteins that functionally interact through a shared metabolite.

We know that in the context of the cell that proteins do not operate in isolation and that epistatic interactions impose an evolutionary constraint on protein sequences. Despite this, application of DMS to study function in more than one protein has been rare. My DMS dataset on the epistasis between DHFR and TYMS will be a major contribution in this technical field.

1.6 A summary of results

As described in the previous section, the landscape of mutational effects in a DMS is dependent on the conditions of the experiment. This process is so important that I dedicate the entirety of Chapter 2 to describing how I chose the final conditions of the selection step of the experiment. A DMS of DHFR has already been done by Thompson et al. in very specific conditions that increased the sensitivity and resolution between near-WT and advantageous mutations. In this regime,

Thompson et al. was able to detect how changes in protein homeostasis modulate the landscape of fitness effects in DHFR. In this project, my goal was to measure epistasis between DHFR and TYMS. Therefore the conditions I ended up choosing were different than those in Thompson et al. DHFR and TYMS were expressed from a different plasmid backbone and at theoretically higher levels. I chose to supplement the media conditions of the assay with both amino acids and thymidine. This relieved selective pressure on both TYMS and folate metabolic enzymes downstream from DHFR. Under these environmental conditions and genetic background, I was able to collect high quality data and measure epistasis in DHFR and R166Q TYMS.

I performed the selection of the DHFR saturation mutagenesis library under these conditions in the background of each TYMS variant, in triplicate. The resulting data is a comprehensive map of fitness effects at every position and every possible single amino acid mutation in the sequence of DHFR in the context of each TYMS variant. The distribution of DHFR fitness effects shifts with TYMS activity. DHFR is most sensitive to mutations in the context of a fully functional TYMS and the least sensitive to mutations when TYMS is inactive. We recapitulated the positive epistasis between DHFR and R166Q TYMS in the DHFR Calibration Curve.

After epistasis was computed for each DHFR mutant in the background of each TYMS mutant, the DHFR positions were categorized into four groups using a K-means clustering algorithm. These categories are: no epistasis, negative epistasis, positive epistasis, and super positive epistasis. In both TYMS mutant backgrounds, the structural distribution of epistasis among these categories are very different. In the R166Q TYMS background, strong positive epistasis is

primarily in the DHFR active site. Positions with positive epistasis surround the active core in a shell-like configuration. There were no positions with negative epistasis. In the Q33S TYMS background, negative epistasis is localized within the active site and in the lower active site domain of DHFR. Positive epistasis was distinctly localized in the upper, adenosine binding domain of DHFR. The strongest signal of epistasis was primarily localized in the active site of DHFR. This result aligns with the idea that the rate of DHFR catalysis is a major constraint that drives the interaction between DHFR and TYMS. Also we observed that the sign of epistasis at the active site was not the same in each TYMS background. Here, DHFR active site positions were negatively epistatic to Q33S TYMS and strongly positively epistatic to R166Q TYMS. This observation, that in this type of functional interaction between two subsequent metabolic enzymes, the pattern of epistasis across in the amino acid sequence in one enzyme is not static and is dependent on the activity of TYMS.

1.7 Figures

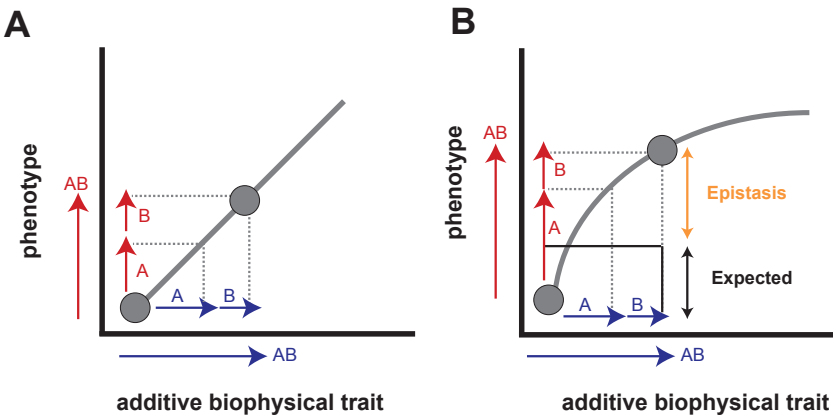

Figure 1.1 A visual description of non-specific epistasis (adapted from Domingo et al.)[19] The relationship between a generic additive biophysical trait (x-axis) and a generic measurable phenotype (y-axis) in two different models of epistasis. **(A)** The solid grey line is the linear relationship between the additive biophysical trait and the phenotype. The arrows in blue show the magnitude of the effect of the single mutations in A and B on the biophysical trait. The longer blue and red arrows are the effects of the single mutants summed together. The grey marker at the origin is the WT. The grey marker at coordinates [AB,AB] is the effect of the double mutation. The mutational effects on the protein are directly proportional to their effects on the observed phenotype in the organism (grey dashed lines). **(B)** The grey line is the non-linear relationship between the biophysical trait and phenotype. The effect of the single mutations, A and B, and the double mutant, AB, on the biophysical trait have the same magnitude in (A). The solid black double-headed arrow marks the magnitude of the expected phenotype of the double mutants, given

a linear model. The dashed grey lines are the measured biophysical trait and phenotypes in the mutants. The orange double-headed arrow is the epistasis, or the difference of the observed phenotype by the double mutants in the linear and non-linear models.

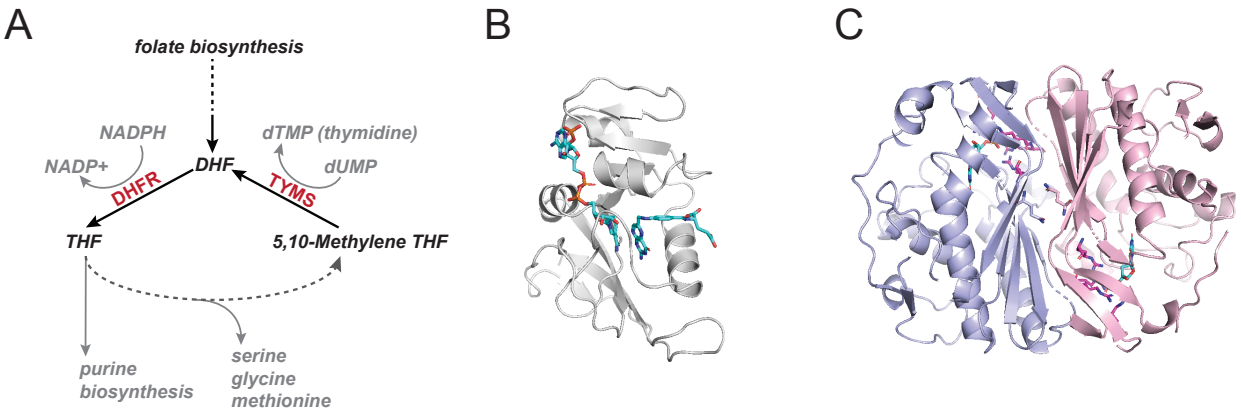

Figure 1.2 The model system. (A) An abbreviated folate metabolic pathway (B) The x-ray crystal structure of *E. coli* Dihydrofolate Reductase (PDBID:1RX2).[35] In cyan sticks are its substrate, dihydrofolate (DHF) and NADPH. (C) The x-ray crystal structure of *E .coli* Thymidylate Synthase, (PDBID:1BID).[36] Each homodimer has a different color (lavender on the left and pink on the right). In the active sites are cyan stick representations of 5,10-methylene THF and dUMP. The four arginine residues are represented in magenta sticks. Right in between homodimers are stick representations of Q33.

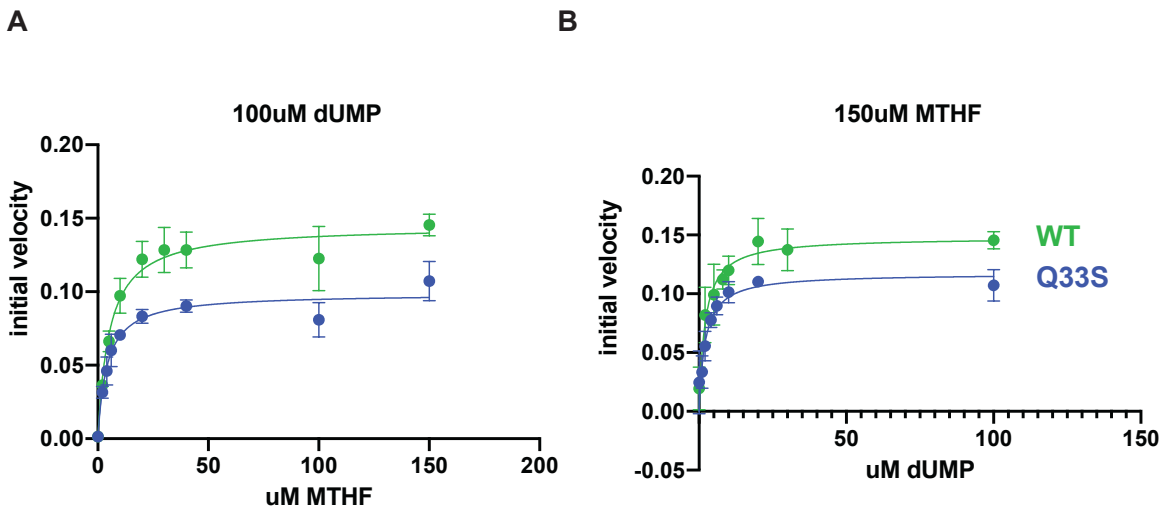

Figure 1.3 *In vitro* Michaelis Menten steady state kinetic measurements. In 100 µM of purified protein (WT TYMS (green) and Q33S TYMS (blue)).

1.8 Tables of enzyme kinetics for WT TYMS and Q33S TYMS. R166Q TYMS activity was not measurable.

Table 1.1 TYMS kinetics in 150 μM MTHF

TYNS	**Vmax**	**error**	**Km**	**error**	**kcat**	**error**	**kcat/Km**	**error**
WT	0.150	0.013	2.237	1.716	2.994	0.256	1.762	1.07
Q33S	0.118	0.009	2.150	0.613	2.355	0.184	1.152	0.299

Table 1.2 TYMS kinetics in 100 μM dUMP

TYMS	**Vmax**	**error**	**Km**	**error**	**kcat**	**error**	**kcat/Km**	**error**
WT	0.145	0.014	5.284	0.827	2.898	0.285	0.556	0.093
Q33S	0.100	0.010	4.425	1.841	1.990	0.193	0.496	0.172

1.9 Methods and Materials for *in vitro* enzyme kinetics of TYMS

1.9.1 Cloning TYMS into protein expression vector

The thyA gene was amplified by PCR from E. coli MG1655 and cloned into pET24A (using XbaI/Xho restriction sites) with no 6-His tag. TYMS point mutants were made using Agilent QuikChange II site-directed mutagenesis kit.

1.9.2 Protein induction and expression

All protein expression was carried out in BL21(DE3) cells transformed with expression vectors above. Transformed cells were grown in LB with Kanamycin (35mg/L) for selection overnight. Following overnight growth, cultures were back-diluted 1:100 in Superbroth and grown at 30°C to an optical density at 600nm (OD600) of 0.5-0.8; IPTG was added to 1mM final concentration. Induced cultures were incubated at 30°C for 4 hours. 0.2ml samples were taken before and after induction; samples were pelleted and supernatants removed. The induced cultures were pelleted at 5000 x g for 15 minutes, 4°C. Supernatants were removed and cell pellets were stored at -80C until purification performed.

1.9.3 Purification

Thawed cell pellets were resuspended in 5ml/50ml culture of lysis buffer (20mM Tris (pH 8.0), 10mM $MgCl_2$, 5mM DTT, 0.2mg/ml lysozyme, 5ug/ml DNAseI, 0.1% deoxycholate); resuspended cells were incubated for 20 minutes at room temperature with gentle rocking. Lysed cells were pelleted at 15K x g, 4°C, for 20 minutes to separate soluble and insoluble fractions. The

soluble fraction was removed from the pellet and combined with 0.3g/ml $(NH_4)_2SO_4$ (50% saturation) and incubated at 4°C with gentle rocking for 10 min before centrifuging at 15K x g, 4°C for 20 min. Supernatants were removed and an additional 0.2g/ml $(NH_4)_2SO_4$ was added to achieve 80% saturation and incubated 10-15 min at 4°C with gentle rocking. Centrifugation at 15K x g, 4°C for 20 min was performed and pellets retained. Pellets were resuspended in 25mM Potassium Phosphate pH 6.5 and dialyzed overnight in 25mM Potassium Phosphate pH 6.5. Anion exchange chromatography was performed using a 1ml HiTrap Q HP column (Cytiva); 25CV gradient to 1M NaCl in 25mM Potassium Phosphate pH 6.5. Fractions were combined, concentrated, and buffer-exchanged into 25mM Potassium Phosphate pH 8, 0.3M NaCl; size exclusion chromatography (HiLoad 16/600 Superdex 75, Cytiva) was performed in the same buffer. Fractions were collected and concentrated prior to making kinetics measurements.

1.9.4 Substrate preparation

(6R)-methylenetetrahydrofolic acid (MTHF) was purchased from Merck & Cie (Switzerland) and dissolved to 100mM in nitrogen-sparged citrate-ascorbate buffer (10mM ascorbic acid, 8.5mM citrate. 30µL aliquots were made in light-safe microcentrifuge tubes, flash-frozen in liquid nitrogen, and stored at -80C. Before use, stock was thawed and diluted to 10mM in TYMS kinetics reaction buffer and quantitated in an enzymatic assay: 50uM MTHF, 200uM dUMP and 1µM TYMS protein were combined and A_{340} measured until steady-state reached. Actual concentration is calculated from the difference in A_{340} before and after the reaction using Beer's Law (MTHF extinction coefficient: 6.4$mM^{-1}cm^{-1}$).

1.9.5 Steady state enzyme kinetics assay

To characterize k_{cat} and K_M for both substrates, we followed the protocol for steady-state Michaelis Menten kinetics adapted from Wang et al. and Agrawal et al.[23,37] Briefly, to measure initial velocity (V_o) of DHF production, we used a UV/Vis spectrophotometer (Perkin Elmer Lambda 650) to monitor increasing absorbance at 340nm ($\Delta\varepsilon_{340}$=6.4 mM^{-1} cm^{-1}). This wavelength is a proxy for DHF production.[23,37] The substrates (dUMP (Sigma) and N^5,N^{10}-methylene-5,6,7,8-tetrahydrofolate (CH_2H_4fol) (Merck & Cie)) were prepared from stocks to appropriate concentrations in TYMS Kinetics Assay Buffer (100mM Tris, 1mM EDTA, 5mM formaldehyde, 50mM DTT, pH 7.5) and allowed to equilibrate to room temperature. Protein was prepared in the same assay buffer and allowed to equilibrate to room temperature for a minimum of 30 minutes prior to making measurements. Measurements for determining Km for 5-10 methylene THF substrates were made using an array of 5-10 methylene THF at saturating concentration of dUMP (100μM). Similarly, to measure the enzyme's Km of dUMP , measurements were made using an array of dUMP concentrations at saturating concentration of 5-10 methylene THF (150μM). In all reactions, DHF accumulation was measured at A_{340} for 15 minutes in a Perkin-Elmer Lambda 650 spectrophotometer alongside a cuvette that contained no-protein as a reference at 25°C. A_{340} data was converted to DHF concentration using Beer's law, and the kinetic parameters (V_{max}, K_m, k_{cat}) were fit with a Michaelis-Menten model in GraphPad Prism software.

CHAPTER TWO
Defining experimental conditions for high resolution measurement of inter-protein epistasis

2.1 Background

Over the past decade, deep mutational scans have revealed the distribution of mutational effects on fitness in a number of individual proteins.[29,31–34,38,39] Fitness is defined as the measurable effect of a mutation in a protein on bacterial growth rate. The distribution of fitness effects from these experiments typically show two modes: one large peak centered near-WT and one much smaller peak with a deleterious fitness. The general interpretation of these distributions is that natural proteins are relatively robust to mutation, with a small number of mutations leading to catastrophic unfolding or loss in activity. However, two examples of DMS in one protein repeated over a range of conditions clearly show that the conditions of the assay modulate the shape of the fitness distribution substantially. In the first example, a DMS on the chaperone, Hsp90, was performed over a range of expression conditions. This study revealed that Hsp90 was increasingly sensitive to mutations at low expression levels.[32] The second example performed DMS on TEM-1 beta lactamase, the protein that confers resistance to the antibiotic, ampicillin to bacteria, across a range of ampicillin concentrations.[34] This study showed that under increasing levels of antibiotic concentration and thus greater selection pressure, more positions in TEM-1 beta lactamase were sensitive to mutations. This pattern of resides stretched from the core to the surface with increasing ampicillin concentration. Thus – consistent with intuition – increasing selection pressure (either by reducing protein expression or increasing the concentration of an antibiotic) decreases how

robust a protein is to mutations. Given this information, it was important to carefully consider the conditions in the selection step of the DHFR deep mutational scan.

For this reason, I performed a series of experiments that helped me choose the condition in which I performed the selection step of the Deep Mutational Scan of DHFR. The condition I chose fit two criteria. First, this condition must show an ability to resolve growth rates of DHFR mutants in the Calibration Curve along a broad range of catalytic activities in DHFR, from wild-type to catalytically inactive. Second, this condition must show that DHFR is epistatic to R166Q TYMS. Taken together, I sought to maximize the dynamic range over which I could detect changes in both DHFR activity and epistasis. My basic strategy was to measure growth rates for select DHFR point mutants in the background of specific TYMS mutations, variations in expression, and variations in media supplementation. Importantly, prior work has established steady-state Michaelis Menten parameters (k_{cat}, K_m) for 11 DHFR mutants (Table 2.1). These point mutants, which I refer to as the "Calibration Curve" mutants, span nearly five log orders in catalytic power. This allows me to closely examine the relationship between *in vitro* DHFR catalytic activity and growth rate, and establish the dynamic range and resolution of my assay. I used the bacterial strain *E. coli* ER2566 that contains genomic knockouts in both DHFR and TYMS.[40] I refer to this strain as ER2566 ΔfolA ΔthyA. For selections, DHFR and TYMS are expressed from pTET-duet, a dual expression vector. In front of DHFR is a T7 promoter. In front of TYMS is the tet promoter. DHFR itself is upstream of TYMS. The terminator of the T7 promoter comes directly after the TYMS coding region. When I refer to a mutation in DHFR or TYMS, I describe mutations in this vector. Lastly, DHFR and TYMS alleles are referred to with the following convention: WT/WT for wild-type

DHFR and wild-type TYMS, respectively; or M42F/R166Q for the mutation M42F in DHFR and the mutation R166Q in TYMS.

2.2 Optimizing conditions in the experimental measurements of epistasis between DHFR and TYMS

2.2.1 Modulating expression of DHFR with an altered ribosome binding site

First, I tested the effect of altering DHFR expression on growth rate. To do this, I used two variants of the ribosome binding site. The first one, named "RBS 1", is the RBS variant with the Shine-Dalgarno sequence, "AAGGAG". The second variant is called "RBS 3" and has the sequence "AATGAG". It has a lower translation initiation rate, leading to DHFR expression that is approximately 0.05x lower than the Shine-Dalgarno RBS 1.[33,41]

For each RBS, I measured the growth rate of all possible single point mutants in the first 40 amino acid positions of DHFR with the NGS-Fit Assay. This is sublibrary 1 of the full DHFR saturation mutagenesis library. This allowed me to evaluate the effects of modulating expression in a simpler and less costly experiment. I provide a deeper description of the NGS-fit assay in Chapter 3 (see section 3.3 and Figure 3.2). In brief, I transformed the library of interest into the knockout E. coli strain, and used next-generation-sequencing (NGS) to monitor the frequencies of thousands of mutant alleles in parallel under conditions in which *E. coli* growth rate is coupled to DHFR activity. During this experiment, the *E. coli* were grown in a turbidostat with M9 minimal, 0.4% glucose media supplemented with 0.4 μg/mL thymidine. From this experiment, I obtained growth rate

measurements in both RBS backgrounds for a total of 762 mutants (~97% coverage of the sublibrary), after excluding the start codon and filtering for mutants that are present at the beginning of the experiment. A mutant was considered present if it had 10 sequencing reads at the t=0 time point and two other later time points.

Across the entire sublibrary, we observed lowering expression with RBS 3 broadly slowed growth. In the correlation between these two data sets, the R^2 is 0.68 (Fig. 2.1B). The mean relative growth rate in the RBS 1 background is 0.531 hr^{-1}. In the RBS 3 background, the mean relative growth rate is much lower at 0.116 hr^{-1}. While the dynamic range of growth rates is greater in the RBS 3 background, it was difficult to accurately observe growth rates for deleterious mutants. In the RBS 3 background, 95 mutants were considered absent and thus filtered from analysis. These mutants had too few sequencing reads, failed to meet the criteria for presence in the population. In contrast, in the RBS 1 background, only 19 mutants were filtered from analysis. By assessing how the RBS background across mutants in the sublibrary as a whole, I was able to identify a tradeoff in lowering DHFR expression that informed my decision to choose RBS 1. The greater dynamic range of relative growth rates in the RBS 3 cost noisier and incomplete growth rate fits in mutants with deleterious growth rate effects.

To evaluate the effects of modulating expression in a more focused way, I considered the subset of mutants in the Calibration Curve (Fig. 2.1A). In both RBS 1 and RBS 3 backgrounds, growth decreased with drops in DHFR catalytic activity. In the case of the DHFR mutants in the RBS 1 background, this decrease in growth is more gradual. The relationship between growth rate and

DHFR biochemical activity is much steeper when protein expression is lowered by RBS 3. This applies to the DHFR variants with catalytic activities that are within an order of magnitude of the wild-type. D27N DHFR, the mutant that sets the floor for both growth and catalytic activity, behaves unexpectedly at lower expression levels. In the context of RBS3, this mutant is apparently more fit than the other DHFR mutants that catalyze their reactions at greater speeds. D27N DHFR in the RBS3 background also appears to be more fit than D27N in the RBS1 background. This is inconsistent with the bulk of the growth rate data, which suggest that decreasing expression leads to more deleterious effects of mutations on growth. These discrepancies in the growth rate effect of D27N DHFR come from lower allelic frequencies in RBS 3 than in RBS 1, leading to large statistical noise. In general, it is challenging to measure the fitness effects of highly deleterious mutations which are present in low frequencies after a short time period of selection. This is clear in the data for DHFR D27N. The increased severity of growth rate defects with lowered protein expression also applies to the rest of the sublibrary of 762 DHFR mutants (Fig. 7.1B).

Ultimately, I opted against lowering protein expression with RBS 3. My logic was that reducing expression would lead to a large number of highly deleterious mutations, making growth rate measurements statistically noisy for a large fraction of my library. Additionally, read coverage and data quality are precious capital in NGS-based assays. Lower DHFR expression meant sacrificing data quality for greater resolution among near-WT and moderate DHFR mutational effects. The RBS 1 did successfully resolve these same mutants but in a narrower range of growth rates. I preferred an outcome where I was able to measure the entirety of the saturation mutagenesis library. RBS 1 would yield higher data quality, greater read depth for a larger portion of the library,

and thus more accurate growth rate measurements. All experiments in the rest of this document were in the context of RBS1.

2.2.2 Modulating environment with media

Before proceeding with the deep mutational scan, it was important to ensure that the epistasis between DHFR and TYMS was robustly detectable. The simplest way to modulate epistasis was by adjusting the media conditions of the experiment. Here, I used the Growth Rate Assay in the plate reader (see section 2.6.1) to measure growth rates of DHFR and TYMS mutants across different conditions in high throughput.

I first focused on the effect of thymidine supplementation on epistasis in just one mutant, G121V DHFR. Here, I grew four mutants in two different M9 minimal media conditions: in the presence or absence of 50 μg/mL thymidine (Fig. 2.2). These four mutants are WT/WT, G121V/WT, WT/R166Q, and G121V/R166Q. This experiment was performed in triplicate in a single 96-well plate. The resulting data in the assay are 24 hour time courses of OD600. The growth rates were linearly fitted over an empirically determined exponential phase of growth. In this experiment, this range was between 0.06 and 0.20. All growth rates were fitted with the same range of OD600s. In the absence of thymidine (Fig. 2.2A), only the WT/WT and G121V/WT were able to grow. As expected, G121V/WT grew slowly. Without the ability to synthesize thymidine nor an exogenous source of thymidine, the R166Q TYMS mutants were not able to survive. In the presence of thymidine (Fig. 2.2B), the pressure on R166Q TYMS is relieved and G121V is buffered by R166Q TYMS. Epistasis of G121V DHFR to this TYMS mutant is computed from these four growth rate

measurements (Fig. 2.2C, equation 1.4). In the presence of thymidine, the signal of epistasis is much greater.

Next, I expanded these measurements of growth and epistasis to the remaining Calibration Curve (Fig. 2.3). To control for day to day variation between experiments, I normalized growth rates with WT/WT. I qualitatively reproduced the pattern of epistasis between DHFR and TYMS reported in Schober et al., where DHFR mutations are either fully rescued or partially rescued by a loss of function in TYMS (Fig. 2.3A). I also observe variation in the magnitude of epistasis across these mutants (Fig. 2.3B). Relieving selective pressure on TYMS function by supplementing thymidine into the media was key to detecting epistasis across a range of DHFR catalytic activities.

I next asked whether further supplementation of the growth media could increase the signal to noise ratio of epistasis. The purpose of amino acid supplementation is to reduce selective pressure on pathways downstream of DHFR that are responsible for synthesizing amino acids and purines. Here, I tested whether supplementation with FolA Mix (see section 2.6.6.3) or 0.4% amicase affected the signal of epistasis. In this experiment, I chose to perform the Growth Rate Assay on three mutants that spanned a range of DHFR catalytic activities (Fig. 2.4). When supplemented with either FolA Mix or amicase, the signal of epistasis was greater for the two more deleterious mutants, F31Y/L54I and D27N. For DHFR mutants with near-WT (M42F) and moderate (F31Y) catalytic activities, the signal of epistasis was slightly diminished. The presence of amino acid supplementation increased the difference of epistasis between these two groups of DHFR mutants.

For the sake of simplicity, I chose to supplement the growth media with 0.4% amicase over FolA Mix.

The final media condition I chose to perform the deep mutational scan in replicates the condition used in Schober et al. Regardless, these experiments were useful to rigorously test what conditions were reasonable for measuring both growth rate and epistasis for a saturation mutagenesis library. This process is necessary in the development of Deep Mutational Scans in general to ensure that the output of the assay actually reports on the effect of mutations in the cell. I chose to take the time to rigorously test these conditions on a smaller set of DHFR mutants to avoid a scenario where I collected NGS data were noisy and uninterpretable for the entirety of the saturation mutagenesis library.

2.3 Evaluating effect of plasmid-based expression of DHFR and TYMS on biochemical activity in cellular lysates

Our model system expresses DHFR and TYMS from a plasmid; not endogenously from the genome. Since *E. coli* in the natural world do not typically do this for essential metabolic enzymes, it worthwhile to ask whether and how much our model system expresses DHFR and TYMS compared to endogenous expression from the genome.

In these experiments, the goal is to compare the activities of DHFR and TYMS in the lysates of strains that express these enzymes endogenously in the genome to strains that do so exogenously in the pTET-duet plasmid vector.

The cell lysate assays for DHFR and TYMS use a spectrophotometer to monitor the absorbance of the product of TYMS and the substrate of DHFR, dihydrofolate (DHF). DHF can be detected spectrophotometrically at an absorbance of 340 nm. In the DHFR lysate assays that monitor DHFR activity, DHF is consumed by DHFR, resulting in a drop of the signal over the course of the experiment. In the TYMS lysate assays, the opposite is true, DHF accumulates over the course of the experiment as TYMS synthesizes it. This drives an increase of the signal over the course of the experiment.

These data were collected in collaboration with Christine Ingle, Research Scientist in the Reynolds Lab collected the following data in this section of the chapter. I provided the reagents, developed the protocol for the cell lysate assay, and analyzed the data she collected.

2.3.1 DHFR Cell Lysate Assay

Four *E. coli* strains were assayed for DHFR activity. The first is ER2566 "WT", which contains genomic versions of DHFR and TYMS. The second is the genomic double knock out of these two enzymes, ER2566 ΔfolA ΔthyA. The third strain is the double knockout with the plasmid containing WT DHFR and WT TYMS, ER2566 ΔfolA ΔthyA + pTET-duet (WT/WT). The last strain is the double knockout with a WT DHFR and R166Q TYMS. These cell lysates were prepared as described in the Materials and Methods section and then pre-incubated with 100 μM NADPH (see section 2.6.2.1.2 on how I prepared this stock). After adding 100 μM of DHF to the cell lysate in the cuvette, the sample was immediately placed in the spectrophotometer where the A340 nm was monitored for 180 seconds. Figure 2.5A-B shows what these time course data look

like after base-line normalization. The strain that expresses DHFR endogenously has a much weaker signal relative to the strains that express DHFR from the plasmid. This difference is so stark that I felt it was necessary to "zoom into" the data for ER2566 "WT" and the double-knockout strains (Fig. 2.5B, Fig. 2.5D). In the signal for the double-knockout that expresses no DHFR, the A340nm signal does decrease until approximately 75 seconds into the experiment. After this, the signal levels off. This shows that outside of the DHF that is added to the sample, there are other molecules in the lysate that are photoactive at an absorbance of 340nm, or that DHF is undergoing some other degradation. Additionally, the presence of the loss of function mutation in TYMS controls sources of DHF outside of the 100 μM added to the sample. Overall, it is reasonable to conclude that the plasmids expresses DHFR at much higher levels (approximately 100-fold greater) than the natural, endogenous ER2566 "WT" strain (Fig. 2.5C-D).

2.3.2 TYMS Cell Lysate Assay

In this assay, the activities of three TYMS variants were expressed from the pTET-duet plasmid in ER2566 ΔfolA ΔthyA: WT, Q33S (intermediate), and R166Q (loss of function). In the TYMS lysate assay, DHF production by TYMS is monitored after adding the cofactor (150 μM MTHF) and substrate (100 μM dUMP) (Fig. 2.6A). To control for DHF abundance, the TYMS variants are paired with an inactive D27N DHFR. Alongside strains that express DHFR and TYMS from the plasmid, we included the strains ER2566 "WT" and ER2566 ΔfolA ΔthyA. TYMS endogenously expressed from the genome of ER2566 "WT" showed the greatest TYMS activity (Fig. 2.6B). The double-knockout strain set the floor of activity and was comparable to activity by R166Q TYMS (Fig. 2.6B). WT TYMS and Q33S TYMS activities from the plasmid were at

similar levels to each other but had less than half of the activity of endogenous TYMS. Like with DHFR, these data show that activities of TYMS expressed from the plasmid and the activity of TYMS expressed from the genome are not similar in magnitude to each other. This experiment also clearly showed that a mutation that renders the enzyme catalytically inactive is reflected by a corresponding drop in TYMS activity in the lysate.

2.4 Conclusions

For the selection step of the DMS, I chose to grow the *E. coli* cultures in minimal media supplemented with thymidine and amicase. I also decided to against using a genetic condition where a mutation in the ribosome binding site lowered expression of DHFR and TYMS off the plasmid. Through these carefully designed growth rate assays, I identified this set of experimental conditions that allowed me to resolve the changes in DHFR activity across nearly 5 orders of magnitude. These environmental conditions also emphasized the epistasis between DHFR and TYMS. After choosing these conditions, I used lysate assays to characterize the activities of DHFR and TYMS in the context of this set of conditions. I found that the plasmid-based expression system in this condition were significantly higher than the activities in the cell lysates of genomic, wild-type DHFR and wild-type TYMS. Along with differences in enzymatic activity, the growth rate effect of mutations in DHFR are also a function of the abundance of the enzyme in the cell. Taken together, this work were the conditions of the growth rate assay to measure epistasis between DHFR and TYMS.

2.5 Figures

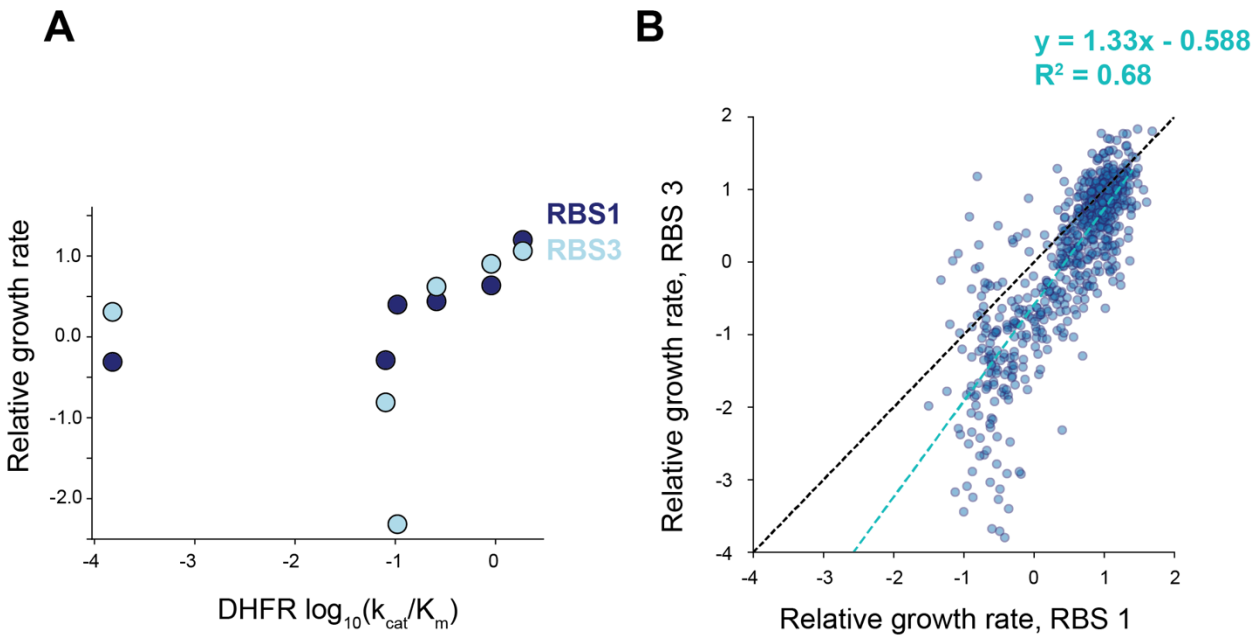

Figure 2.1. The effect of a mutation in the ribosome binding site (RBS) on growth rate effects of single point mutants of DHFR. **(A)** The Calibration Curve DHFR point mutants in the context of two RBS variants: the Shine-Dalgarno sequence (RBS 1, navy blue) and a mutant that lowers transcriptional activity (RBS 3, light blue). **(B)** Comparing RBS variant effect on the relative growth rates of mutants in sublibrary 1 of DHFR. Relative growth rates are normalized such that a mutant with a growth rate of 1 is equivalent to WT. RBS 1 is on the x-axis, RBS 3 is on the y-axis. The navy markers are the relative growth rates for individual mutations. The black dashed line shows Y = X. The cyan dashed line is the correlation fitted with a linear regression. The cyan text reports the equation of the fit (Y = slope*X + intercept) and the R-Squared statistic (R^2).

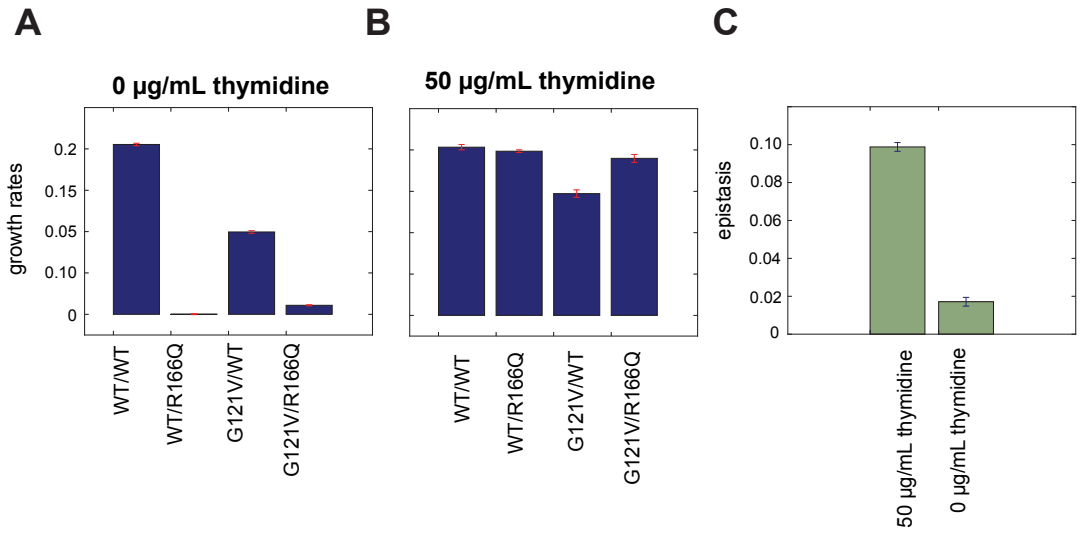

Figure 2.2 Thymidine supplementation increases signal of epistasis. Growth rates of *E. coli* strains with G121V DHFR and/or R166Q TYMS mutations in media **(A)** with no thymidine supplementation or **(B)** with 50 µg/mL thymidine. **(C)** Epistasis of G121V DHFR to R166Q TYMS in the presence and absence of thymidine.

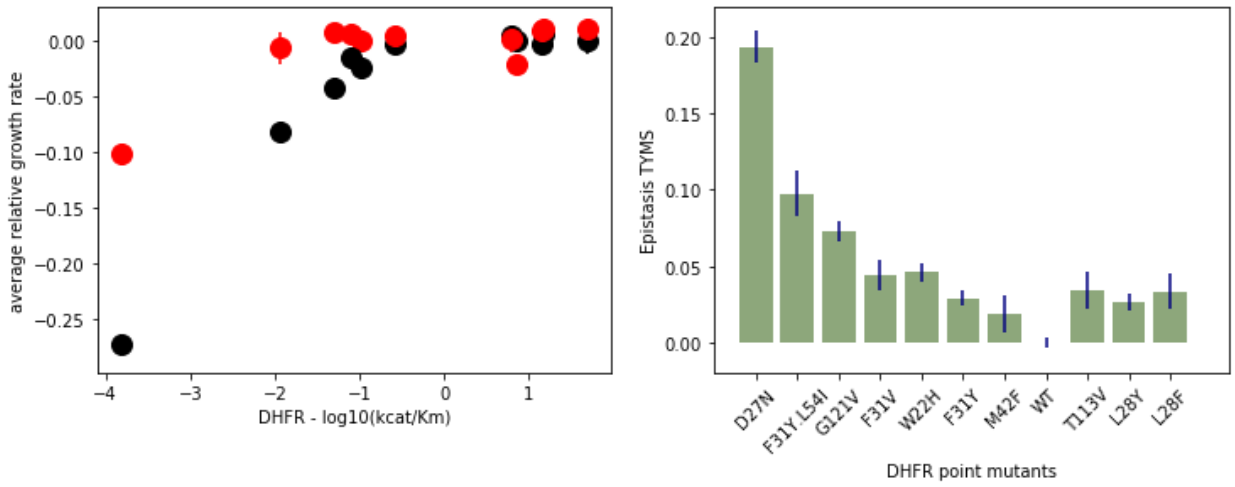

Figure 2.3 The relationship between DHFR catalytic activity, growth, and epistasis in minimal media supplemented with 50 µg/mL thymidine. **(A)** Calibration curve in media with 50 µg/mL thymidine. The relative growth rates were normalized by WT. In black are the single mutants, with a WT TYMS background. In red are the double mutants with a R166Q TYMS background. **(B)** Epistasis of each DHFR mutant to R166Q TYMS, ordered by DHFR catalytic activity.

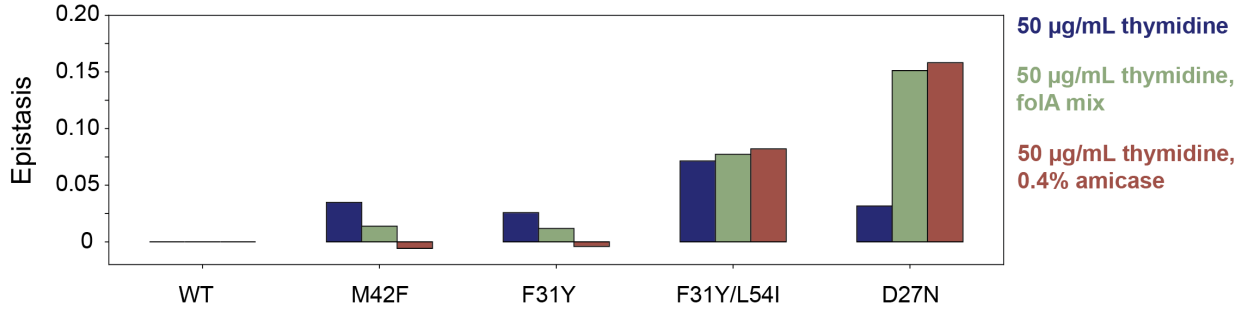

Figure 2.4 The effect of amino acid supplementation on epistasis select DHFR mutants. These experiments were done in media with only thymidine (navy), thymidine and 1X FolA mix (green), or thymidine with 0.4% amicase (brown).

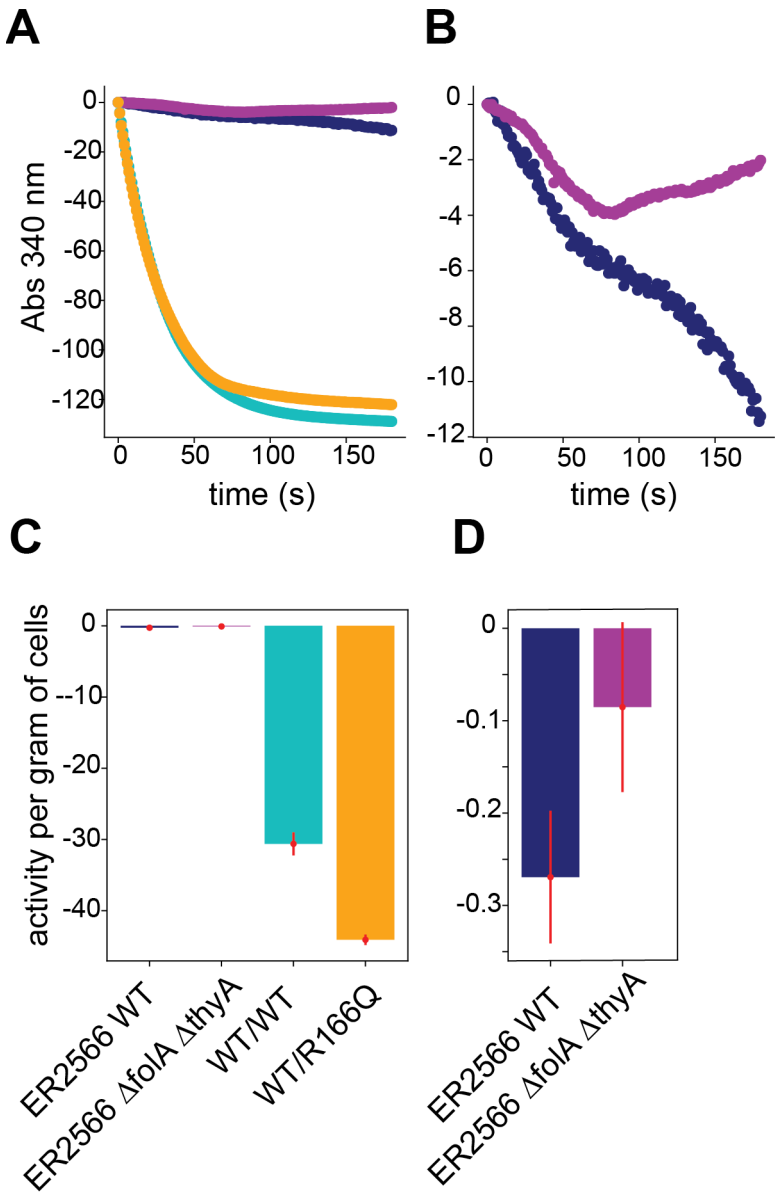

Figure 2.5 Estimated DHFR activity in cell lysates. (A-B) Time course of absorbance at 340 nm in cell lysates. (A) ER2566 WT in navy, ER2566 ΔfolA ΔthyA in magenta, the double knockout expressing WT DHFR and WT TYMS from pTET-duet plasmid in cyan, WT DHFR and R166Q TYMS from pTET-duet in orange. **(B)** Zoomed in view of ER2566 WT (navy) and ER2566 ΔfolA ΔthyA (magenta). **(C-D)** Activity, or rate of DHF consumption was inferred from

the slopes of the curves in A and B. These activities were then normalized by cell mass in the sample. For strains with plasmids, the slopes were fitted with a linear regression over the linear parts of the curves (first 20 seconds for WT/WT, first 40 seconds for WT/R166Q). In the strains without plasmids, a linear regression was fitted over all of the data. **(D)** Zoomed in view of ER2566 WT (navy) and ER2566 ΔfolA ΔthyA (magenta).

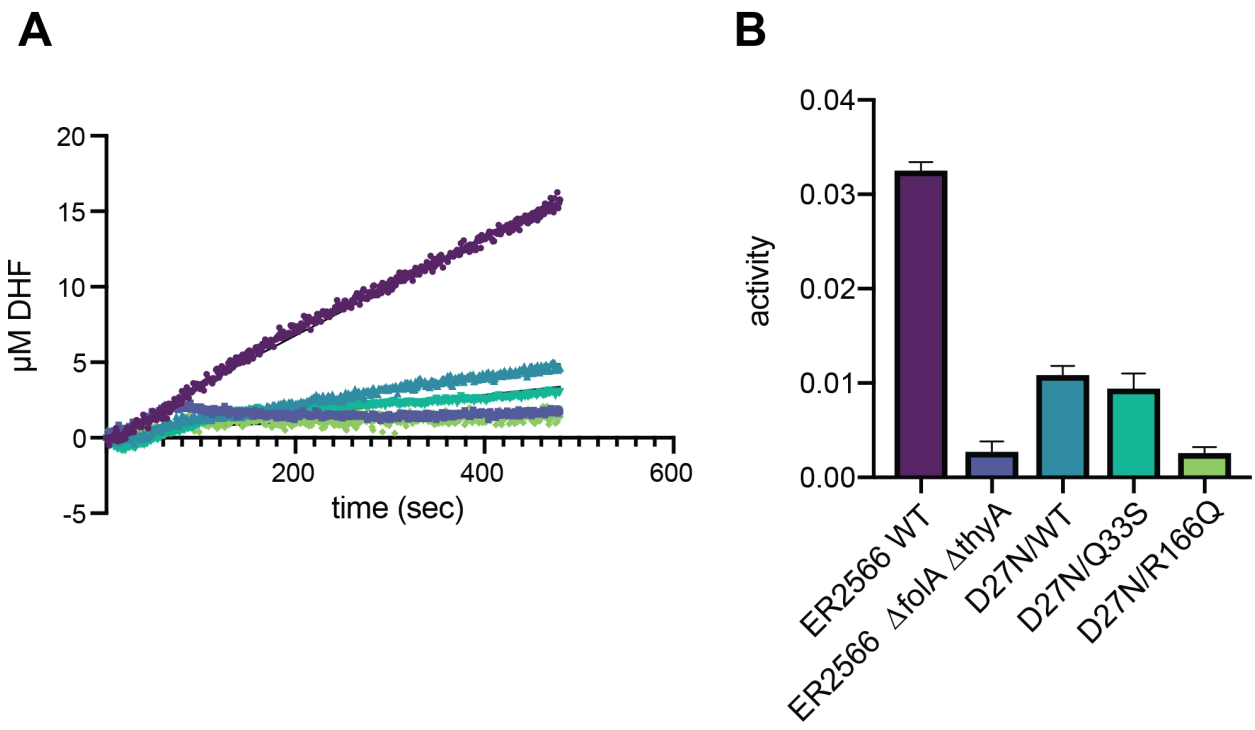

Figure 2.6 Estimated TYMS activity in cell lysates. (A) A time course of DHF production in cell lysates of the following *E. coli* strains: ER2566 WT (dark purple), ER2566 ΔfolA ΔthyA (genomic knockouts of DHFR and TYMS, respectively; in light purple), the double knockout with pTET-duet with D27N DHFR and WT TYMS (dark blue-green), or D27N DHFR and Q33S TYMS (light blue-green), or D27N DHFR and R166Q TYMS (lime green). The lines through the markers linear regressions for respective strains. **(B)** Rate of DHF production, or the activity in the cell lysate are the slopes of the linear fits of each strain. The colors of the bars correspond with the color of the traces in (A).

2.6 Materials and Methods

2.6.1 Growth rate measurements of individual DHFR/TYMS mutations

For each mutational variant, a streak of colonies from LB agar plates was grown overnight at 37°C in M9 minimal media with 0.4% glucose (with or without supplementation). The next morning, all overnight cultures were washed and back-diluted to an optical density at 600 nm (OD600) of 0.1. into the same minimal media described above for adaptation at 30°C for 4 hours (220 rpm shaking). Following adaptation, each culture was back-diluted to OD600 0.1. 10 µL of this back-diluted cultures was used to inoculate a 200 µL volume to a final OD600 of 0.005 in a 96 well plate. The plate was incubated at 30°C in a plate reader (Perkin Elmer, VictorX3), that the OD600 was monitored for 24 hours with a program that repeated throughout the course of the experiment. A description of the program follows: the plate would shake (30 seconds, normal speed, orbital type, 1.80 mm diameter), read the OD600 of each well for 0.5 second, and delay for 570 seconds before shaking the plate again. After every third shake, OD600 measurement, and delay, 5 µL of sterile dH_2O was dispensed into each well. The edge wells of each plate were filled with media, but not inoculated to avoid artifacts due to evaporation. Growth rates were estimated for culture by linear regression of the log-linear portion of each growth curve.

2.6.2 Assay of DHFR activity in cell lysates

Cell lysates were prepared by growing the following *E. coli* strains overnight at 37°C. ER2655 was cultured in LB only. ER2566 ΔfolA ΔthyA were cultured in LB and supplemented with 50 µg/mL thymidine. ER2566 ΔfolA ΔthyA containing pTET-duet that expressed DHFR/TYMS

point mutants were cultured in LB supplemented with 50 µg/mL thymidine and 30 µg/mL chloramphenicol. The overnight cultures were back-diluted to an OD of 0.1 in respective LB media and then grown for 4 hours at 37°C. The OD600 of these cultures were recorded to permit normalization of lysates in later steps. These cultures were pelleted at 2348 rcf for 5 minutes at room temperature. The supernatants were discarded and the pellets were stored in -20°C freezer for at least one night so that the cells undergo an initial lysis via a freeze/thaw cycle. On the day of the lysate assay, the cells are thawed at room temperature and then returned to ice. Meanwhile, the lysis buffer was prepared in ddH_2O (20 mM Tris, pH 7.5, 50 mM MgSO4, 0.1% DOC, 2 mg/mL lysozyme, 5 ng/mL DNAse, 500 µM DTT). Each thawed pellet was resuspended with a volume of lysis buffer to normalize equal amount of cells in each sample pellet. The culture with the lowest OD600 received the smallest volume of lysis buffer (in these experiments, 1.2 mL of lysis buffer).

After resuspension with lysis buffer, the lysate is then incubated at room temperature on a nutator for 15 minutes. Then the debris from the pellet was pellet at 15,000 rcf at 4°C for 15 minutes in a microcentrifuge. The supernatants were immediately transferred to fresh Eppendorf tubes. This is the undiluted cell lysate to be used in the spectrophotometric assay. Alongside cell lysates, blank lysis buffer was also pelleted. This blank lysis buffer was used to prepare diluted cell lysates in triplicate. These samples were: undiluted, 1:2, 1:4, 1:8, and 1:16.

During the lysis and pelleting steps, the remaining reagents for the assay were prepared. The assay buffer was prepared by diluting 1M DTT with MTEN, pH 7.0 (50 mM MES, 25 mM Tris Base,

100 mM NaCl, 25 mM ethanolamine hydrochloride) in a 1:20 dilution and then placed on ice. The assay buffer was then used to dilute the frozen DHF stock to a working concentration of 1000 μM. The prepared DHF substrate was covered with foil and placed on ice (see section 2.6.2.1.1 for how we prepared the DHF stock). In a separate conical tube, the NADPH substrate was prepared by diluting the frozen NADPH stock to a working concentration of 1000 μM (see section 2.6.2.1.2 for how to prepare NADPH stock). The diluted and undiluted cell lysates were pre-incubated in 100 μM NADPH – 900 μL of the cell lysate was combined with 100 μL of the 1000 μM NADPH working stock in an Eppendorf tube. 30 minutes ahead of measurements of 340 nm in the spectrophotometer, the 1000 μM DHF working stock and the cell lysate pre-incubated with 100 μM NADPH were equilibrated in a 25°C water bath for 30 minutes (in the dark). A blank sample containing no cell lysate was prepared from 900 μL pelleted lysis buffer with 100 μM NADPH and 100 μL of the DHF working stock. For each reaction, 900 μL of the cell lysate + 100 μM NADPH were combined with 100 μL of the 1000 μM DHF (for final DHF concentration of 100 μM) into a quartz cuvette. The UV-vis spectrophotometer then immediately recorded the absorbance of the sample at 340nM at 25°C for 2-5 minutes. The quartz cuvettes were thoroughly cleaned between each sample with dH_2O three times, acetone two times, dried with compressed air, and wiped free of smudges with lens paper.

2.6.2.1 Preparation of key DHFR lysate assay reagents. Adapted from Reynolds et al. [28]

2.6.2.1.1 DHF stock

MTEN, pH 7.0 was pre-chilled at 4°C. The amber ampule containing 10 mg of dihydrofolic acid (CAS number: 4033-27-6) was kept on ice throughout preparation of the stock solution. As soon

as the ampule was broken open, 14 µL of beta-mercaptoethanol was added to the dihydrofolic acid powder to prevent degradation. Then 1 mL of pre-chilled MTEN buffer was added to the ampule and resuspended. This was transferred to a foil-wrapped conical vial (which limited light exposure) and placed on ice. This resuspension step in MTEN was repeated 3 more times for a total of 4 mL of DHF stock. To quantify the concentration of DHF, 10 µL of the prepared stock was used in a series of 20 µL volume serial dilutions at 1:1, 1:10, 1:100, and 1:000 ratios of DHF:MTEN. Each of these serial dilutions were measured on the nanodrop at absorbance of 282 nm. The absorbance of the undiluted stock was back calculated for each sample and averaged. The concentration of the undiluted stock was calculated from this absorbance using Beer's Law (extinction coefficient of DHF is 28 $mM^{-1}cm^{-1}$). The stock was then aliquoted into 150 µL volumes in black opaque Eppendorf tubes. These aliquots were then flash frozen in liquid nitrogen and then immediately transferred to -80°C for storage.

2.6.2.1.2 NADPH stock

A 4 mM stock of NADPH (CAS Number: 2646-71-1) was prepared in 10 mL pre-chilled MTEN, pH 7.0 buffer. The concentration of this stock was verified using the same method as in the quantification of DHF. The absorbances of the serial diluted samples were measured at 340 nm. The extinction coefficient of NADPH is 6200 $M^{-1}cm^{-1}$. Once quantified, the samples were aliquoted into clear walled Eppendorf tubes, flash frozen with liquid nitrogen, and then stored at -80°C.

2.6.3 Assay TYMS activity in cell lysates

The cell pellets and lysates of *E. coli* strains ER2566, ER2566 ΔfolA ΔthyA, and ER2566 ΔfolA ΔthyA with pTET-duet expressing DHFR/TYMS mutants (D27N/WT, D27N/R166Q, D27N/Q33S) were prepared in the same manner as the pellets for the DHFR cell lysate assay. The blank lysis buffer was used to prepare 1:2 and 1:4 dilutions of cell lysate, in triplicate.

During the lysis and pelleting steps, the remaining reagents for the assay were prepared. The TYMS Kinetics Assay Buffer (100mM Tris, 1mM EDTA, 5mM formaldehyde, 50mM DTT, pH 7.5) was diluted with 1M DTT in a 1:20 dilution and then placed on ice. This diluted TYMS Kinetics Assay Buffer was then used to prepare 1000 μM dUMP and 1500 μM MTHF from frozen -80°C stocks (MTHF stock preparation previously described in section 1.9.4). The conical tube holding the substrate, which contained 10x dUMP and MTHF in assay buffer, was wrapped in foil and placed on ice. A blank was prepared from pelleted lysis buffer.

30 minutes ahead of measurements in the UV-vis spectrophometer, the cell lysate and substrate (the conical vial with dUMP and MTHF) were incubated separately in a 25°C water bath. For each sample measurement, 900 μL of the cell lysate and 100 μL of the substrate (for a final concentration of 100 μM dUMP and 150 μM MTHF) were combined in a quartz cuvette. This sample was immediately placed in the UV-vis spectrophotometer and the absorbance at 340 nM was monitored for 2-5 minutes.

2.6.4 Tables

Table 2.1. Calibration Curve Michaelis-Menten enzyme kinetics

Mutant	**kcat (s^{-1})**	**Km (μM)**	**Reference**
WT	7.95	1.1	Reynolds et al. Cell 2011
W22H	1.89	18	Reynolds et al. Cell 2011
L28F	18.5	9.9	Thompson et al. eLife 2020
L28Y	19.2	21.2	Thompson et al. eLife 2020
F31V	8.65	108	Reynolds et al. Cell 2011
F31Y	20.61	80	Reynolds et al. Cell 2011
M42F	79.2	13	Reynolds et al. Cell 2011
L54F	6.3	0.7	Huang et al. Biochemistry 1994
L54I	7.88	35	Reynolds et al. Cell 2011
T113V	32.9	21.4	Fierke and Benkovic Biochemistry 1989
G121V	0.3	6.1	Reynolds et al. Cell 2011
F31Y/L54I	1.94	168.3	Reynolds et al. Cell 2011

Table 2.2. Calibration Curve plasmids (pTET-duet, RBS 1)

pTN	DHFR	TYMS
312	WT	WT
342	W22H	WT
343	D27N	WT
344	L28F	WT
345	L28Y	WT
346	F31Y	WT
347	M42F	WT
337	G121V	WT
339	F31V	WT
341	F31Y/L54I	WT
335	T113V	WT
315	WT	R166Q
336	F31Y	R166Q
349	W22H	R166Q
350	D27N	R166Q
351	L28F	R166Q
352	L28Y	R166Q
353	F31V	R166Q
354	M42F	R166Q
355	T113V	R166Q
338	G121V	R166Q
341	F31Y/L54I	R166Q
408	WT	Q33S
409	D27N	Q33S

Table 2.3. DHFR sublibrary 1 (pTET-duet)

pTN	RBS	DHFR	TYMS
325	1	SL1	WT
330	3	SL1	WT

Table 2.4. *E. coli* strains

gTN	Strain	Description
101	ER2566	"WT"
331	ER2566 dfolA dthyA	Genomic double knockouts

2.6.5 M9 minimal media recipe

The following autoclaved media components were combined to the following final concentrations: 1X M9 salts, 0.4% glucose (w/v), 2 mM $MgSO_4$, and 0.2% amicase (w/v).

The pH of the media was adjusted to pH 6.5 and sterile ddH_2O was added to volume. The media was sterile filtered with a 0.22 μm, PVDF filter. Supplements to the media were withheld until the start of the experiment.

2.6.6 Supplement recipes

2.6.6.1 50 mg/mL thymidine (1000X stock)

0.5 g thymidine was dissolved into 10 mL ddH_2O by vortexing then incubated in a 55°C water bath for 10-15 minutes. After incubation, the mixture was vortexed to dissolve thymidine into solution. I repeated the incubation and vortex steps as necessary until the thymidine fully dissolved. The solution was sterile filtered with a 0.22 μm PVDF filter into a sterile conical vial and stored at room temperature. The stock was not used once precipitated (approximately 1 week).

2.6.6.2 30 mg/mL chloramphenicol (1000X stock)

0.3 g of chloramphenicol was dissolved into into 10 mL 100% etOH by vortexing and then stored at -20°C.

2.6.6.3 FolA mix (250X, 50 mL, in ddH_2O)

The following components were dissolved into ddH_2O by incubating at 37°C for 10-15 minutes: 475 mg glycine, 943.8 mg methionine, 12.5 mg pantothenate, and 250 mg adenosine. After the

components are dissolved into solution, the stock was sterile filtered with 0.22 μm PVDF filter. The stock was then aliquoted into Eppendorf tubes and stored at -20°C. Each experiment used a fresh stock because freeze/thaw cycles of the FolA Mix were poorly tolerated.

CHAPTER THREE
THE STRUCTURAL DISTRIBUTION OF EPISTASIS BETWEEN A PAIR OF ESSENTIAL METABOLIC ENZYMES

3.1 Introduction

Mutational scanning provides an approach to map residue-level interactions both within and between proteins. Historically, this strategy has been used to applied to map the pattern of constraints between physically binding protein pairs For example, consider Clackson and Wells' classic study, "A hot spot of binding energy in a hormone-receptor interface".[6] In this work, they performed an alanine scanning mutagenesis assay over many positions in the interface between human growth hormone (hGH) and human growth hormone binding protein (hGHbp). They measured the thermodynamic free binding energy of the mutants. They discovered that not all residues were acting equally. Only a few residues, or "hot spots" were contributing to most of the free binding energy that drove the formation of the interface. Without this method of alanine scanning mutagenesis of all residues at interface, it would have been impossible to observe this heterogenous pattern of epistatic interactions within the interface of the hGH-hGHbp physical complex.

The study of the interface between hGH and hGHbp focused on the amino acids that form the interface, not the rest of the protein. This pattern of epistasis at the amino acid sequence level in physical complexes beyond the interface was outstanding until 2018 when Diss and Lehner published the first full saturation mutagenesis between Fos and Jun, a pair of proteins that physically bind to form a transcription factor. In this work, a thermodynamic model was able

predict the effect of double mutants on the ability of Fos and Jun to physically interact. The interactions between mutants that were not captured by the thermodynamic model localized at interfacial positions between the two proteins. As of November 2021, this study was the first instance of a comprehensive saturation mutagenesis between two proteins in a physical complex. These mutational studies of physical protein complexes shaped our understanding of how proteins recognize one another with high specificity and affinity, and profoundly influenced strategies for engineering new complexes and drugging protein interfaces.

In contrast, the pattern of epistasis between proteins that interact functionally but not physically, like DHFR and TYMS, remains relatively unexplored. Because these two enzymes functionally interact through a constraint on their relative activities, the structural distribution of epistasis is non-intuitive. However, mapping the pattern of epistasis between metabolic enzymes is a key step towards understanding the evolution of metabolic pathways, and defining principles for engineering them. This chapter describes the first study of a deep mutational scan in a pair of proteins that are not known to form a physical complex.

I measured the fitness effects for nearly every single mutation of DHFR in the context of three alleles of TYMS. These are the fully functional WT TYMS, the slightly less functional TYMS Q33S, and the fully inactive loss of function mutant, R166Q TYMS (see Chapter 1 for an introduction to these mutations, 1.8 for a table of enzyme kinetics, and section 1.9 for a protocol of TYMS enzyme kinetics).

3.2 The saturation mutagenesis libraries.

I began these experiments by characterizing the DHFR saturation mutagenesis libraries. The DHFR saturation mutagenesis library was divided into four sublibraries: named SL1, SL2, SL3, and SL4. This is to facilitate growth rate measurements using short-read sequencing (NGS typically covers a 300 base-pair region). Each sublibrary contained all possible amino acid single point mutations for 40 positions in DHFR. For instance, SL1 contained mutants for only the first 40 positions in DHFR (excluding the start codon), SL2 carried mutants for positions 41-80. SL3 does for 81-120. SL4 contained mutants for the remaining 39 positions (excluding the stop-codon). The sublibraries were originally generated by Samuel M. Thompson during his PhD work in the Kortemme lab (UCSF). To characterize the dependence of DHFR mutations on TYMS background, I duplicated these sublibraries two times through subcloning: once in the pTET-duet construct that expressed R166Q TYMS and another time in the same vector with the mutation, Q33S TYMS. This yielded a total of three DHFR saturation mutagenesis libraries (12 sublibraries), each of which are paired with a different TYMS variant: WT, Q33S, and R166Q.

Before proceeding with measuring growth rates of the DHFR mutants, it was important to establish library completeness. To do this, I transformed each sublibrary into E. coli and deep sequenced a 1 mL sample of culture for each sublibrary, replicate, and TYMS background at the start of selection. A given mutant is considered present if the NGS sequencing data returned greater than 10 observations (or counts) of the mutant in the population. Table 3.1 summarizes basic statistics of each TYMS background and replicate. These include: the number of missing mutants ($N_{missing}$),

the number of present mutants ($N_{present}$), and a percentage of total possible mutants that are present (% present).

There were 3,002 total possible single point amino acid mutants in the DHFR saturation mutagenesis library. This comes from taking the number of positions in DHFR (160), excluding the stop and start codons (158), and multiplying by the total number of amino (158 x 20 = 3,160). Lastly the mutations that are redundantly lableled as WT at each position in the sequence (I2I, M42M, G121G, etc.) are removed (3,160 – 158 = 3,002). Across all TYMS backgrounds and replicates, more than 95.9% of mutants were present. Notably, the percent of the libraries that were present are much greater in the first two replicates of the WT and R166Q TYMS backgrounds: at least 99% of mutants are present. In the first two replicates of the library with a Q33S TYMS background, this was only marginally lower with 96.87% mutants present.

Figure 3.1A visualizes mutant counts in libraries from replicate 3, the replicate with the lowest percentage of present mutants, as heat maps. Across all three TYMS backgrounds, most mutants had between 100-1000 counts. For R166Q TYMS, the number of counts was more uniform. In WT TYMS, there were mutants with fewer counts represented by darker blue pixels (10-100 counts). In the Q33S TYMS heatmap, the pattern was less homogenous with even more dark blue and black pixels. The histograms of these mutant counts show the distribution of mutant counts in each TYMS background (Fig. 3.1B). The mean mutant count in each distribution were: 2413 for WT TYMS, 1827 for Q33S TYMS, and 2435 for R166Q TYMS. The median mutant count in each distribution was: 1653 for WT TYMS, 983 for Q33S TYMS, and 1726 for R166Q TYMS.

The Q33S TYMS library contained the greatest number of missing mutants, ranging from 49 to 123 mutants with fewer than 10 counts. The missing mutants (black pixels) in the heatmap in Figure 3.1A were not systematically dispersed along a particular position nor amino acid mutation. Thus, I opted against supplementing these individual mutants into the library. I chose to continue using this library because a vast majority of the mutants were present (95%) and likely to have at least 100 counts.

3.3 Measuring growth rates of mutants in a mixed library.

3.3.1 The NGS-Fit Assay measures growth rates of individual alleles in a mixed mutant library.

After assessing library completeness, I measured the growth rate effects of each mutant in the DHFR library in the context of each TYMS variant with the NGS-Fit Assay (Fig. 3.2). This assay works by sampling the population of each sublibrary over time during the selection phase of the experiment, and using sequencing to quantify the frequency of each allele. During the selection phase, the culture is maintained at a constant density and volume with the turbidostat, a continuous culture device. When the culture exceeds a target OD600 of 0.15, fresh media is automatically dispensed to dilute the culture. To maintain a constant volume, excess culture is removed as waste. In addition, the Reynolds Lab turbidostat can maintain up to 15 vials of cultures in parallel. For the data in this chapter, I performed this assay on a total of 36 sublibrary cultures over the course of four separate experimental sessions: one for each combination of four sublibraries, three TYMS

variants, and three biological replicates. I sampled the culture over the course of selection at 6 time points. These samples were pelleted, frozen, and then lysed in water. Each cell lysate is a template for generating amplicons with two rounds of PCR. An amplicon is generated for each sample from two rounds of PCR. Amplicons are DNA sequences that contains the subset of DHFR to be deep sequenced. In this case, each amplicon spanned the relevant 40 amino acid positions in DHFR in a given sublibrary. In each of the three NGS runs I prepared samples for, the relevant amplicons were equally mixed and then deep sequenced with 300-cycle run on an illumina HiSeq machine by the Sequencing-Only service at GeneWiz. The output of the NGS run is a collection of FASTQ files, each file corresponding to a sample taken during the experiment. This FASTQ file contains a list of sequences, or reads. Each read was quality score filtered and then mapped to a DHFR mutation. After processing each FASTQ file, the output was a raw count of mutants (and WT) for each sublibrary and replicate at each time point.

3.3.2 How was growth rate computed from mutant counts?

From the NGS-Fit Assay sequencing data, mutants and the WT were counted at 6 time points during the experiment at 0, 4, 8, 12, 20, and 24 hours. First, a relative frequency was computed for each mutant at every time point (equation 3.1).

$$f(t)_{mut} = log_2\left(\frac{N_{mut,t}}{N_{WT,t}}\right) - log_2\left(\frac{N_{mut,t=0}}{N_{WT,t=0}}\right) \qquad \text{(equation 3.1)}$$

The relative frequency simplifies interpretation of how fit an allele is in the population compared to the WT. Across the entire time course, the relative frequency of WT is set to zero. At t = 0, all mutants start out with relative frequencies of zero as well. When $f(t)_{mut}$ is positive, the mutant

is present in the population at greater numbers than WT. When $f(t)_{mut}$ is negative, the mutant is present at a lower numbers than WT. Figure 3.2.6 shows the trajectory of relative frequencies of two example mutants clearly: The pink mutant's relative frequency drops with each subsequent timepoint, showing that it is out competed by other variants in the population that are more fit (better growing), like the navy mutant.

In order to estimate an accurate growth rate from the time course of relative frequencies, mutants were further filtered by requiring that the mutant must be present at t = 0 and must be present in at least two other time points. The threshold for presence is 10 counts. For mutants that are meet or exceed these criteria, the relative frequencies are fitted with a linear regression. These fits are normalized by sample size or the total number of counts of the mutant at each time point. The slope of this line is the growth rate relative to WT. Figure 3.3 shows example fits for a subset of mutants in the Calibration Curve in each TYMS background. Here, this method of fitting a growth rate qualitatively reproduces the result that mutations in DHFR that are deleterious to bacterial growth (like D27N and G121V) are rescued by R166Q TYMS.

To minimize variation between culture vials in the turbidostat, I normalized the relative growth rates (units per hour) by the growth rates of the bulk culture vials (generations per hour). The resulting normalized relative growth rate is in units of per generation. I then scaled these normalized relative growth rates in the range between 0 (minimum growth rate) and 1 (WT-like growth). We used the relative growth rate of D27N DHFR, the catalytically dead mutant as the minimum growth rate, in equation 3.2 to scale the data.

$$g_{scaled} = \frac{g + g_{D27N}}{g_{D27N}}$$ (equation 3.2)

3.3.3 Reproducibility of relative growth rates between replicates

Fortunately, these criteria for fitting a growth rate did not filter out a significant number of mutants across replicates. For WT TYMS, triplicate relative growth rates were fit for 95.6% of mutants. For Q33S TYMS, relative growth rates were fit for 92.1% of mutants. For R166Q TYMS, relative growth rates were fit for 98.1% of mutants. For each TYMS background, mean relative growth rates and a standard deviation were computed for DHFR across triplicates. For mutants with triplicate growth rate measurements, the reproducibility of these growth rates from replicate to replicate was assessed in the 2-D correlation scatter plots in Figure 3.4. Among every replicate and TYMS background, mutants with deleterious effects on growth rate were noisier. This is an expected, as slow-growing mutants are by definition lower in frequency and thus more subject to statistical counting noise. In particular, Q33S TYMS showed the strongest reproducibility among all three replicates: the slope of the linear relationship between the growth rates of any two replicates was greater than 0.9. The R-squared of the linear regressions between replicates in the Q33S TYMS background ranged from 0.87 to 0.91, indicating strong replicate to replicate reproducibility. In the WT TYMS and R166Q TYMS, the 1:1 reproducibility between replicates were weaker, with R-squared of the linear regressions between replicates in these two TYMS backgrounds ranged from 0.73 to 0.89. I speculated that splitting replicates across different experimental days and separate NGS runs may have non-trivially contributed to weaker reproducibility among replicates in the WT TYMS and R166Q TYMS backgrounds. For the NGS-Fit Assay of the library in the background of Q33S TYMS, the samples were collected during a

single run of the turbidostat and sequenced on a single NGS run. This was not the case for the WT and R166Q TYMS libraries: the replicates were divided among two separate runs. This hypothesis is still untested because I proceeded with analyzing these data because the I was satisfied with the overall result: that the growth rates between replicates are positively correlated with each other.

3.3.4 Assessing selection for DHFR catalytic activity in the Calibration Curve

In each TYMS background, I first analyzed the average growth rates of DHFR mutants in the Calibration Curve (Fig. 3.5). These data act as an internal control on selective pressure on DHFR catalytic activity and epistasis to R166Q TYMS. Figure 3.5 shows that the relationship between growth and DHFR catalytic activity was monotonic in the background of WT TYMS and Q33S TYMS. Mutations that are deleterious to biochemical activity were also deleterious to growth. In the background of R166Q TYMS, I was reasonably reproduce the result that a loss of function TYMS fully or partially rescues these effects on DHFR activity. DHFR mutants with slower catalytic activities are epistatic to each TYMS mutant. Interestingly, the epistasis of several DHFR mutations to TYMS R166Q and Q33S has opposite signs: Q33S TYMS appears to aggravate growth rate effects (negative sign epistasis) and R166Q TYMS buffers these growth rate effects (positive sign epistasis).

3.4 The pattern of fitness effects across the sequence of DHFR in the context of each TYMS variant.

All of the average relative growth rates in the DHFR library in the background of each TYMS variant are represented in the heatmaps in Figure 3.6A. Three features of these heatmaps indicated

that the experiment was successful. The first is that among all three TYMS backgrounds, mutations to the stop codon were either missing or deleterious. The second is that most mutations in the WT TYMS and Q33S TYMS background were either near-WT (white, very light blue, and very light orange pixels) or are deleterious to fitness (medium to dark blue pixels). This is consistent with the expectation that few mutations will be beneficial relative to wildtype. The third feature is that positions previously established as critical to the biochemical function of DHFR were sensitive to mutations. For example, the DHFR position D27 is essential for the hydride transfer step of the catalytic cycle. In the WT TYMS background, all mutations except for D27E were deleterious to fitness. Additionally, key positions in the loops that undergo conformational changes associated with catalysis — the Met20 loop at positions 9-24, the F-G loop at positions 116-132, and the G-H loop at positions 142-150 — were sensitive to mutations in the WT TYMS and Q33S TYMS backgrounds.[42]

These heatmaps were useful in comparing the effect of each TYMS background on the DHFR saturation mutagenesis library. It is clear that R166Q buffered the mutational effects of DHFR. Most mutations with deleterious fitness effects were partially or fully rescued in the context of a loss of function TYMS. The effect of Q33S TYMS on the mutations in DHFR library was less obvious. Mutations at positions that were key to DHFR catalytic function are similarly deleterious to growth. However, some positions where mutations had neutral effects on fitness were more positive in the context of Q33S TYMS.

The same data in the heatmaps were also represented as histograms fitted with a double gaussian distribution (Fig. 3.6B, Table 3.2). In all three TYMS backgrounds, the distribution of DHFR mutational effects on fitness were divided among two modes: Mode 1 was centered near 1 with WT-like growth. The second, much smaller, Mode 2 was centered around a much slower growth rate at ~0.5. The height of each peak varied among each TYMS background. When TYMS is catalytically inactive (R166Q), the majority of mutational effects were in the first mode, where DHFR is very robust to mutations. Deleterious mutations were rarest in the background of R166Q TYMS. When TYMS is fully active in the WT form or moderately active with the mutation, Q33S, the distribution of fitness effects shifted such that the height of Mode 1 was nearly halved. In the backgrounds where TYMS was active, DHFR was much more sensitive to mutations. Mode 2 in the WT and Q33S TYMS backgrounds were more flat and spanned a larger range of growth rates than Mode 1. The distributions revealed that slightly more mutants had neutral or positive fitness effects with Q33S TYMS than in the WT TYMS background.

How TYMS modulates the growth rates of each individual mutant in the DHFR library is represented in the correlation scatter plots in Figure 3.7. Broadly, mutations in DHFR had similar fitness effects in both the WT and Q33S TYMS backgrounds (Fig. 3.7B). However, in the R166Q TYMS background, there was no correlation to the growth rates in the WT TYMS background (Fig. 3.7A). Instead, the mutations in this correlation plot clearly clustered into two groups: The first group of mutants had neutral or beneficial fitness effects, regardless of the state of TYMS. The second group contains mutants that are deleterious to DHFR in the background of the WT TYMS and are partially rescued when paired with the inactive R166Q TYMS. In this analysis, I

showed that pattern of growth rates of the DHFR mutants changes in the context of different TYMS activities.

In summary, I mapped the fitness effects of all possible single point mutations in DHFR in the context of three different alleles of TYMS. I observed that in the context of a fully functional WT TYMS, DHFR was more sensitive to mutations, particularly at positions that are involved in the catalysis of DHF to THF. This pattern of growth rate effects was largely preserved in the context of the moderately active Q33S TYMS mutant. However, we observed that in the context of a catalytically inactive TYMS, DHFR was remarkably robust to mutations. The sequence constraints of DHFR varied across different alleles of the same enzyme, TYMS.

3.5 The patterns of epistasis in DHFR to two different alleles in TYMS differ in both magnitude and sign.

Epistasis describes how the effect of combinations of mutations differs from the effects of the mutations considered independently. Prior to my thesis work, epistasis between DHFR and TYMS was measured for a handful of alleles (Fig. 2.3). The growth rate measurements described in 3.4 are sufficient to now compute and analyze epistasis across the entire amino acid sequence of DHFR.

We computed epistasis using with equation 3.3, where G is the relative growth rate, a is the single mutation in DHFR, b is a single mutation in TYMS, ab is the double mutant. For every DHFR

mutant in the saturation mutagenesis library, we computed epistasis twice, once for R166Q TYMS and once for Q33S TYMS.

$$Epistasis = G_{ab} - G_a * G_b \quad \text{(equation 3.3)}$$

Under the conditions of my assay (50 ng/ml thymidine), the individual TYMS mutations do not effect growth rate. Given that WT growth rate is normalized to one, this simplifies epistasis to equation 3.4 below.

$$Epistasis = G_{ab} - G_a \quad \text{(equation 3.4)}$$

The epistasis data were organized into heatmaps that show how DHFR, at the amino acid sequence level, was epistatic to two different alleles of the same enzyme (Fig. 3.8). Within each heatmap, the pattern of epistasis was heterogenous, with both negative and positive epistasis. In both TYMS backgrounds, a majority of DHFR mutations were not epistatic (Fig. 3.9). In the Q33S TYMS background, we observed both positive and negative epistasis (Fig. 3.9A). Most mutants within the first 40 positions in the active site of DHFR were negatively epistatic, or were less fit in the background of Q33S TYMS. Conversely, positions 50-60 of DHFR, most mutations were positively epistatic, or were buffered in the background of Q33S. In the background of R166Q TYMS, mutations with positive epistasis were much more pervasive and negative epistasis was much rarer (Fig. 3.9B). By comparing these two heatmaps side-by-side, it is clear that (1) the magnitude of epistasis to R166Q TYMS was greater and (2) for positions in DHFR that are

involved in the biochemistry of the enzyme, Q33S TYMS was slightly negatively epistatic and R166Q TYMS was positively epistatic. This flipped sign epistasis was prevalent in the first 40 positions of DHFR near the active site. This included positions that form the Met 20 loop, which undergoes conformational changes to hold on to and release substrate and co-factor during the catalytic cycle. These first 40 positions also includde D27, which as discussed earlier, is essential in the hydride transfer step of in the redox reaction that DHFR performs. Another instance where epistasis is in opposite directions in Q33S TYMS and R166Q TYMS is G121 in the F-G loop, which forms hydrogen bonds to the Met 20 loop, holding the active site in the *closed* state. The mutation, G121V works at reducing DHFR activity by reducing its affinity for the co-factor, NADPH.[42] The heatmaps themselves reveal a pattern of epistasis that can be analyzed in the context of the structure of DHFR.

3.6 The structural distribution of epistasis in DHFR to TYMS.

3.6.1 Categorizing positions in DHFR by their epistasis using a simple K-means Clustering algorithm

Next, we sought to map the pattern of epistasis to the structure. This required some measure that summarized the effect of all 20 mutations at a single protein position. To accomplish this, we carried out K-means clustering of each DHFR position according to the profile of epistasis across all substitutions. To start, we represented each DHFR position (in each TYMS background) as a vector of epistasis measurements. We computed the pairwise “distances” between all pairs of DHFR positions and TYMS backgrounds. The goal here is to assess how similar or different the

epistasis is in a pair of DHFR positions. In one TYMS mutant, each DHFR position has a 20 amino acid long vector of epistasis measurements (this is a single column in the epistasis heatmap in Fig. 3.8A). Prior to computing distance, we concatenate the matrices of epistasis in each TYMS background. The length of this heatmap is now twice the length of the number of positions in DHFR (L). For every pair of positions, i *and* j in L, we pull the vectors of epistasis E_i and E_j, respectively. The vectors are sorted in descending order of magnitude, ignoring amino acid identity. Not all positions will have epistasis measurements for every amino acid mutation. Thus, these "null" mutants are removed from the each vector. Then longer vector is trimmed to match the length of the shorter one. The distances were computed by subtracting these two vectors sorted and trimmed from each other.

Additionally, we initialized four seed vectors, each with the length of 20 amino acids, each represent a category centered by an empirical epistasis value. These clusters and their epistasis values are: no epistasis at 0, negative epistasis at -0.05, positive epistasis at 0.05, and strong positive epistasis at 0.15. We computed a distance between each seed vector and epistasis vector at each positions in L. These distances to the seeds were appended to the distance matrix, D, are four seed vectors.

This final matrix is a square matrix with the length of L (318), plus 4 for each epistasis cluster. This gives a matrix with the dimensions of 322 by 322. Along the diagonal of this matrix are the first 159 pairs of positions, which represent distances between epistasis vectors within Q33S TYMS. The next 159 pairs of positions represent distances between epistasis vectors within

R166Q TYMS. The last 4 positions along the matrix is the distance between each position and the seed cluster.

In the first iteration of the K-Means Clustering Algorithm, each position in L is assigned to a cluster based on the minimum distance between the epistasis vector and the seed vector. In the next iteration, each position, *k*, we determine whether it needs to be reassigned to a different cluster. To determine if *k* needs to be reassigned, we compared the average distance between *k* and its 10 nearest neighbors in each cluster. The position, k, gets assigned to the cluster with the minimum average distance of each of these four values. At the end of this iteration, all positions in L may or may not be reassigned. We repeated step of the algorithm 14 more times. Note that we observed that no positions switched to another epistasis clusters after 8 iterations. After categorizing these epistasis data into these clusters with this K-Means Clustering Algorithm, the make-up of each clusters in each TYMS mutant background are described in Tables 3.3.1 and 3.3.2. In both Q33S and R166Q TYMS mutant backgrounds, the clusters with the largest number of DHFR positions are neutral epistasis (59 in Q33S TYMS, 68 in R166Q TYMS) and positive epistasis (71 in Q33S TYMS, 70 in R166Q TYMS). No positions in the R166Q TYMS background were in the negative epistasis cluster, while 22 positions in the Q33S TYMS background were in the negative epistasis cluster.

3.6.2 The structural pattern of epistasis between DHFR and TYMS

In the Q33S TYMS background, I focused on three clusters of positions with negative epistasis, positive epistasis, and neutral epistasis (Fig. 3.10). The negative epistasis cluster contained the

fewest number of positions and was mostly localized within the active site of DHFR (Fig. 3.10A-B). The positive epistasis cluster contained the largest number of positions. These positions formed a physically contiguous unit that is predominately localized in the adenosine binding domain of DHFR (Fig. 3.10E-F). A subset of positions were within the back of the active site and away from the loops that undergo conformational change during catalysis. DHFR positions with no epistasis are distributed along the entire structure of DHFR (Fig. 3.10C-D). In this cluster, the positions that lie at the surface of the enzyme surround the core residues in the active site.

The structural pattern of epistasis in the R166Q TYMS background was much different than in the Q33S TYMS background. Positions with positive epistasis were pervasive, covering positions across in both the active site and adenosine binding domain of DHFR (Fig. 3.11 C-D). The positions in the super positive epistasis cluster were concentrated in the core of the active site and include residues essential for catalysis (Fig. 3.11A-B). Remarkably, 9 of these positions (4, 16, 19, 27, 34, 120, 121, 132, and 147) are in both the negative epistasis cluster in the Q33S TYMS background and super positive epistasis cluster in the R166Q TYMS backgrounds. Mutations in these 9 positions are deleterious to bacterial growth and even more so when TYMS is moderately active. A non-functional TYMS fully buffers the effect of these active site mutations. From these data, we can conclude that epistasis between these two enzymes is primarily within the active site of DHFR. Furthermore, the direction of epistasis is dependent on the activity of TYMS.

3.7 Conclusions

Our experiment is not the first instance where the effects of double mutants spanning a pair of subsequent metabolic enzymes was studied. In 1986, a study from Dykhuzien et al. used metabolic control theory to model the relationship between the relative activities in two lactose metabolic enzymes and their effects bacterial growth rate.[43] Though their measurements are sparse relative to technology in the late 2010s, their model appears to fit this relationship between enzymatic activity and growth well. Here, we performed this experiment on a much larger scale with far more mutants. It is theoretically possible to test if their model of metabolic control theory is relevant to our data set.

The observation that epistasis between genes varies among different alleles of the same gene is not new.[44] Here, Xu et al. used Flux Balance Analysis to assess how the experimental measurements of epistasis among *S. cerevisiae* metabolic genes respond to simulated perturbations in flux. They found that epistasis is allele-specific, and that epistasis measurements between a given gene pair is dynamic, or is sensitive to simulated perturbations in metabolic flux. This idea is recapitulated in our data, where DHFR varies within the structure of the enzyme and varies among the two TYMS mutant alleles.

This chapter describes how I used deep mutational scans of DHFR to study the structural pattern of epistasis to TYMS. Here I observed that the pattern of epistasis in DHFR is heterogenous and appears to be dependent on the functional state of TYMS. DHFR is remarkably robust to mutations in the context of a non-active TYMS but is much more sensitive to mutations in the context of active TYMS. The strongest signal of epistasis in DHFR to either TYMS mutant is localized within

the active site. In 9 positions within the active site, the sign of the epistasis is in opposite directions. In the background of Q33S TYMS, we observe negative, aggravating epistasis in the active site and among these sign-flipping positions. In the background of R166Q TYMS, the epistasis is strongly positive, buffering the effect of mutations at these sites.

3.8 Figures

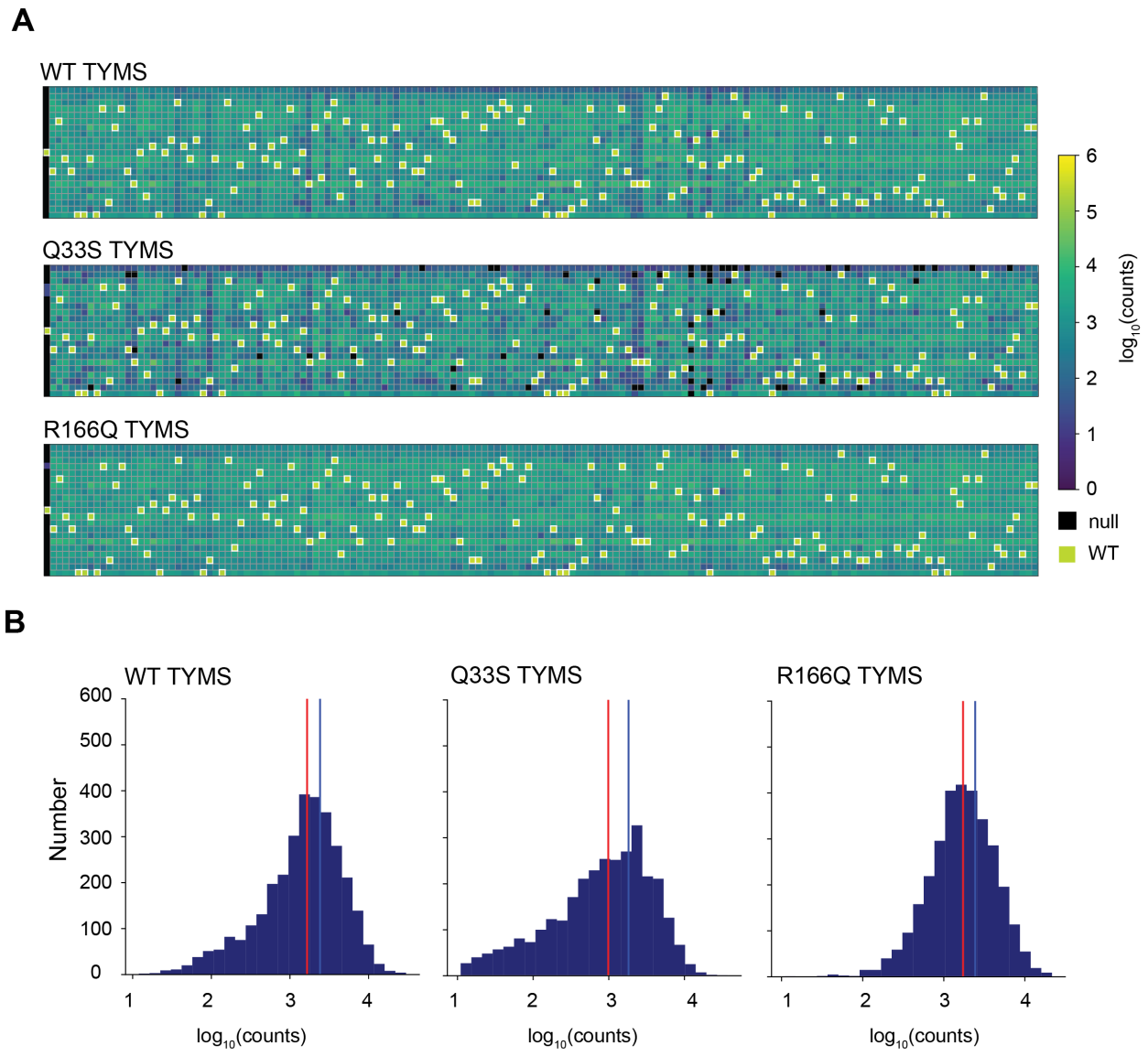

Figure 3.1 Completeness of DHFR saturation mutagenesis libraries. Log_{10} - normalized mutant counts of DHFR libraries in each TYMS background from t = 0, replicate 3. (A) Heatmap of mutant counts of DHFR saturation mutagenesis libraries in each TYMS background (top - WT TYMS, middle – Q33S TYMS, bottom – R166Q TYMS). X-axis from left to right represent positions in the DHFR amino acid sequence (stop codon, position 160 is excluded). Y-axis

represent amino acids, in alphabetical order, with Ala on the bottom and stop codon on the top. Black pixels represent mutants with fewer than 10 counts. Pixels outlined with a lime green box represent WT position (e.g. I2I, M42M, D27D, G121V, etc.). (B) Distributions of $\log_{10}$ normalized mutant counts in each DHFR library. The red lines mark the mean of the distribution (2413 - WT, 1827 - Q33S, 2435 - R166Q). The blue lines mark the median of the distribution (1653 - WT, 983 – Q33S, 1726 – R166Q).

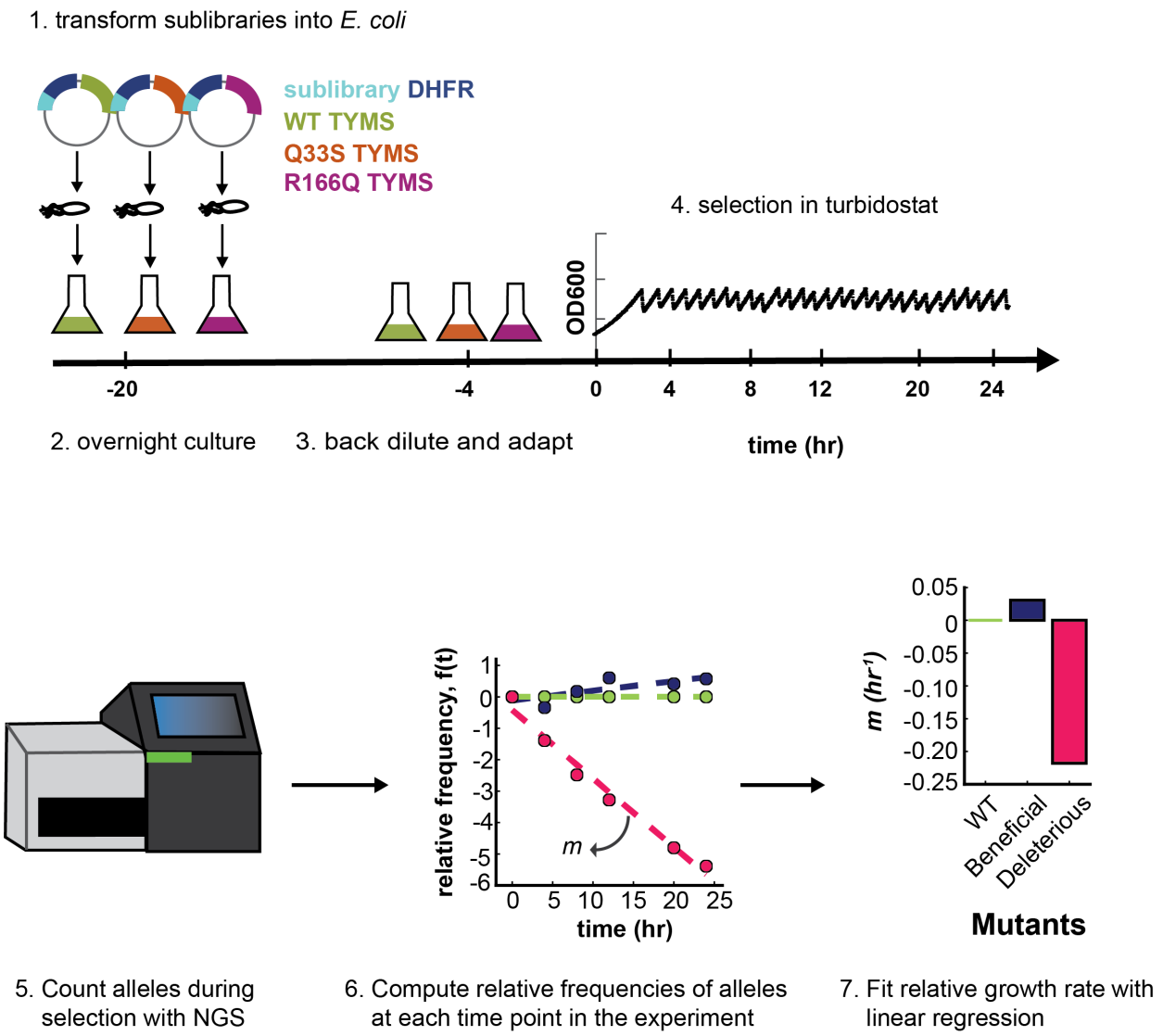

Figure 3.2 Experimental workflow for DMS of DHFR. (1) Each culture in the experiment contained one of the four sublibraries of the DHFR saturation mutagenesis library (cyan part of the plasmid – sublibrary, navy blue part of the plasmid – rest of the DHFR coding region) paired in the background of TYMS variant (green – WT TYMS, orange – Q33S TYMS, purple – R166Q TYMS). This plasmid sublibrary was transformed into ER2566 ΔfolA ΔthyA by electroporation. (2) Afterward, the culture is grown overnight in minimal media supplemented with 50 µg/mL

thymidine and 30 μg/mL chloramphenicol (see 2.5.5) at 37°C. (3) In the morning, the culture is back-diluted to an OD600 of 0.1 in the same media. This back-diluted culture is grown at 30°C for four hours to adapt to the selection temperature. (4) Afterwards, 1 mL of the culture is sampled, pelleted, and frozen at -20°C. This is the first time point of the experiment. Using the same media, the adapted culture is back-diluted to an OD600 of 0.1 and transferred into the turbidostat which will maintain the cultures at a target OD600 of 0.15 for 24 hours at 30°C. Here, I show an example OD600 time course in the turbidostat. 1 mL of the culture is sampled, pelleted and frozen at -20°C after 4, 8, 12, 20, and 24 hours. (5) To prepare the sample library for deep sequencing, the cell lysates of the samples were used as templates for synthesizing amplicons of the sublibrary region of DHFR by PCR (see section 3.9.2). (6) The output of this experiment are counts of each mutant over the course of the experiment, which are normalized by WT to compute a relative frequency. This value describes the trajectory of a mutant during the experiment in relation to WT. A mutant with a positive relative frequency has a larger number of alleles in the population than WT (green). A mutant with a negative relative frequency shows that its allelic population is depleting (pink). Mutants with WT-like effects on fitness will show a flat trajectory at zero (navy). (7) This time course is then fitted with a linear regression. The slope of this fit (m) is the growth rate of the mutant, relative to WT.

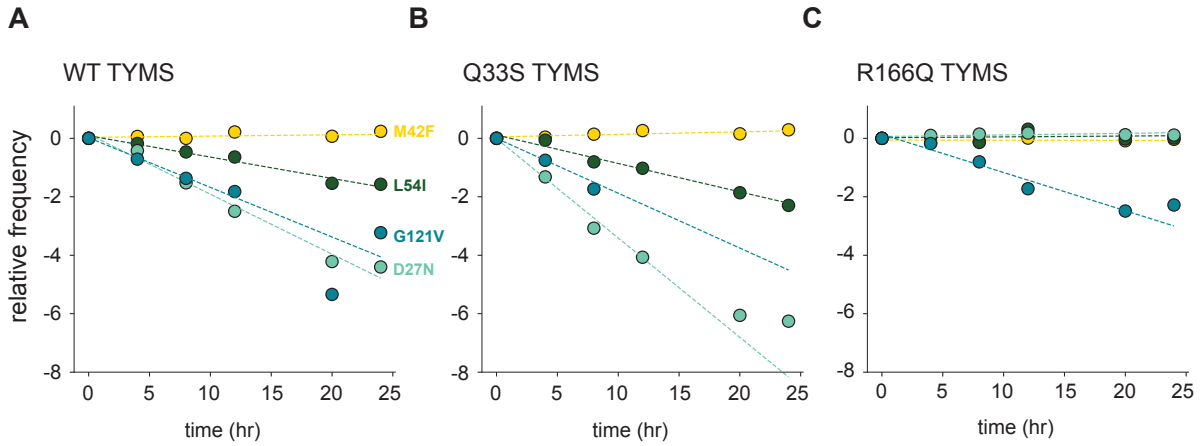

Figure 3.3 Examples of relative growth rate fits in a subset of the Calibration Curve in replicate 3. A-C show relative frequency time course traces for M42F (yellow), L54I (dark green), G121V (teal), and D27N (mint green) in the background of WT TYMS (A), Q33S (B), and R166Q TYMS (C). The markers are relative frequencies computed from experimental mutant counts using equation 3.1. The dashed lines are fits of the data with ordinary least squares linear regression. These fits are normalized by the sample size (the total, non-normalized mutant counts across time points).

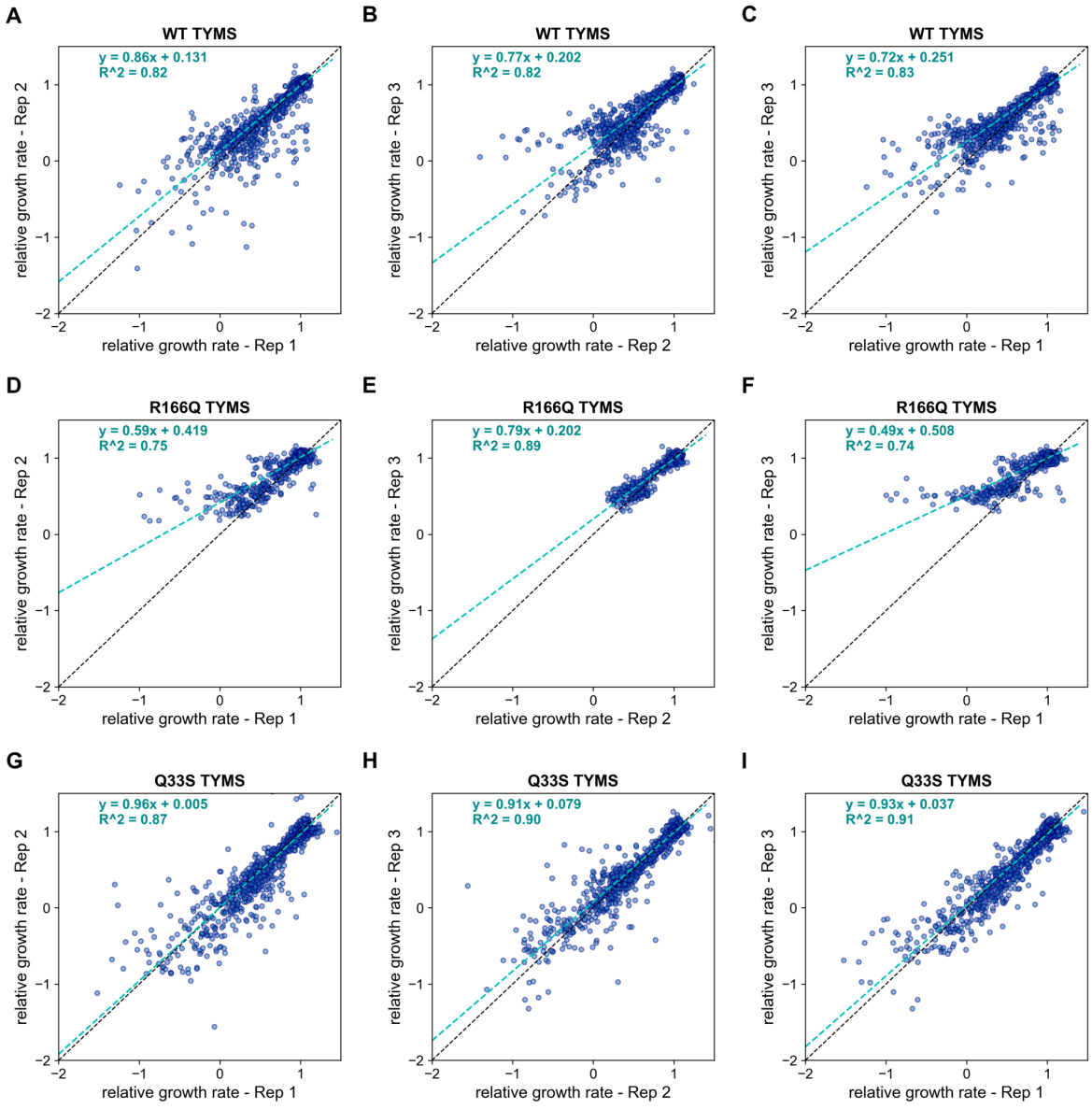

Figure 3.4 Reproducibility of relative growth rates between biological replicates. All correlation scatter plots here compare relative growth rates of DHFR mutants in a pair of replicate measurements (navy blue markers). In each plot, a linear regression was fitted (dashed cyan line). The linear equation of these fits and R-squared are reported on the plot itself. The black dashed

line is a perfect correlation: Y=X. TYMS backgrounds among replicate measurements of DHFR relative growth rates are consistent: (A-C) WT TYMS, (D-F) R166Q TYMS, (G-I) Q33S TYMS.

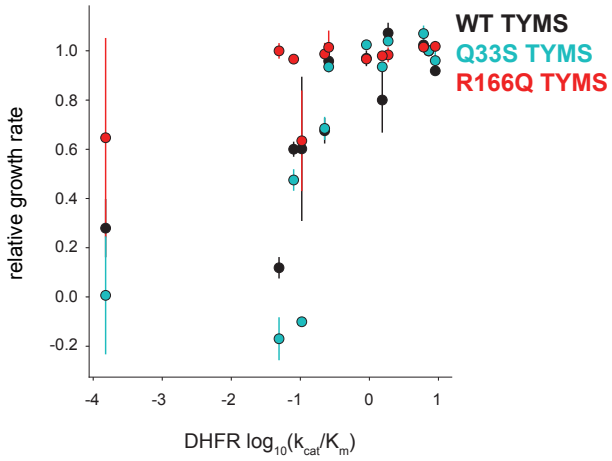

Figure 3.5 The Calibration Curve of DHFR point mutations in the background of each TYMS variant. The relationship between DHFR catalytic activity and relative growth rates in the background of TYMS variants: WT – black, Q33S – cyan, R166Q – red. Error bars are standard deviations.

The relationship between growth and DHFR catalytic activity is monotonic in the background of WT TYMS and Q33S TYMS. Mutations that are deleterious to biochemical activity are also deleterious to growth. In the background of Q33S TYMS, mutants with moderate and deleterious catalytic activities correspond are aggravating to growth rate. In the background of R166Q TYMS, we reasonably reproduce the result that a loss of function TYMS fully or partially rescues these effects on DHFR activity.

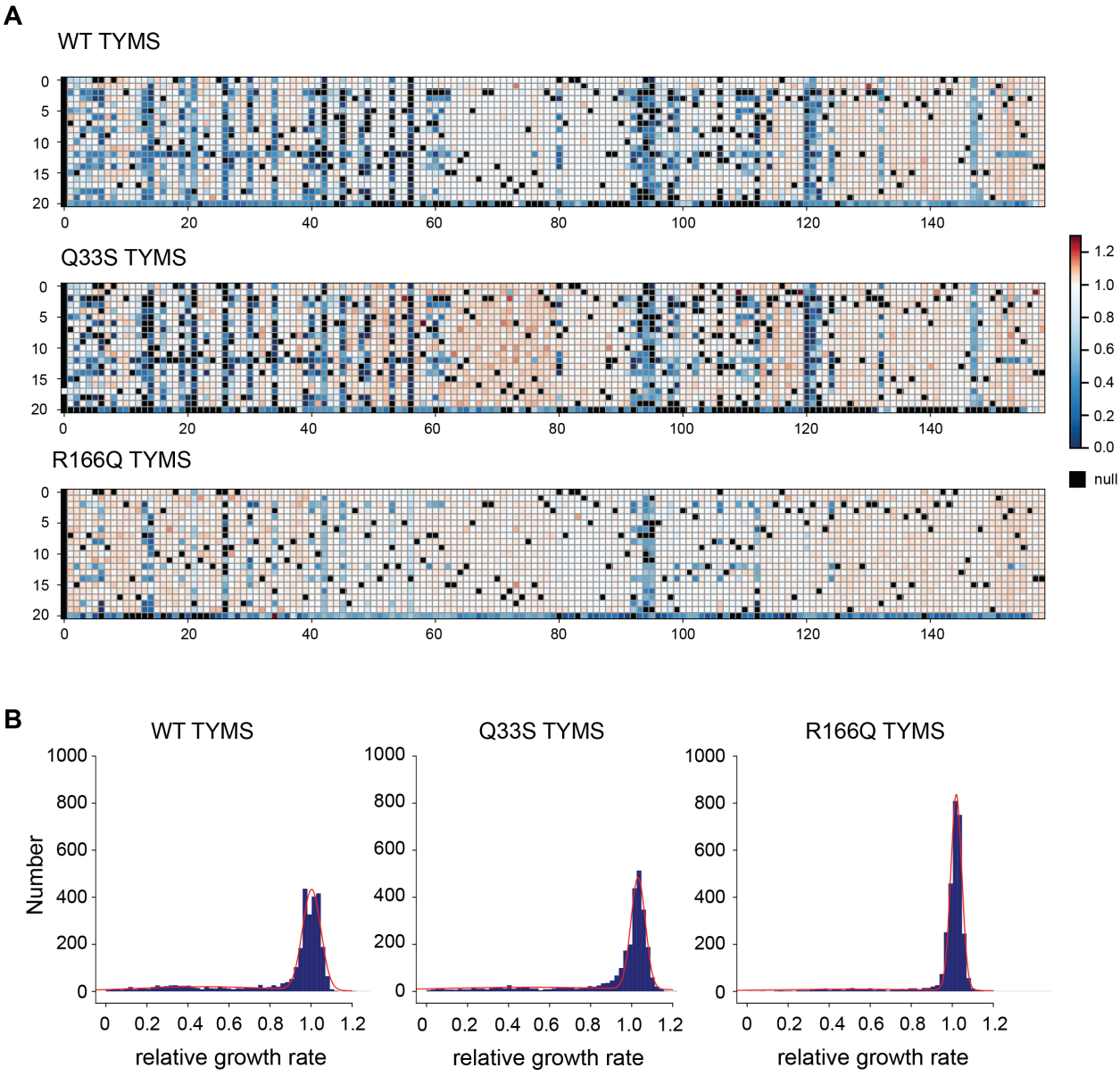

Figure 3.6 Two representations of average relative growth rates of mutants in the DHFR saturation mutagenesis library. (A) Heatmaps of average relative fitness of DHFR in the WT TYMS (top), Q33S TYMS, and R166Q TYMS backgrounds. The amino acid positions of DHFR are along the x-axis. All possible amino acid mutations are on the y-axis (Ala is on the top, stop codon is at the bottom). Black pixels indicate a mutant that is not present. Mutations with white

pixels had neutral effects on growth. Mutations with blue pixels are deleterious to growth. Red pixels indicate beneficial growth rate effects. (B) The distributions of average relative growth rates in WT TYMS (left), Q33S TYMS (middle), and R166Q TYMS (right) backgrounds. The red line are double gaussian fits of the distributions (See Table 3.2 for double gaussian parameters).

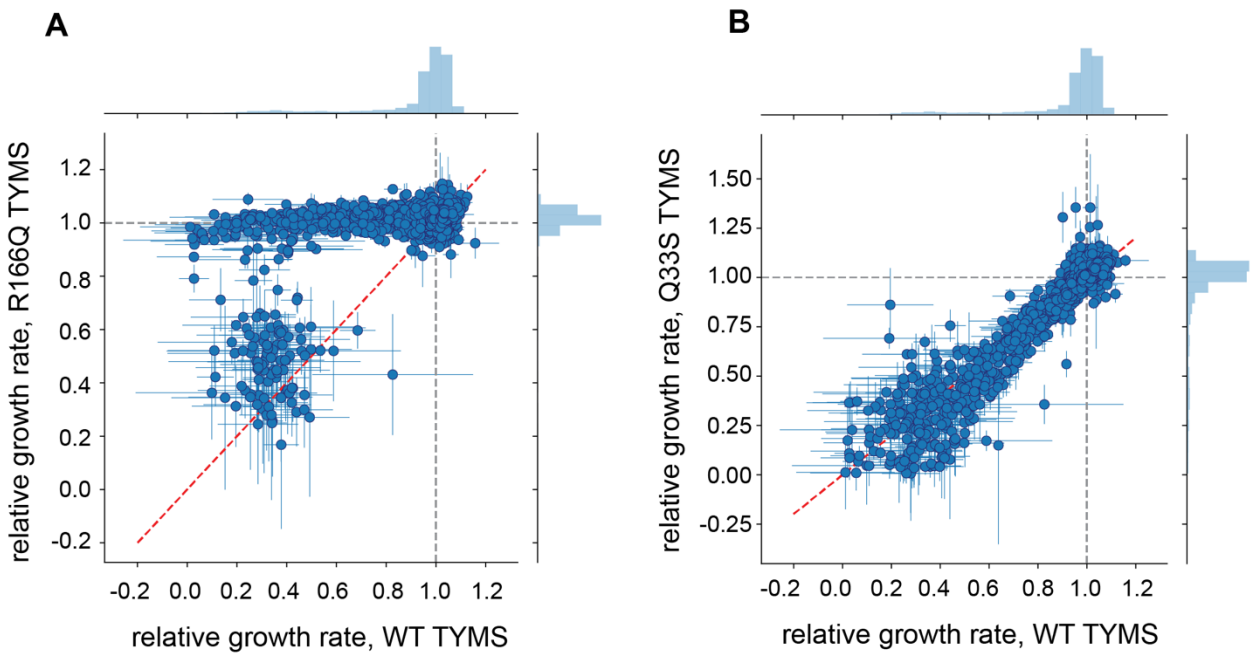

Figure 3.7 Comparing the effect of a mutation in TYMS on relative growth rates of DHFR mutants. Blue markers are mean relative growth rate from average of triplicate measurements. Error bars are standard deviations. Red dashed line is a perfect correlation (Y=X). The dashed grey lines on 1 on both axes mark the growth rate of WT DHFR. Relative growth rates of DHFR mutants in the background of WT are on the x-axis. A histogram of these growth rates are above the scatter plots. (A) Relative growth rates of these same DHFR mutants in the context of R166Q TYMS on the y-axis. A histogram of these data are on the right of the scatter plot. (B) Relative growth rates of these same DHFR mutants in the context of Q33S TYMS on the y-axis. A histogram of these data are on the right of the scatter plot.

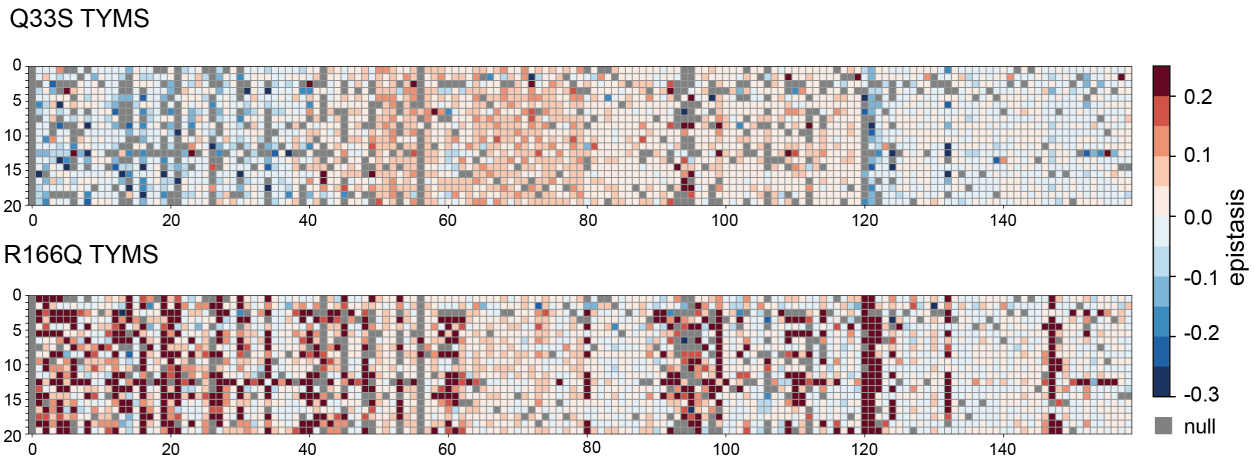

Figure 3.8 The pattern of DHFR epistasis to Q33S TYMS (top) and R166Q TYMS (bottom). Epistasis could not be calculated for mutations indicated in grey. DHFR mutants with white pixels have neutral epistasis. Mutants with red pixels were positively epistatic to the TYMS mutant. Mutants with blue pixels were negatively epistatic to the TYMS mutant.

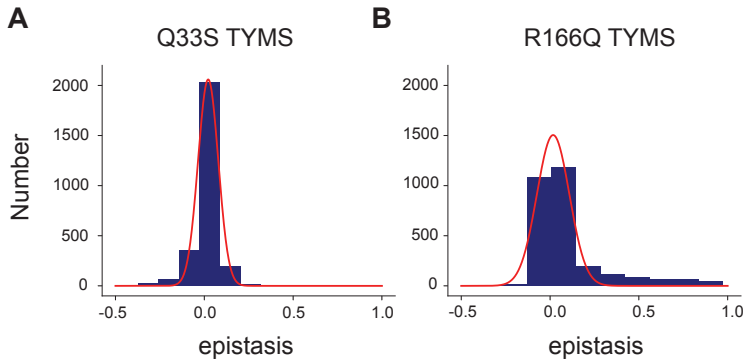

Figure 3.9 The distribution of DHFR epistasis to two TYMS variants. A single gaussian was fitted to each distribution (red line). (A) Epistasis distribution in the Q33S TYMS background. The parameters of the single gaussian are: peak height = 2059, mean = 0.022, sigma = 0.057 (B) Epistasis distribution in the R166Q TYMS background. The parameters of the single gaussian are: peak height = 1504, mean = 0.017, sigma = 0.092.

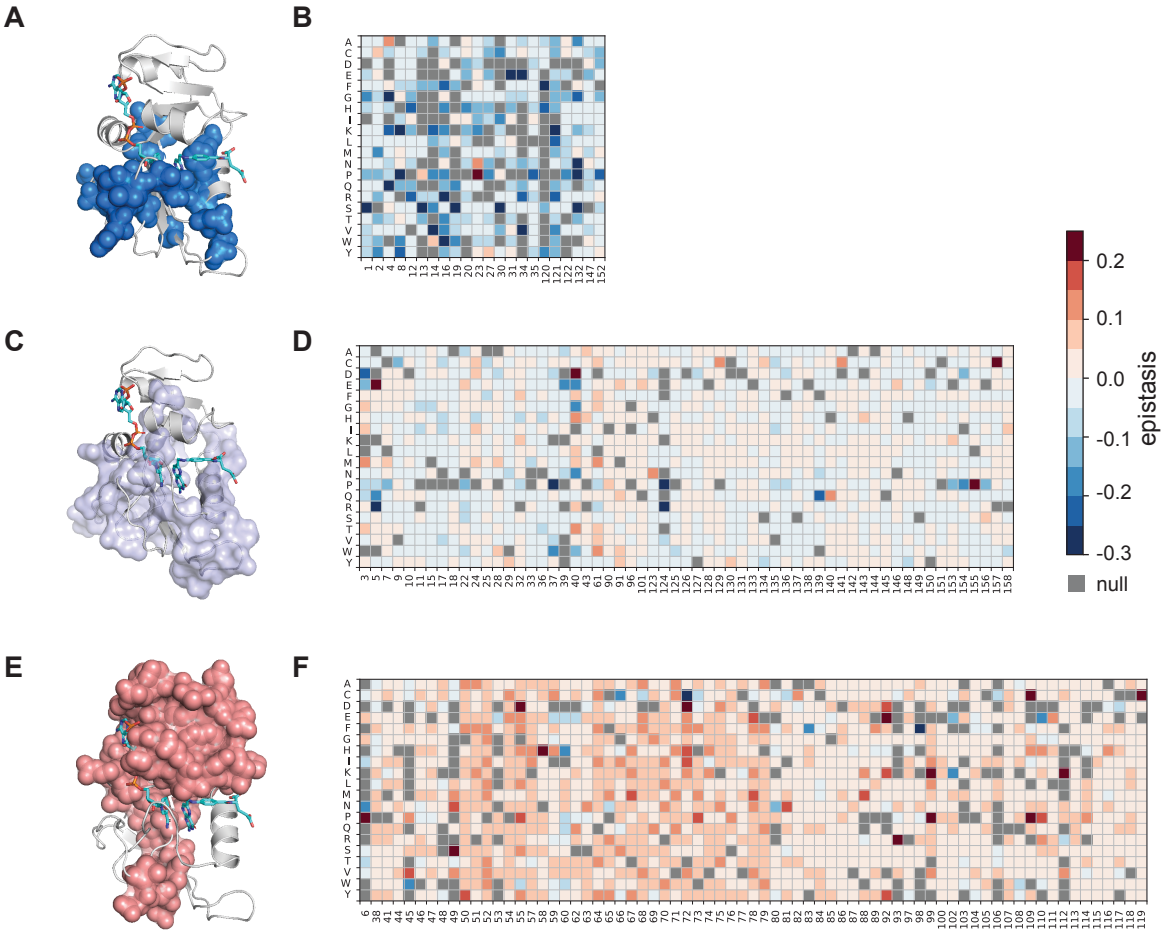

Figure 3.10 Epistasis to Q33S TYMS can be categorized into three clusters of DHFR positions. (A-B) negative epistasis, (C-D) neutral epistasis, and (E-F) positive epistasis). (A,C,E) The structure of DHFR (PDBID: 1RX2) is in a grey ribbon. The substrate and co-factor, DHF and NADPH, respectively are in represented as cyan sticks. (A) DHFR positions with negative epistasis in blue spheres. (C) DHFR positions with positive epistasis in pink spheres. (E) DHFR positions with neutral epistasis in light purple spheres. (B,D, and E) The epistasis heatmap from Fig. 3.7A is grouped according to positions clustered by the K-means clustering algorithm.

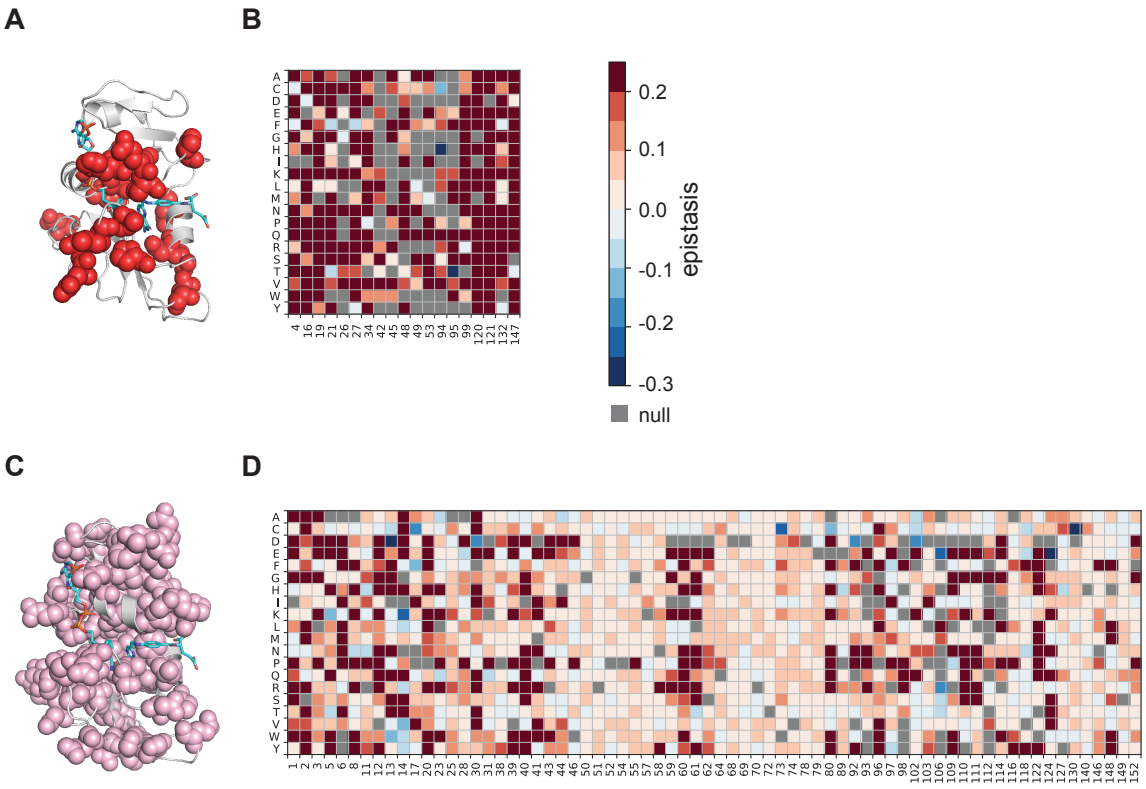

Figure 3.11 Epistasis to R166Q TYMS can be categorized to two groups: (A-B) super-positive epistasis and (C-D) positive epistasis. (A,C) The structure of DHFR (PDBID: 1RX2) with epistasis clusters in red (A) or pink (C) spheres. (B,D) The epistasis heatmap from Fig. 3.7A is grouped according to positions clustered by the K-means clustering algorithm.

3.9 Materials and Methods

3.9.1 Sub-cloning the saturation mutagenesis library

The saturation mutagenesis of DHFR was performed by Samuel M. Thompson.[33] After receiving the sub-libraries of DHFR, restriction digest and ligation was used to clone each sub-library into a pTet-Duet plasmid upstream of TYMS. The entire DHFR coding region containing restriction sites, NotI and EcoNI was amplified by PCR. These library inserts and target plasmids were double digested with NotI and EcoNI for 3 hours at 37°C. The digested plasmid was treated with Antarctic phosphatase for 1 hour at 37°C. The DHFR insert and treated plasmid were ligated with T4 DNA ligase overnight at 16°C. The concentrated ligation product was then transformed into *E. coli* XL1-blue (homemade competent cells with a minimum transformation efficiency of $\geq 10^8$ CFU/ µg DNA) by electroporation, and recovered in SOB for 1 hour at 37°C. To estimate library coverage, 20 µL of the recovery culture was diluted 1:10 into 180 µL SOB. This was used for downstream serial dilutions of 1:100 and 1:1000 in SOB. 100 µL of each dilution was plated onto LB agar + 30 µg/mL chloramphenicol, incubated at 37°C overnight. The remaining recovery culture was grown in liquid LB + 30 µg/mL chloramphenicol at 37°C, 220 rpm overnight. In the morning, the colonies on each plate were counted. The remaining recovery culture was grown in liquid LB overnight at 37°C and plasmid purified the following day. Colony counts of the plated dilutions were used to estimate coverage of library mutants in the liquid culture.

3.9.2 NGS sample preparation

Each pellet from the NGS-Fit assay (section 3.3.1 and Fig. 3.2) was thawed and lysed by resuspending the cells with 100 µL dH_2O and incubated at 95°C for 5 minutes. These cell lysates were then spun at 21,130 x g for 10 minutes in a room temperature bench top microcentrifuge. Supernatants containing the plasmids were isolated from the pellet and used for downstream NGS sample preparation.

Each sub-library was sequenced as an amplicon using Illumina TruSeq-HT i5 and i7 indexing primers to identify each sub-library and time-point. The amplicon for each sample is made using two subsequent rounds of PCR. The first round amplified the DHFR coding region of the sub-library (sequences of these primers are in Table 3.5) and the second used Illumina primers to add flanking sequences. This yielded final barcoded products ranging from 298 – 315 bp, depending on the sub-library. The amplicons were then individually quantified using picogreen and mixed equimolarly, with a final target amount of ≥ 2000 ng. This mixture of amplicons, the sample library, was gel-purified. To assess purity, the A260/A80 nm and A260/A230 nm absorbance ratios of the sample library were measured on a nanodrop. The sample library DNA concentration was measured using the Qubit Assay. This mixed and quantified library was sequenced with a 150 x 2 cycle paired-end Illumina HiSeq at GeneWiz using their Sequencing-Only service.[45]

3.9.3 Computational pipeline for analysis of the Next - Generation Sequencing data

All scripts referenced below are in the path: smb://lamella.biohpc.swmed.edu/project/greencenter/Reynolds_lab/shared/tnn/. Data from NGS for each amplicon contain forward and reverse reads in a pair of FASTQ files. These reads are

overlapping and are merged using the program, USEARCH (see tnn_pythonUCombiner.bsh, splitFastqs_tnnHiSeq.py, and RunUSearch_tnnHiSeq.py).[46] Each read is quality score filtered (Q-Score ≥ 20) and identified as a wild-type (WT) or mutant of DHFR (see monsterBash.bsh and countingAlleles.py). The mutants are then counted and adjusted according to noise by hamming distance to other mutants (JM_hamming_analysis-tnnNotes.ipynb). The python script, 1_CalcGrowthRates.ipynb was used to analyze these mutant counts, fit and analyze growth rates. Epistasis was then computed, analyzed, and then categorized with the K-Means Clustering Algorithm with 2_CalcEpistasis.ipynb.

3.10 Tables

Table 3.1 Library completeness at t = 0

TYMS	Replicate	Nmissing	Npresent	% present
WT	1	5	2997	99.83
WT	2	26	2976	99.13
WT	3	57	2945	98.10
R166Q	1	2	3000	99.93
R166Q	2	13	2989	99.57
R166Q	3	13	2989	99.57
Q33S	1	49	2953	98.37
Q33S	2	94	2908	96.87
Q33S	3	123	2879	95.90

Table 3.2 Parameters of double gaussians of the distributions of relative growth rates

	Mode 1			**Mode 2**		
TYMS	**peak height**	**mean**	**sigma**	**peak height**	**mean**	**sigma**
WT	428	1.003	0.045	20	0.476	0.344
Q33S	474	1.031	0.035	17	0.500	0.522
R166Q	831	1.019	0.027	9	0.5	0.449

Table 3.3.1 Epistasis clusters in the background of Q33S TYMS

Epistasis Clusters	Npositions	DHFR positions
Neutral = 0	59	3+5+7+9+10+11+15+17+18+22+24+25+28+29+32+33+ 36+37+39+40+43+61+90+91+96+101+123+124+125+ 126+127+128+129+130+131+133+134+135+136+137+ 138+139+140+141+142+143+144+145+146+148+149+ 150+151+153+154+155+156+157+158
Negative = -0.05	22	1+2+4+8+12+13+14+16+19+20+23+27+30+31+34+ 35+120+121+122+132+147+152
Positive = 0.05	71	6+38+41+44+45+46+47+48+49+50+51+52+53+54+ 55+57+58+59+60+62+63+64+65+66+67+68+69+70+ 71+72+73+74+75+76+77+78+79+80+81+82+83+84+ 85+86+87+88+89+92+93+97+98+99+100+102+103+ 104+105+106+107+108+109+110+111+112+113+ 114+115+116+117+118+119
Super Positive = 0.15	4	26+42+94+95

Table 3.3.2 Epistasis clusters in the background of R166Q TYMS

Epistasis Clusters	Npositions	DHFR positions
Neutral = 0	68	7+9+10+15+18+22+24+29+32+33+35+36+37+47+63+ 65+66+67+71+75+76+77+81+82+83+84+85+86+87+ 88+90+91+100+101+104+105+107+108+113+115+ 117+119+123+125+126+128+129+131+133+134+ 135+136+137+138+139+141+142+143+144+145+ 150+151+153+154+155+156+157+158
Negative = -0.05	0	n/a
Positive = 0.05	70	7+9+10+15+18+22+24+29+32+33+35+36+37+47+ 63+65+66+67+71+75+76+77+81+82+83+84+85+ 86+87+88+90+91+100+101+104+105+107+108+ 113+115+117+119+123+125+126+128+129+ 131+133+134+135+136+137+138+139+141+ 142+143+144+145+150+151+153+154+155+ 156+157+158
Super Positive = 0.15	19	4+16+19+21+26+27+34+42+45+48+49+ 53+94+95+99+120+121+132+147

Table 3.4 DHFR plasmid sublibraries (pTET-duet, RBS 1)

pTN#	DHFR sublibrary	TYMS
405	SL1	WT
356	SL2	WT
403	SL3	WT
358	SL4	WT
406	SL1	R166Q
360	SL2	R166Q
404	SL3	R166Q
362	SL4	R166Q
409	SL1	Q33S
410	SL2	Q33S
411	SL3	Q33S
412	SL4	Q33S

Table 3.5 Custom primers for amplicon generation.

oTN	Sublibrary, Direction	Primer sequence
118	SL1, FWD	CACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTTAATAATGAGATATACCATG
301	SL1, REV	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNGATTGATTCCCAGGTATG
334	SL2, FWD	CACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCACCTTAAATAAACCCGTG
335	SL2, REV	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNATCAGCATCGTGGAA
336	SL3, FWD	CACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGTGAAGTCGGTGGATG
337	SL3, REV	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNGGAAATGGGTGTCGC
338	SL4, FWD	CACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGCATATCGACGCAGAAGTGG
339	SL4, REV	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNCTTGTCGACGCCTG

CHAPTER FOUR
Discussion and Future Directions

4.1 A structural map of epistasis in a pair of metabolic enzymes

In my thesis work, I sought to address the question of how functional coupling between two metabolic enzymes is encoded within the amino acid sequence and physical structure of one of them. To answer this question, I performed deep mutational scanning assays on the enzyme DHFR while varying TYMS. More specifically, I considered all DHFR single mutations in the context of a wild-type, moderately active mutant (Q33S), and non-functional TYMS variant (R166Q). The result is a rigorous dataset with epistasis measurements over the entire amino acid sequence of DHFR.

This epistasis dataset was analyzed with a simple clustering algorithm to group DHFR positions into discrete epistasis categories. This analysis resulted in two distinct views of the distribution of epistasis in the crystal structure of DHFR, one for each TYMS mutant. In both views, the positions with the greatest magnitude of epistasis lied at the active site. In the context of an active TYMS mutant, the positions in the active site had negative epistasis. When in the background of a non-functional TYMS mutant, these positions in the DHFR active site had strong positive epistasis. Beyond the active site, the distribution of DHFR positions with positive epistasis was also context dependent. In the background of R166Q TYMS, positions with po sitive epistasis formed a concentric shell around the active site. In the background of Q33S TYMS, positions in the positive

epistasis cluster formed a structurally distinct cluster in the adenosine binding subdomain of DHFR, away from the negative epistasis cluster in the active site subdomain.

The finding that epistasis across enzymes is localized at the active site is consistent with the mechanism of positive epistasis, where the matched rates of catalysis in both enzymes prevents the accumulation of their shared intermediate DHF. DHFR appears to be highly robust to mutations in the context of a TYMS that does not catalyze the reaction that synthesizes DHF. Presumably this is because DHF does not accumulate, and TYMS is not present to deplete the pools of reduced folate. In the presence of active TYMS that continues to makes DHF, DHFR appears to be much more sensitive to mutations within the active site. Here, active site mutations reduce the capacity for DHFR to consume DHF and then maintain a flux of reduced folates throughout the pathway.[20] Epistasis was also localized to the active sites in a pair of proteins in bacterial Quorum Sensing that interact through the synthesis and binding a small molecule across different cells in a population.[13] Therefore, we can think of the active site as a non-physical "interface" between protein pairs that do not form a physical complex but share an intermediate.

4.2 A mechanistic model of epistasis between DHFR and TYMS

The prevalent positive sign epistasis we observed in the R166Q TYMS background was consistent with our expectations from prior work.[20] However, the negative epistasis we observed in the background of Q33S TYMS was unexpected. The observed negative epistasis means that in the context of a less active or non-active TYMS, the fitness effect of certain DHFR mutations is more deleterious. A potential hypothesis for this negative epistasis is that these mutations lower DHFR

affinity for DHF (K_m) but not the rate of the catalytic reaction (k_{cat}). The fitness effect of such mutations would be masked in the context of WT TYMS that synthesizes enough DHF to saturate DHFR active sites in the cell. When paired with a TYMS variant that produces less substrate, the deleterious effect of DHFR mutants with lower affinity for DHF on growth rate becomes apparent. This hypothesis can be tested by purifying a random sample of mutants with negative epistasis and then performing steady-state enzyme kinetics assays to measure parameters of catalytic activity: k_{cat}, K_m, and V_{max}.

In prior work, a simple model was able to describe the relationship between DHFR catalytic activity, metabolite abundance, and *E. coli* growth rate.[47] Here, the authors studied three point mutants in DHFR that conferred resistance to the antibiotic, trimethoprim. They measured the effect of cellular protein abundance, catalytic activities, and growth rates of the single, all possible double mutants, and the triple mutant of DHFR across a range of trimethoprim concentrations. They make the assumption that the intra-cellular concentration of DHF is at a steady-state level, which is not true, given the folate metabolomics data from Schober et al. Even so, they were able to fit a Michaelis Menten-like function that accurately predicted *E. coli* growth rate from the effect of trimethoprim on DHFR catalytic activity, DHFR protein abundance, and DHFR stability. This work motivates further development of this model so that it can accurately reflect the non-steady state DHF abundance in DHFR mutants and epistasis with TYMS. Such a model would effectively map how the epistasis between DHFR and TYMS relate to bacterial growth rate and fitness.

More broadly, creating a mechanistic mathematical model of epistasis would provide a framework for interpreting my data, generating hypotheses, and predicting the fitness effects of DHFR mutations in various TYMS backgrounds. To create such a model, future work should characterize an analogous Calibration Curve that relates catalytic activity to growth rate for a larger set of TYMS mutants. One could draw on ideas from Metabolic Control Theory, as discussed below, or potentially try fitting the data to the Goldbeter-Koshland equation, a model for two enzymes that perform cyclic coupled reactions. Ultimately, the goal of a mechanistic model is to comprehensively understand and capture the positive and negative epistasis we observe between DHFR and TYMS. Such a model would be able to predict the growth rate from DHFR and TYMS catalytic activities.

An application of Metabolic Control Theory on a pair of lactose metabolic enzymes provides a framework for developing such a mechanistic model. Metabolic Control Theory (MCT) uses a series of equations to describe the epistasis between enzymes in a metabolic pathway.[48] In brief, the equations of MCT state that the overall flux through an enzyme in a pathway is due to its individual activity and abundance *and* is dependent on the activities and abundances of other enzymes in the pathway. The flux control coefficients, a measure of how the abundance of an enzyme affects the flux, must sum together to equal one. To more concretely understand MCT, let's take a look at an example from 1986 that mapped the relationship between fitness to the activity of two enzymes in *E. coli* lactose metabolism, beta galactosidase and beta galactosidase permease.[43] According to MCT, the flux through the pathway is a function of the ratio of concentrations of substrate and products of the pathway, the enzyme activities defined by

Michaelis-Menten kinetic parameters K_m and V_{max}, and equilibrium constants between the intermediate substrates. Dykhuizen and Hartl defined fitness from measurements of *E. coli* growth rate in a lactose-limited chemostat, a continuous culture device.[43] Under this growth condition, the growth rate is directly proportional to the flux through the permease and lactase enzymes in lactose metabolism. Because the final derived relationship computes a relative fitness, the function becomes independent of substrate and product (these parameters cancel each other out). The MCT equation was used to derive the relationship between the relative fitness of a mutant in a given enzyme and the relative activities of both enzymes in the pathway. *The result is a three-dimensional surface that can predict the relative fitness from the relative growth rates of the lactase and permease.* In this application of MCT, the fitness of a lactase or permease mutant is dependent on the enzymatic activities of both enzymes. This relationship was tested with experimental measurements of the relative activities and relative fitness in a small number of mutations in both enzymes. were characterized with *in vitro* kinetic assays and *in vivo* growth rate measurements. Though the number of experimental measurements were limited to 4 permease mutants and 14 lactase mutants, the model fit the data well.

A major hinderance in applying MCT to model epistasis between DHFR and TYMS is a limitation on high-throughput measurements of catalytic activity. We made thousands of measurements of the effect of mutations in DHFR on growth rate. An analogous DMS of direct and high-throughput biochemical assays of catalytic activity does not yet exist. A way one can circumvent this problem is to randomly sample a handful of mutants in each category of epistasis, protein purify these mutants individually, and then perform steady-state kinetics assays on each individual mutant. This

low-throughput approach could be sufficient for testing the predictions of growth rate in the development of a mechanistic model.

In the Introduction of this Dissertation, I discuss at length about the idea of non-specific epistasis. Briefly, this kind of epistasis is a function that captures global non-linearities in the relationship between the biophysical traits of the mutations and phenotype.[19] The mechanistic model would be the function that describes the non-specific epistasis between DHFR and TYMS. It would describe the relationships between DHFR catalytic activity, TYMS catalytic activity to accurately predict the growth rates in any given combination of mutants.

4.3 Considering the effect of environment and genetic background on measurements of epistasis

Unfortunately, for experimentalists, epistasis is not a static measurement. The magnitude and direction of epistasis are highly context dependent. For example, the severity of the mutation affects the magnitude of epistasis measurements. This idea is supported by the lack of reproducibility between a pair of high-throughput screens of genetic interactions.[49] One screen uses CRISPRi to knockdown transcription. The other screen, synthetic genetic array (SGA), uses temperature—sensitive genetic deletions. Of the 5,072 gene pairs measured in both studies, only 149 have the same measured genetic interaction. This variation between the two data sets suggests that the differences in the epistatic profiles are due to differences in the severity of the genetic perturbations in their respective screens.

The context dependence of epistasis also applies at a smaller scale within a single gene. The dynamic feature of epistasis was initially identified from alleles computationally generated using flux-balance analysis to titrate WT activity through a single gene. These alleles within the same gene had different epistatic effects to each other.[44] Differential epistasis of alleles within the same gene was later recapitulated experimentally using mutations in a transcriptional repressor that altered expression level. Here, epistasis among these mutants in the same gene varied in both sign and magnitude according to expression level.[39]

Given that epistasis itself is a context-dependent measurement, the "correct" experimental conditions for measuring epistasis is subjective to the experimenter. In my case, the selection conditions I chose are far from a natural environment. DHFR and TYMS were expressed from a plasmid at higher levels than in the genome (Fig. 2.4, Fig. 2.5). The saturation mutagenesis library was transformed into a laboratory expression strain of *E. coli,* ER2566 ΔfolA ΔthyA, which lacks Lon protease, an enzyme that some strains of natural *E. coli* use in protein quality control.[50] The selection step was also performed under highly controlled laboratory conditions where both the optical density, temperature, and growth media of the bacterial cultures remained consistent for the entirety of the experiment. One way to address this discrepancy between the selection conditions of a laboratory and natural environments is with the method Phylogenetics informed by Deep Mutational Scanning (phyDMS).[51] PhyDMS first computes the amino acid preferences of the fitness effects in the saturation mutagenesis library and visualizes this profile of this as a seqLogo plot. A MSA of the protein family of interest is then analyzed to generate a likely phylogenetic tree. The amino acid propensities of the sequence under the conditions of natural

selection are statistically inferred from this phylogenetic tree. If these profiles of amino acid site preferences are wildly different, then the experimental condition of the DMS also is far from the conditions under which natural selection occurred. This analysis can be readily applied to the five different DMS assays of DHFR. These include the three TYMS backgrounds described in Chapter 3 and two in the presence and absence of Lon protease in Thompson et al. This analysis would assess which experimental conditions more closely reflects the condition of natural selection in the molecular evolution of DHFR.

4.4 Epistasis dataset to test a model of sequence co-evolution, Positional Mirror Tree

Evolutionary statistics describes an umbrella of statistical methods to infer co-evolution within or between proteins. One of these methods is Mirror Tree, which was designed to identify both physical and non-physical protein-protein interaction pairs by comparing similarities of their respective phylogenetic trees.[52,53] The idea underlying Mirror Tree is that if a pair of interacting proteins undergo similar selection pressures over evolutionary time, their respective phylogenetic trees will be similar to each other. In practice, Mirror Tree identifies whether a pair of proteins are interacting from correlated sequence similarities. The “interaction score” is a Pearson correlation between a pair of vectors of pairwise sequence similarities computed from an individual MSA; one for each protein family. To account for noise due to shared historical speciation events during evolutionary time, a phylogenetic correction is applied.[53] This phylogenetic correction is estimated by a vector of sequence similarities in essential Housekeeping genes like the subunits of the 16s RNA polymerase. After applying this correction, this interaction score is representative of the functional correlations between the two proteins.

Mirror Tree provides a way to identify whether or not a pair of proteins are interacting with each other in a binary fashion. This approach does not provide fine-grain detail on co-evolution at the level of the amino acid sequence. Positional Mirror Tree extends Mirror Tree to identify co-evolutionary relationships between amino acids across proteins.[54]

Broadly, Positional Mirror Tree computes a Pearson correlation for each pair of amino acid positions within and across a pair of proteins. For a single amino acid position, *a,* the amino acid identities of each pair of sequences in MSA of one protein family are compared to each other. If they are the same amino acid, this is assigned a 1. If they are not the same amino acid, this is assigned a 0. The result is a binary square matrix with the dimensions of the number of sequences in the alignment by the number of sequences in the alignment. The upper diagonal of this binary matrix represents is then linearized into a vector. This vector represents the pairwise identities of the sequences at position *a*. This process is repeated for amino acid position *b*. A Pearson correlation is then calculated between identify vectors for *a* and *b.* Like in Mirror Tree, the phylogenetic correlation is applied to reduce phylogenetic noise. Now, this Positional Mirror Tree score represents how strongly *a* and *b* are co-evolving with each other. This process of generating a pair of binary matrices and computing their interaction score is repeated for every single pair of positions within and between both proteins.

The result of Positional Mirror Tree is a matrix of co-evolutionary relationships between all pairs of amino acids within and between the sequences of DHFR and TYMS. A slice of this matrix

along a single position in TYMS (e.g. R166), contains the correlations of every single position in amino acid sequence of DHFR to this single TYMS position. In other words, this vector shows how the entire sequence of DHFR is constrained by this one position in TYMS. Future work should use the dataset on the epistasis between DHFR and TYMS to test whether or not the model of sequence co-evolution in Positional Mirror Tree is predictive of real, functional relationships across proteins.

Currently, Positional Mirror Tree would be the method that can infer relationships between proteins at the amino acid sequence level. My dataset of the epistasis in the sequence of DHFR to TYMS would be essential to validating such a model of sequence co-evolution. If Positional Mirror Tree or another model of sequence co-evolution between proteins is effective at representing real sequence constraints in protein-protein interactions, such a framework would also provide some hope for biologists everywhere. With this method, a biologist can first use these evolutionary statistical methods to identify whether their proteins of interest interact and how this interaction is encoded in the sequence. These models could then inform targeted experiments to closely study the mechanism driving the protein-protein interaction.

REFERENCES

1. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855–867 (2008).
2. Horovitz, A. & Fersht, A. R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *Journal of Molecular Biology* **214**, 613–617 (1990).
3. Karlin, S. General two-locus selection models: some objectives, results and interpretations. *Theor Popul Biol* **7**, 364–398 (1975).
4. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *PNAS* **99**, 14878–14883 (2002).
5. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).
6. Clackson, T. & Wells, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science* **267**, 383–386 (1995).
7. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
8. Heyne, M. *et al.* Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein–Protein Complexes. *J. Am. Chem. Soc.* (2021) doi:10.1021/jacs.1c08707.
9. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *PNAS* **110**, 15674–15679 (2013).
10. Rivoire, O., Reynolds, K. A. & Ranganathan, R. Evolution-Based Functional Decomposition of Proteins. *PLOS Computational Biology* **12**, e1004817 (2016).

11. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072–1080 (2012).
12. Skerker, J. M. *et al.* Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell* **133**, 1043–1054 (2008).
13. Wellington Miranda, S. *et al.* A covariation analysis reveals elements of selectivity in quorum sensing systems. *eLife* **10**, e69169 (2021).
14. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
15. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
16. Babu, M. *et al.* Global landscape of cell envelope protein complexes in Escherichia coli. *Nat Biotechnol* **36**, 103–112 (2018).
17. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* **32**, 285–290 (2014).
18. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
19. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genom. Hum. Genet.* **20**, 433–460 (2019).
20. Schober, A. F. *et al.* A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism. *Cell Reports* **27**, 3359-3370.e7 (2019).

21. Ducker, G. S. & Rabinowitz, J. D. One-Carbon Metabolism in Health and Disease. *Cell Metabolism* **25**, 27–42 (2017).

22. Kwon, Y. K. *et al.* A domino effect in antifolate drug action in Escherichia coli. *Nat Chem Biol* **4**, 602–608 (2008).

23. Agarwal, P. K., Billeter, S. R., Rajagopalan, P. T. R., Benkovic, S. J. & Hammes-Schiffer, S. Network of coupled promoting motions in enzyme catalysis. *PNAS* **99**, 2794–2799 (2002).

24. Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **313**, 1638–1642 (2006).

25. Finer-Moore, J. S., Santi, D. V. & Stroud, R. M. Lessons and Conclusions from Dissecting the Mechanism of a Bisubstrate Enzyme: Thymidylate Synthase Mutagenesis, Function, and Structure. *Biochemistry* **42**, 248–256 (2003).

26. Stroud, R. M. & Finer-Moore, J. S. Conformational Dynamics along an Enzymatic Reaction Pathway: Thymidylate Synthase, "the Movie". *Biochemistry* **42**, 239–247 (2003).

27. Sotelo-Mundo, R. R., Changchien, L., Maley, F. & Montfort, W. R. Crystal structures of thymidylate synthase mutant R166Q: Structural basis for the nearly complete loss of catalytic activity. *Journal of Biochemical and Molecular Toxicology* **20**, 88–92 (2006).

28. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* **147**, 1564–1575 (2011).

29. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801–807 (2014).

30. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biology* **18**, 150 (2017).

31. McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).

32. Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B. & Bolon, D. N. A. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLOS Genetics* **9**, e1003600 (2013).

33. 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**, e53476 (2020).

34. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160**, 882–892 (2015).

35. Sawaya, M. R. & Kraut, J. Loop and Subdomain Movements in the Mechanism of Escherichia coli Dihydrofolate Reductase: Crystallographic Evidence,. *Biochemistry* **36**, 586–603 (1997).

36. Stout, T. J., Sage, C. R. & Stroud, R. M. The additivity of substrate fragments in enzyme–ligand binding. *Structure* **6**, 839–848 (1998).

37. Wang, Z. *et al.* Mg2+ binds to the surface of thymidylate synthase and affects hydride transfer at the interior active site. *J Am Chem Soc* **135**, 7583–7592 (2013).

38. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

39. Li, X., Lalić, J., Baeza-Centurion, P., Dhar, R. & Lehner, B. Changes in gene expression predictably shift and switch genetic interactions. *Nat Commun* **10**, 3886 (2019).

40. Lee, J. *et al.* Surface Sites for Engineering Allosteric Control in Proteins. *Science* **322**, 438–442 (2008).

41. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**, 946–950 (2009).

42. Schnell, J. R., Dyson, H. J. & Wright, P. E. Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 119–140 (2004).

43. Dykhuizen, D. E., Dean, A. M. & Hartl, D. L. Metabolic Flux and Fitness. *Genetics* **115**, 25–31 (1987).

44. Xu, L., Barker, B. & Gu, Z. Dynamic epistasis for different alleles of the same gene. *PNAS* **109**, 10420–10425 (2012).

45. GENEWIZ from Azenta | Standalone NGS Solutions. https://www.genewiz.com/Public/Services/Next-Generation-Sequencing/Standalone-NGS-Solutions/.

46. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

47. Rodrigues, J. V. *et al.* Biophysical principles predict fitness landscapes of drug resistance. *PNAS* **113**, E1470–E1478 (2016).

48. Kacser, H. & Burns, J. A. The control of flux. *Biochem Soc Trans* **23**, 341–366 (1995).

49. Jaffe, M. *et al.* Improved discovery of genetic interactions using CRISPRiSeq across multiple environments. *Genome Res.* **29**, 668–681 (2019).

50. Gur, E. & Sauer, R. T. Recognition of misfolded proteins by Lon, a AAA+ protease. *Genes Dev.* **22**, 2267–2277 (2008).

51. Hilton, S. K., Doud, M. B. & Bloom, J. D. phydms: software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* **5**, e3657 (2017).

52. Pazos, F., Ranea, J. A. G., Juan, D. & Sternberg, M. J. E. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol* **352**, 1002–1015 (2005).

53. Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**, 3482–3489 (2005).

54. Schober, A. F. Using Evolutionary Statistics to Understand Cellular Systems. (University of Texas Southwestern Medical Center, 2019).