EXPLORING THE MOLECULAR EVOLUTION OF PROTEINS WITH DEEP MUTATIONAL SCANNING

By

Matthew Steven Faber

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology – Doctor of Philosophy

2019

ABSTRACT

EXPLORING THE MOLECULAR EVOLUTION OF PROTEINS WITH DEEP MUTATIONAL SCANNING

By

Matthew Steven Faber

In this thesis, deep mutational scanning is expanded and applied to better understand the molecular evolution of proteins. Protein evolution is a complex process where subtle changes in molecular architecture can have massive impacts on biophysical properties, altering how well-adapted a protein is to a specific task or environment. Deep mutational scanning provides a finer level of understanding of molecular evolution by assessing the effect of every possible single-mutation on a protein's function. The technique combines site saturation mutant libraries, high throughput selections, and deep sequencing to tabulate the changes in mutant frequencies. From these changes the impacts of the mutations on protein function are characterized. This technology allows for efficient exploration of the local evolutionary landscape of a protein, making it a powerful tool for understanding evolution.

Here, I use deep mutational scanning to study how the initial likelihood of obtaining the native folded state of an enzyme *in vivo* constrains its evolution. We designed two unique single-point mutants of AmiE, an aliphatic amidase from *Pseudomonas aeruginosa*. These mutant enzymes are significantly less likely to reach the native folded state *in vivo* than the unmutated precursor and have catalytic efficiencies that are statistically indistinguishable from the initial unmutated enzyme. I tested the impacts of nearly all single-point mutations for the two impaired enzymes using high-throughput growth selections and compared them to the precursor enzyme. These comparisons provided insights into how evolutionary outcomes are changed following

decreases in the likelihood of native folding, and on how the impacts of single mutations combine to influence function.

The other primary goal of this thesis is the development of a new method that expands the utility of deep mutational scanning studies. This method assembles comprehensive single-site saturation, and large multi-point, mutant genome libraries of the bacteriophage ϕX174. To assemble the mutant genome libraries we combine nicking scanning mutagenesis and Golden Gate cloning. With these viral genome libraries, deep mutational scanning experiments can be performed *in situ*. These libraries are a valuable tool for studying the molecular determinants of viral host switching, the combination of inter- and intra-subunit mutations, and other aspects of the molecular evolution of viruses.

This thesis is dedicated to my family and friends who have inspired and encouraged me to pursue a life of exploration through science.

ACKNOWLEDGEMENTS

I want to acknowledge my graduate advisor, Tim Whitehead, for his guidance, support, and patience in helping me mature my scientific practices. Our time working together has significantly impacted my character in many positive ways and imbued me with the skills necessary to succeed. I also want to acknowledge my labmates - Angelica Medina-Cucurella, Carolyn Haarmeyer, Caitlin Stein, Emily Wrenbeck, James Stapleton, Justin Klesmith, Matilda Newton, Matt Bedewitz, Monica Kirby, and PJ Steiner - for their support and friendship. I want to thank my parents and siblings for their unending love, support, and for always encouraging me to be a curious and creative individual. I want to thank my friends for the great times spent together, and for the good times still to be had. Finally, I want to thank my wife Fabiola for her constant love, positivity, support, and inspiration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AmiE, Amidase E

CDR, complementarity determining regions

DFE, distribution of fitness effects

DMS, deep mutational scanning

FACS, fluorescence activated cell sorting

FPLC-SEC, fast protein liquid chromatography with a size exclusion chromatography gel

IMAC, immobilized metal affinity chromatography

KS, Kolmogorov-Smirnoff

NSM, nicking scanning mutagenesis

scFv, single chain fragment of variable regions

V_H-V_L, variable heavy and variable light chains

TROP2Ex, tumor associated calcium signal transducer 2 extracellular domain

YSD, yeast surface display

CHAPTER 1

An introduction to deep mutational scanning, its limitations, and products

Abstract

Deep mutational scanning combines saturation mutagenesis, high throughput selections, and deep sequencing to characterize the impacts of individual mutations on a respective protein. This technology allows for the characterization of thousands of mutations in parallel and as such has greatly expanded our abilities to understand protein evolution and perform rational design. This chapter will provide an introduction to deep mutational scanning, its methodological limitations, and what the outputs can and cannot tell us about the impacts of the tested mutations.

Introduction

Deep mutational scanning (DMS) is a platform technology for efficiently assessing the impact of thousands of individual mutations on a protein of interest[1,2] (**Figure 1.1**). A DMS experiment requires a high throughput selection (HTS) for the protein of interest, the ability to generate comprehensive single-site saturation mutant libraries, and access to a deep sequencing platform. DMS is facilitated by clever cellular and molecular biology methods for HTS and for library preparations, as well as by the drop in deep sequencing costs to pennies per million base-pairs. The outputs of DMS are fitness metrics for each mutation that describe the magnitude by which a given mutation is beneficial or deleterious in the HTS performed. This chapter is not a comprehensive review of deep mutational scanning as there are already several excellent reviews that do so[3,4]. Additionally, this chapter will not focus on the application of the obtained datasets for studying evolution, or for forward engineering as these aspects have been reviewed in Wrenbeck et al.[5] and Faber and Whitehead[6] (Chapter 2). This chapter will provide a brief introduction to the technical aspects of DMS, what and where the limitations in the methods are, and what the obtained datasets can and cannot tell us about the mutations studied.

The technology

As mentioned above DMS is the combination of three different processes: preparation of comprehensive site-saturation mutant libraries, high throughput selections, and deep sequencing (**Figure 1.1**). Here I will briefly describe advances in methods for preparing the mutant libraries, and the HTSs most used in our laboratory.

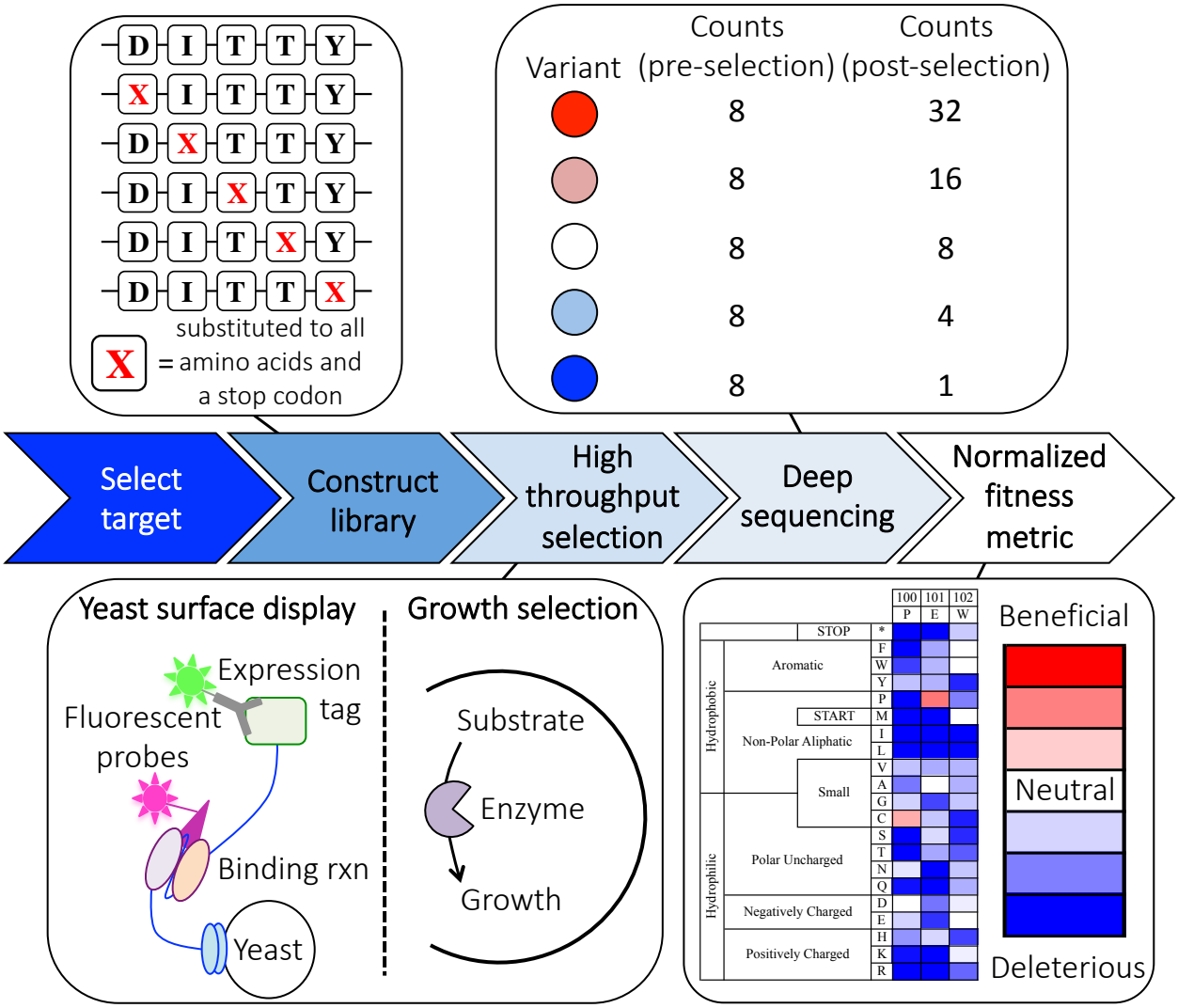

Figure 1.1: Deep mutational scanning overview. A DMS experiment begins with target selection, the selected target has to be compatible with a HTS. Next, comprehensive site-saturation mutant libraries are created. The HTS is applied to the mutant libraries to cause the weaker mutants to drop out, and beneficial mutations to proliferate. Counts of mutants in the pre- and post-selection libraries are tabulated using deep sequencing, and normalized fitness metrics are calculated based on the performance of the unmutated protein.

Library preparation

Single site saturation (SSM) mutagenesis libraries are constructed by template based mutagenesis using plasmid ssDNA and mixtures of mutagenic primers that encode the desired mutation(s). The mutant strand is ligated, the parent DNA strand digested, and a second

replication and ligation is performed[7]. Finally, the library is transformed into a cloning strain of *E. coli* to generate enough plasmid library for downstream applications. Initial library preparations using the Pfunkel method required preparation of ssDNA from bacteriophage and used uracil-containing DNA, which increased time and effort[8]. During my PhD studies my lab invented Nicking Mutagenesis, which allowed for the production of ssDNA by nicking a single strand of a dsDNA plasmid and digesting it with exonucleases[8]. This key advance allowed SSM library preparation from plasmid dsDNA in a single day. We have further advanced the production of these mutant libraries by replacing the degenerate mutagenic oligos with unamplified ink-jet printed oligo pools that contain user-defined mutations[9] (Chapter 5). Using unamplified oligo pools in place of degenerate oligos both simplifies the mutagenesis procedure and results in libraries with greater representations of all of the programmed mutations[9].

High throughput selections

The type of HTS used in a DMS experiment is dependent on the protein(s) being studied and the information one wants to obtain[5,10]. The two primary screens used in our laboratory are yeast surface display (YSD) paired with fluorescence activated cell sorting (FACS)[9], and growth-based selections – where weaker variants become depleted and stronger variants enriched - for studying enzymes[10-12] (Chapter 3; **Figure 1.1**). Other selections exist, such as phage display in place of yeast display, lytic virus growth selections, or reporter based assays for enzyme function. Pairing YSD and FACS can be used to study protein-protein interactions, improve/engineer affinity and specificity[13-16], and also to map paratopes and epitopes[17-18]. YSD and FACS can also be used to improve/study the stability of proteins and enzymes[19-21]. DMS experiments with growth-based selections can be used to inform the affinity/specificity

engineering of enzymes, and to improve enzymes for a respective reaction environment[12] (Chapter 3). Both YSD-FACS, and growth-based selections, can be used to answer molecular evolution questions because these methods map local fitness landscapes for a given environmental condition[12,20] (Chapter 3). Thus, DMS is a valuable tool for answering both applied and fundamental research questions.

The limitations

Experimental

DMS experiments require a suitable HTS, which significantly restricts the reach of these techniques. For YSD paired with FACS the desired target for scanning must be able to be displayed on the yeast cell surface, and the purified binding partners must be able to be produced[10,17,18] . Additionally, the reaction conditions have to be compatible with FACS and non-toxic to the displaying cells. Altered glycosylation patterns – inherent in YSD – can also impose significant issues. The mannose rich O- and N-linked glycosylation patterns in yeast are distinct from those in mammals, and cytosolic proteins are naturally unglycosylated. The attachment of non-native glycosylations can disrupt binding and alter other biophysical parameters for a protein. Also restricting the use of FACS is the oligomeric state of the protein being displayed and our group has successfully displayed up to homotrimers on surface of yeast[17].

Growth-based HTSs require that cell growth be dependent on the reaction catalyzed by the enzyme of study. In Chapter 3 we study an amidase using DMS. When supplied with an amide the amidase releases ammonium, allowing for nitrogen restricted growth selections. Growth selections also require a growth rate ratio – growth rate in selective media over growth rate in unselective media – that is within a window of values[10]. Outside of the window the

growth of the cell is not proportional to the function of the enzyme. Tuning the growth rate ratio to be within this window requires certain biophysical parameters of enzymes like solubility, and catalytic efficiency to be met, as well as transcription and translational tuning with synthetic elements. Therefore, enzymes with very low catalytic efficiencies, and/or low solubilities, are unlikely to be compatible with DMS.

DMS is subject to the limits of the current mutant library preparation methods, cell sorting restrictions, and limitations in sequencing technologies[9,10,18,22]. The information obtained from DMS experiments is also restricted to a range of fitness metrics. DMS experiments performed using the pipelines developed in our lab can only scan an ~5-fold range of differences in fitness metrics[20,23]. DMS can discriminate sign differences with great accuracy and fidelity, but the ability to accurately quantify the impact of very deleterious mutations is dependent on sequencing depth[10-12] (Chapter 3). Often very deleterious mutations can only be qualitatively described.

When analyzing obtained datasets we are limited in the conclusion we can draw about the biophysical impacts of mutations. This limitation is typical in HTSs as these screens are often general competition assays that do not discriminate between the various biophysical properties that can be modified. Additional HTS screens[19,20] that select for a specific biophysical property, or experimental characterization of individual mutants[12] are required to determine how mutations impact the biophysical properties of the protein.

Evolutionary insights

DMS datasets have to be approached with the understanding that these are limited test tube evolution experiments. Additionally, in any selection you get what you select for, meaning

that the conclusions and insights we obtain are biased for, and limited to, the selections performed[10,12,20] (Chapter 3). In Chapter 3, we perform our selections in a highly controlled environment at 37°C. However, our model enzyme - Amidase E from *P. aeruginosa* - has naturally evolved to function in a wide range of temperatures[24]. It is likely that mutations that are beneficial or deleterious at 37°C could have drastically different outcomes at different temperatures[25]. Additionally, our model enzyme is promiscuous and able to digest many short chain amides[12,24], yet the selections we perform in Chapter 3 only use one substrate, acetamide. Therefore, we have selected for mutants that are only better at catalyzing the digestion of acetamide, or that increase the amount active enzyme *in vivo* at 37°C in the specific cellular host. It is possibly that the beneficial mutations found in Chapter 3 will have decreased catalytic efficiencies with other substrates, and that some of the deleterious mutations found might be beneficial for digesting other aliphatic amides[12] (Chapter 3).

Selection conditions inherently constrain and bias the outcomes of test tube evolution experiments[11,12,26,27]. In laboratory evolution experiments the use of a single set of environmental conditions – which produce fitness values that can accurately describe mutational impacts only in the respective environment – is often required to limit the complexity of evolution such that it can be studied and understood[27]. For example, the Lenski evolution experiments have studied molecular and organismal evolution in several *E. coli* strains, individually, at a single temperature and constant chemical environment for over 60,000 generations[28,29]. The evolutionary insights obtained from this experiment are extremely valuable, yet they are biased for the organisms studied and the selection conditions used. By comparison, DMS approaches to studying molecular evolution are also subject to these limitations and biases. The environment used in a DMS experiment is user defined within certain parameters, and the selection

environment can be modified as long as the HTS is maintained[10]. However, altering the conditions of the HTS to test other environments – for example by performing growth-based selections at several different temperatures – can be extremely challenging and may not be feasible.

Finally, DMS is currently unable to model evolution over thousands of years and in highly complex environments. In nature there are many different environments a single organism can pass through, and many other organisms and pathogens to compete with, making the process of evolution massively more complex than in our laboratory experiments[30-33]. Therefore, while the data obtained from DMS is very informative, we must never lose sight of the fact that what is true in the test tube is an oversimplification of what is possibly occurring in nature.

The products

Deep mutational scans provide local fitness landscapes typically defined as nearly all single mutational steps in the evolutionary space[1,2] (**Figure 1.1**). These can be used to gain insights into molecular evolution[12,20] (Chapter 3) and can be used in forward engineering proteins toward a desired purpose[5,6] (Chapter 2). As described previously, the definition of fitness is dependent on the HTS used and the environment of the assay. The obtained normalized fitness metrics describe how well a respective mutation is able to compete with the unmutated predecessor protein in the respective HTS[10].

Growth selection products

With growth based selections for studying enzymes, the normalized fitness metric describes how well a cell hosting a respective mutant is able to compete with – replicate faster

than – a cell hosting the unmutated enzyme in the assaying conditions at a respective temperature[10]. The mathematics for calculating normalized fitness metrics for growth based selections – as in Chapter 3 - from deep sequencing data was published by Kowalsky et al.[10] and is as follows. First, the frequency of a respective mutant (i) in the pre- (f_{io}) and post-selection (f_{if}) populations is calculated with the tabulated counts of mutants from the deep sequencing data:

$$f_{oi} = \frac{x_{oi}}{\sum x_{oi}} \quad (1)$$

$$f_{fi} = \frac{x_{fi}}{\sum x_{fi}} \quad (2)$$

Where x_{io} and x_{if} are the number of counts of a specific variant in the pre- and post-selection populations respectively, and Σx_{oi} and Σx_{fi} are the total number of counts for all variants in the pre- and post-selection populations respectively. Next, the enrichment ratio, ε_i, for a mutant in the population is calculated from frequencies of the mutants obtained in (1) and (2):

$$\varepsilon_i = log_2\left(\frac{f_{fi}}{f_{oi}}\right) \quad (3)$$

The equation for calculating the enrichment ratios can be rewritten as:

$$\varepsilon_i = log_2\left(\frac{x_{fi}}{x_{oi}}\right) - log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right) \quad (4)$$

We also write the equation for the specific growth rate (μ_i) (5) of a cell as:

$$\mu_i = \ln\left(\frac{x_{fi}}{x_{oi}}\right)\frac{1}{t} \quad (5)$$

Where t is the difference in time between the end of the selection (f) and initiation of the selection (o). Combining the equations (4) and (5) gives:

$$\mu_i log_2 e = \frac{1}{t}\left(\varepsilon_i + log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right)\right) \quad (6)$$

To simplify the equation we can first generate the average doubling period (g_p) as a function of the change in the total number of counts for all variants in the pre- and post-selection populations:

$$g_p = Number\ of\ doublings = log_2(\frac{\sum x_{fi}}{\sum x_{oi}}) \quad (7)$$

We can also remove *t* through redefining it:

$$t = \frac{ln2 * g_p}{\mu_p} \quad (8)$$

Where μ_p is the bulk average growth rate of the population for the time between the initiation and end of the selection. Next we combine (7) and (8) into (6) to make the growth rate of a mutant a function of its respective enrichment ratio:

$$\mu_i = \mu_p\ (\frac{\varepsilon_i}{g_p} + 1) \quad (9)$$

Finally, we can calculate our normalized fitness metric (ζ_i) where the growth rate of the respective mutant protein (i) is normalized to that of the unmutated predecessor protein (wt):

$$\zeta_i = log_2\ \left(\frac{\mu_i}{\mu_{wt}}\right) \quad (10)$$

We can also set the normalized fitness metric as a function of the enrichment ratios and average doubling period:

$$\zeta_i = log_2 \left(\frac{\left(\frac{\varepsilon_i}{g_p}\right)+1}{\left(\frac{\varepsilon_{wt}}{g_p}\right)+1}\right) \quad (11)$$

The normalized fitness metric provides a quantitative description of how well a cell harboring a mutant *i* is able to compete with a cell hosting the predecessor enzyme in the respective environment. This data does not provide information on how the mutation impacts the biophysics of the enzyme.

For enzyme-based DMS, there are two general biophysical factors that are impacted by mutations and that determine fitness outcomes: specific velocity, and the amount of active enzyme being expressed within the cell. Both of these general properties are composed of a variety of different factors. Specific velocity is dependent on the standard Gibbs free energy of the reaction being catalyzed, substrate and product concentrations, flux of upstream and downstream reactions, the Michaelis constant (K_M), and the maximum reaction velocity (k_{cat})[25,34]. The probability of the enzyme being expressed and properly folded is a function of: the folding rate, the fidelity of the association process for oligomeric proteins, and the thermodynamic stability of the tertiary and quaternary structures and intermediates[35,36]. A growth selection is unable to discriminate between any of the above-mentioned factors, and additional biophysical analysis is required to understand why a mutation is deleterious or beneficial.

YSD-FACS products

This chapter will not present the mathematics used in calculating the normalized fitness metrics for DMS experiments with YSD-FACS, those calculations can be found in the referenced work[10,17]. While I have performed DMS with YSD-FACS for epitope mapping[17], paratope mapping (results incorporated into a patent in preparation), and affinity maturation (results incorporated into a patent in preparation), these projects are not included in this thesis. Normalized fitness metrics obtained from YSD-FACS experiments provide a quantitative measure of how well a mutant protein competes with its unmutated ancestor in either binding to a given target, or in displaying on the surface of the yeast in a respective environment[10,17,20]. Unlike the growth-based selections, YSD-FACS experiments can select for two different general biophysical properties: affinity/specificity, and surface display (**Figure 1.1**).

Comparing the obtained fitness landscapes from selections for binding/display with those from selections only for display allows for the discrimination of mutations that impact stability/solubility from those that impact affinity/specificity[17,19,20]. Selections for display can be used to select for more stable variants, and to allow us to identify mutations that alter folding probabilities. To aid in discriminating the impacts of the mutations Kowalsky et al.[17] calculated Shannon entropy metrics for each position in the protein being studied for both the displaying sorted and binding/display sorted populations. These Shannon entropy metrics provide a quantitative measure of how well the respective position tolerates mutations. Positions with the lowest Shannon entropy scores are the least able to tolerate mutations. Comparison of Shannon entropy scores from selections for display with selections for binding/display allows for the identification of residues critical for the binding reaction. This comparative analysis is the foundation for mapping protein-protein interactions[17,18] with DMS. As described in Chapter 2, comparison of DMS datasets from different HTS with the same target protein allows for greater discrimination of the biophysical implications for respective mutations, and how this data can be used for forward engineering.

REFERENCES

REFERENCES

1. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S: **High-resolution mapping of protein sequence–function relationships**. *Nat Methods* 2010, 7:741-746.

2. Hietpas RT, Jensen JD, Bolon DNA: **Experimental illumination of a fitness landscape**. *Proc Natl Acad Sci U S A* 2011, 108:7896-7901.

3. Araya CL, Fowler DM (2011) **Deep mutational scanning: Assessing protein function on a massive scale**. *Trends Biotechnol* 29(9):435–442.

4. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat Methods* 2014, 11:801–807.

5. Wrenbeck EE, Faber MS, Whitehead TA: **Deep sequencing methods for protein engineering and design**. *Curr Opin Struct Biol* 2017, 45:36-44.Wrenbeck 2019

6. Faber MS, Whitehead TA: **Data-driven engineering of protein therapeutics**. *Curr Opin in Biotech* 2019, 60:104-110.

7. Firnberg E, Ostermeier M**: PFunkel: efficient, expansive, user- defined mutagenesis**. *PLoS ONE* 2012, 7:e52031.

8. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat Methods* 2016, 13:928-930.

9. Medina-Cucurella AV, Steiner PJ, Faber MS, Beltrán J, Borelli A, Kirby MB, Cutler SR, Whitehead TA: **User-defined single pot mutagenesis using unpurified oligo pools**. *Prot. Eng. Des. Sel.* 2019, TBD:TBD.

10. Kowalsky CA, Klesmith JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA: **High-Resolution Sequence-Function Mapping of Full-Length Proteins**. *PLoS One* 2015, 10:e0118193.

11. Klesmith JR, Bacik JP, Michalczyk R, Whitehead TA: **Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli**. *ACS Synth Biol* 2015, 4:1235–1243.

12. Wrenbeck EE, Azouz LR, Whitehead TA: **Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded**. *Nat Comm* 2017, 8:15695.

13. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, Mattos CD, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D: **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing**. *Nat Biotechnol* 2012, 30:543-548.

14. Strauch E-M, Fleishman SJ, Baker D: **Computational design of a pH-sensitive IgG binding protein**. *Proc Natl Acad Sci U S A* 2014, 111:675-680.

15. Procko E, Berguig GY, Shen BW, Song Y, Frayo S, Convertine AJ, Margineantu D, Booth G, Correia BE, Cheng Y et al.: **A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells**. *Cell* 2014, 157:1644-1656.

16. Koenig P, Lee CV, Walters BT, Janakiraman V, Stinson J, Patapoff TW, Fuh G: **Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding**. *Proc Natl Acad Sci U S A* 2017, 114:E486-E495.

17. Kowalsky CA, Faber MS, Nath A, Dann HE, Kelly VW, Liu L, Shanker P, Wagner EK, Maynard JA, Chan C et al.: **Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing**. *J Biol Chem* 2015, 290:26457-26470.

18. Medina-Cucurella AV, Whitehead TA: **Characterizing Protein-Protein Interactions Using Deep Sequencing Coupled to Yeast Surface Display**. *Met in Mol Biol* 2018, 1764:101-121.

19. Julian MC, Li L, Garde S, Wilen R, Tessier PM: **Efficient affinity maturation of antibody variable domains requires co-selection of compensatory mutations to maintain thermodynamic stability**. *Sci Rep* 2017, 7:45259.

20. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA: **Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning**. *Proc Natl Acad Sci U S A* 2017, 114:2265-2270.

21. Wrenbeck EE, Bedewitz MA, Klesmith JR, Noshin S, Barry CS, Whitehead TA: **An automated data-driven pipeline for improving heterologous enzyme expression**. *ACS Synth Biol* 2019, 15:474-481.

22. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, Bradbury ARM: **Deep sequencing in library selection projects: what insight does it bring?** *Curr Opin Struct Biol* 2015, 33:146-160.

23. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS: **The stability effects of protein mutations appear to be universally distributed.** *J mol biol* 2007, 369:1318-1332.

24. Kelly M, Clarke PH: **An inducible amidase produced by a strain of *Pseudomonas aeruginosa***. *J gen Microbiol* 1962, 27:305-316.

25. Somero GN: **Adaptation of enzymes to temperature: searching for basic ‘‘strategies’’**. *Comp Biochem and Physio* 2004, 139:321-333.

26. Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability**. *Proc Natl Acad Sci U S A* 2006, 103:5869–5874.

27. Lenski RE: **Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations**. *The ISME J* 2017, 11:2181-2194.

28. Wiser MJ, Ribeck N, Lenski RE: **Long-term dynamics of adaptation in asexual populations**. *Science* 2013, 342:1364-1367.

29. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM: **The dynamics of molecular evolution over 60,000 generations**. *Nature* 2017, 551:45-50.

30. Holmes EC: **What can we predict about viral evolution and emergence?**. *Curr Opin in Vir* 2013, 3:180-184.

31. Neher RA: **Genetic draft, selective interference, and population genetics of rapid adaptation**. *Annu Rev Ecol Evol Syst* 2013, 44:195-215.

32. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R: **Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures**. *Nature Gen* 2014, 46:82-87.

33. Liao M, Somero GN, Dong Y: **Comparing mutagenesis and simulations as tools for identifying functionally important sequence changes for protein thermal adaptation.** *Proc Natl Acad Sci U S A* 2019, 116:679-688.

34. Siddiqui KS: **Defying the activity–stability trade-off in enzymes: taking advantage of entropy to enhance activity and thermostability**. *Critical Reviews in Biotechnology* 2017, 37:309-322.

35. Baker D, Agard DA: **Kinetics versus thermodynamics in protein folding**. *Biochemistry* 1994, 33(24):7505-7509.

36. Shakhnovich EI: **Theoretical studies of protein-folding thermodynamics and kinetics**. *Curr Opin in Struc Biol* 1997, 7:29-40.

CHAPTER 2

Data-driven engineering of protein therapeutics

This chapter is adapted with permission from the article “Data-driven engineering of protein therapeutics” in *Current Opinion in Biotechnology* 60:104–110 by Matthew S. Faber and Timothy A. Whitehead.

Abstract

Protein therapeutics requires a series of properties beyond biochemical activity, including serum stability, low immunogenicity, and manufacturability. Mutations that improve one property often decrease one or more of the other essential requirements for therapeutic efficacy, making the protein engineering challenge difficult. The past decade has seen an explosion of new techniques centered around cheaply reading and writing DNA. This review highlights the recent use of such high throughput technologies for engineering protein therapeutics. Examples include the use of human antibody repertoire sequence data to pair antibody heavy and light chains, comprehensive mutational analysis for engineering antibody specificity, and the use of ancestral and inter-species sequence data to engineer simultaneous improvements in enzyme catalytic efficiency and stability. We conclude with a perspective on further ways to integrate mature protein engineering pipelines with the exponential increases in the volume of sequencing data expected in the forthcoming decade.

Introduction

Over 100 therapeutic proteins are currently FDA-approved for use as drugs, with functions as variegated as enzymes that rob tumors of necessary nutrients to antibodies that block signaling mechanisms important for clinical presentation of rheumatoid arthritis. Regardless of the exact category of protein, there are a large number of requirements necessary for therapeutic efficacy: the biologic must possess sufficient affinity or catalytic efficiency toward its intended target while minimizing non-specific interactions, harbor a sufficient in vivo half-life, and maintain low immunogenicity and aggregation propensity. Additionally, each protein must satisfy a number of manufacturability constraints such as the capability of expression and purification in sufficient quantities, sufficient storage stability, and so forth. Naturally occurring proteins usually are suboptimal in at least one of these properties. While alterations to these properties can be imparted by mutations (protein engineering), mutations are pleiotropic and often a property can only be improved at the expense of another important requirement. Such competing constraints on protein function render brute force screens – used extensively by contemporary protein engineers – inefficient; therefore, new insights into navigating sequence space efficiently are always of interest.

This past decade has seen the cost of reading and writing DNA drop precipitously[1]. These advances have led to completely new ways to interrogate biology, including construction of thousands of synthetic genes[2], evaluation of functional effect of tens of thousands of mutations in a protein in massively parallel experiments[3], sequencing of thousands of homologues across the tree of life for any given gene, and even ways to sequence human antibody repertoires[4]. Such data-rich reservoirs are beginning to be tapped by protein engineers.

Here, we review recent progress in the use of data-driven protein engineering of potential therapeutics in order to satisfy multiple competing constraints. We restrict this review to protein engineering by amino acid substitutions. There are topical reviews concerning the de-immunization of protein therapeutics[5], engineering competing trade-offs in antibody function[6], and a comprehensive overview of engineered therapeutic enzymes[7]. Thus, we will highlight representative and instructive examples utilizing de novo gene synthesis, deep mutational scanning, phylogenetic analysis, or antibody repertoire sequencing (**Figure 2.1**); we also offer

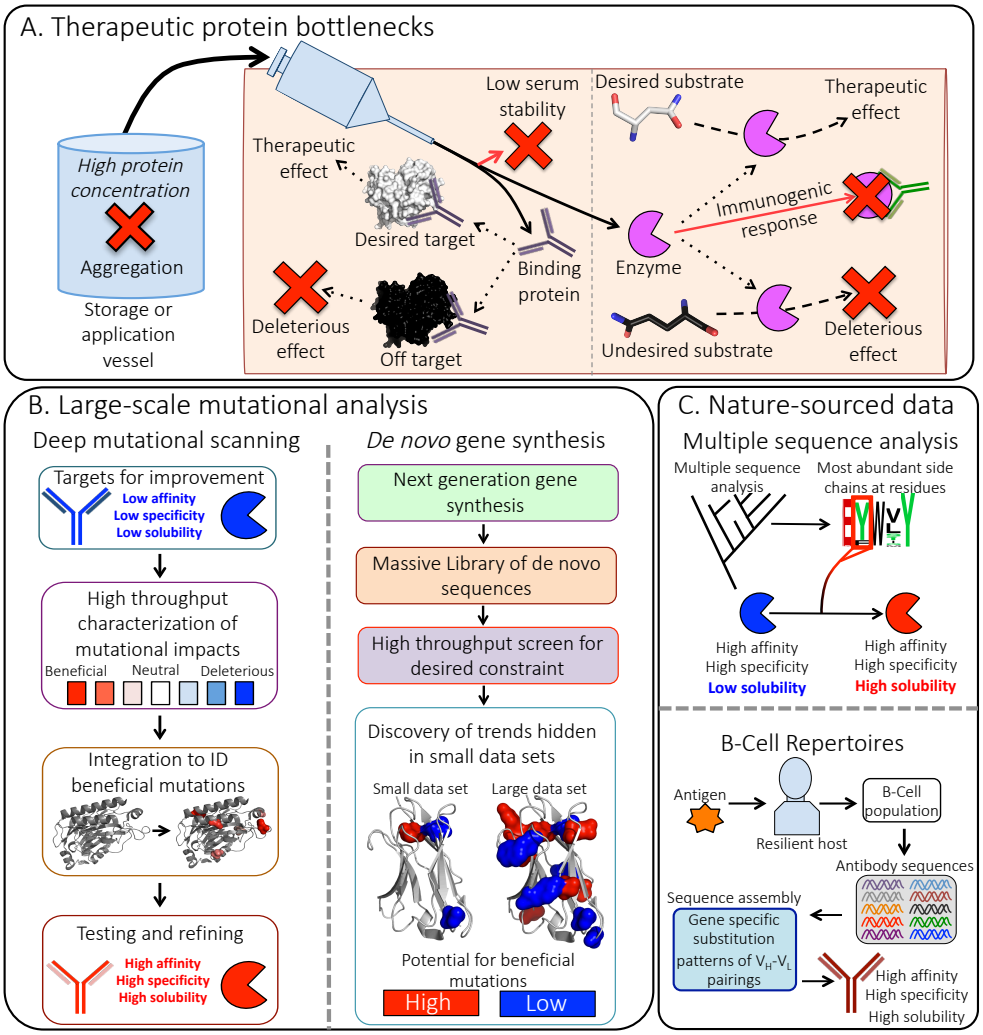

Figure 2.1: High throughput techniques used to overcome therapeutic protein engineering bottlenecks. A. Different bottlenecks encountered in developing therapeutics. B–C. Different high throughput techniques used for protein engineering.

perspective on future ways to harness the extraordinary torrents of data that we anticipate in the coming decade.

Large-scale mutational analysis

Integrating deep sequencing with user-defined saturation mutagenesis[8] and screens such as yeast surface display sorting enables the determination of the functional effect of tens of thousands of mutations on a given protein sequence in a massively parallel fashion; this technique is known as deep mutational scanning[3]. The mutational effect on each functional property of the protein can be assessed separately (if a screen exists), and mutations can be incorporated only if all screens give a positive result. Here, we give the latest examples of leveraging such datasets to engineer proteins while considering multiple constraints.

Antibody deep mutational scanning

Antibodies often bind two or more related proteins, and frequently the protein engineer is tasked with engineering specificity for one protein over the other. A recent notorious example involved the use of an antibody to monitor age-dependent levels of GDF11 in mice[9]. However, this antibody also bound the closely related homolog GDF8 (myostatin), which resulted in erroneous conclusions[10] in the original paper. This multi-specificity problem is general, as many potential targets have very similar homologs in humans with some differing by only 10% or so in pairwise sequence identity[11-14]; antibodies, therefore, must have exquisite specificity for therapeutic and diagnostic applications.

Deep sequencing combined with a suitable screen can be used to identify specificity-modulating mutations in protein binders[15]. A recent excellent example comes from Koenig et al.

from Genentech[16], who sought to prevent binding to angiopoietin-1 (Ang-1) for a candidate antigen binding fragment (Fab) with desired binding to Ang-2 and vascular endothelial growth factor (VEGF). Using phage display, the authors determined the relative binding profile of many possible single point mutants in the six complementarity determining regions (CDR)s (483 light chain, 609 heavy chain) for each protein in parallel. From these 1092 mutations, 25 (2.3%) were shown in vitro to result in no or severely reduced binding to Ang-1. We note that all of these mutations slightly decreased binding affinities for Ang-2 and/or VEGF, illustrating the general difficulty – and in some cases, impossibility – of finding specificity-modulating mutations without a functional trade-off even when the local mutational space is comprehensively sampled.

Thermal stability is another parameter that is often negatively correlated with binding affinity[17]. Reduced Fab thermal stability results in lower expression titers[18]. The necessity of engineering specificity-affinity while maintaining stability is critical for therapeutic application of binding proteins[19], and state of the art methods involve using co-screening for stability and affinity simultaneously[20]. Alternatively, one can screen for stability and affinity in parallel to uncover rare, globally optimal mutations. Recently, the same Genentech team used this approach to scan all possible Fab framework and CDR single point mutants for affinity and stable expression in a phage display context[21]. Affinity-enhancing mutations were identified using selections against VEGF, while stabilizing mutations were identified with selections against protein A or protein L. There were a handful of mutations improving both stability and affinity. In particular, mutation from Phe to Ala at a single framework mutation 25 Å from the binding site, light chain residue 83 (LC-F83A), was found to strongly improve thermostability and affinity. The authors attributed this improvement to alteration of the interface between the variable and constant regions on the light chain, which was confirmed by hydrogen–deuterium

exchange mass spectrometry. Intriguingly, LC-83 is one of the 5–10% highest somatically mutated LC positions as determined by deep sequencing of over a thousand human lymphoid tissues. This suggests that LC-83 mutations can generally improve thermostability in Fabs, which the authors confirmed by incorporating LC-F83A in several unrelated mAbs. This is a great example of integrating naturally sourced deep sequencing data on a single target to obtain general insights into the functional tradeoffs of mutations in a broader set of antibodies.

Enzyme deep mutational scanning

Enzymes are applied in the treatment of diverse disorders such as cancer therapeutics to deplete essential amino acids or metabolic precursors needed by the cancer cells[22,23], or as replacement therapies to return specific metabolites to healthier levels in a patient[24]. Engineering non-immunogenic, serum-stable, and active enzymes can be challenging. For example, human kynurenine-degrading enzyme has low serum stability, and engineering has proven difficult because mutations that increase catalytic efficiency have decreased serum stability (J. Blazeck, personal communication). Understanding the trade-offs between stability and catalytic efficiency is critical for simplifying enzyme engineering.

In contrast with protein binders such as antibodies, generalizable high-throughput screens for directly testing enzyme function do not exist. Thus, deep mutational scanning pipelines are not available for all enzyme classes. A creative solution for forward engineering stability and activity was presented by Klesmith et al., who presented the comparative analysis of deep mutational scanning datasets for solubility/stability and activity for two different model enzymes: levoglucosan kinase, and TEM-1 beta-lactamase[25]. From these datasets, Klesmith et al. were able to extract features common to mutations that improve stability while not hampering

catalytic activity. These mutations were sampled in the evolutionary history of the enzyme, were at least 15 angstroms from the active site, were in positions of the tertiary protein sequence without a large number of other residues in close contact, and were not mutations to or from proline. Combining these filtering metrics yields a greater than 90% probability of choosing a stabilizing, catalytically neutral mutation in any given enzyme. This method has been developed into a Rosetta-based script and applied to increase the thermal stability and in vivo expression yield of a Type III polyketide synthase (E. Wrenbeck and T. Whitehead, unpublished results).

Writing libraries of synthetic genes

Compared with more limited datasets of proteins with 1 or 2 amino acid changes described above, datasets with wider tranches of sequence space may be more useful for identifying general engineering rules. For example, several enzyme discovery efforts evaluate candidates by synthesizing hundreds of genes encoding enzymes spread throughout the protein superfamily[26,27]. Over the past five years the cost of writing DNA for kb-size genes has plateaued at about $0.10 per base pair[28], which means that libraries of several hundred genes can be written and tested on a medium-sized lab's budget.

A stunning example of the use of large protein datasets for obtaining engineering insights was recently described by Adimab scientists[18]. Monoclonal antibodies represent the largest class of engineered therapeutic proteins by revenue, with $98 billion in worldwide sales in 2017[29]. Even with dozens of FDA-approved antibodies, some still fail in late-stage clinical trials for reasons unrelated to their binding affinity to their respective target. Understanding how a primary sequence determines manufacturability could help guide antibody-drug development akin to the Lipinski 'rule of five' edict for small molecule drugs[30]. To that end, Adimab scientists produced

a set of nearly all mAbs commercialized or in advanced stage clinical trials described in the patent literature through mid-2009 (137 in total)[18]. They then assayed this set for a dozen biophysical properties including aggregation propensity, non-specific binding, and melting temperature. The first major insight was that – contrary to expectations – many of these antibodies had unfavorable scores in one or more measured biophysical property. The second major insight was the significantly better biophysical properties of the approved subset of mAbs compared to those in Phase 2 clinical trials, suggesting potential difficulties in translating some of these Phase 2 mAbs.

Although this mAb dataset was only published in 2017, researchers are already mining it for predicting the sequence determinants of poor biophysical properties. For example, the Tessier group has used the Adimab dataset to find that positive CDR net charge correlates with mAb self-association using simple sequence-based scoring methods[31]. This same group, following work on individual antibodies[32], has shown that reducing CDR net charge also decreases antibody nonspecificity (P. Tessier, personal communication).

Going from hundreds to tens of thousands of proteins – with greater underlying predictive power – requires a continued decrease in the cost of writing DNA. To that end, recent advances in DNA synthesis using microarray-derived oligo pools[1] have been utilized to synthesize thousands of designed small proteins[2,33]. Producing longer genes of 0.8-2 kb in length from such oligo pools is difficult. The best approaches have a success rate of ~2%[34], and increasing the fidelity rate is essential for these emerging techniques to reach wider application.

Nature-sourced data

Protein design by phylogeny

The cumulative sequencing data of thousands of protein homologs across the tree of life, for nearly all known therapeutically relevant proteins, are a rich resource for the protein engineer. It has been known for decades that, given any position in a protein family, the consensus residue is likely to be stabilizing[35]. This insight enables the engineering of stable enzymes by simultaneously incorporating mutations at consensus positions in a protein sequence. An interesting variation on this idea comes from Nguyen et al., who used consensus sequence information to refine the specificity of an L-asparaginase from Erwinia chrysanthemi (ErA)[36]. The therapeutic effect of wild type ErA is diminished by its dual activity as an L-glutaminase. While asparagine depletion starves cancer cells, glutamine depletion is associated with many negative side effects[37]. To reduce the L-glutaminase activity of ErA, conserved stretches of residues were identified within the active site. Nonconserved adjacent residues were targeted for saturation mutagenesis with the assumption that such mutations could modulate specificity without disrupting desired catalytic function. This strategy resulted in a multi-point mutant with conserved L-asparaginase activity and a $25-fold reduction of L-glutaminase activity. In vivo experiments using the engineered ErA have determined it is an effective therapy for both T and B-cell acute lymphoblastic leukemia without the side effects of glutamine depletion[38].

Consensus mutations informed by phylogeny are increasingly incorporated into structure-based computational design algorithms to engineer stabilized proteins[39]. To cite as the best example, Goldenzweig et al. developed a Rosetta-based method to stabilize human enzymes[40]. Evolutionary conservation data were converted to a position specific scoring matrix, and only mutations above a certain conservation threshold were considered. Mutations passing additional

structure-based filters were combined into new designs. They used this approach to increase the expression yield and stability of a human acetylcholinesterase variant, which can be used for organophosphate detoxification following nerve agent exposure. Use of this method led to an acetylcholinesterase variant with 51 mutations, $2000-fold increased bacterial expression compared to the original enzyme, and with nearly the same specific activity.

The consensus mutation approach has reached its logical end-state with the use of ancestral sequence reconstruction to improve protein function[41-43]. Resurrecting ancient proteins often leads to enzymes with incredible stabilities[44]. Zakas et al. used this general understanding to resurrect several ancestral Factor VIII proteins[45]. By doing so, they engineered a therapeutically effective ancestral Factor VIII variant with higher stability than extant forms.

Immunogenicity is always a major concern for candidate therapeutic proteins. For example, a highly engineered Factor VII variant failed late stage clinical trials because of immunogenicity[46]. Interestingly, the reconstructed ancestral Factor VIII variants share 95% sequence identity with human Factor VIII, but have reduced cross-reactivity with known anti-human Factor VIII antibodies[45]. This reduced cross-reactivity was verified in in vitro analyses. It will be interesting to see if these in vitro experiments translate to lower immunogenicity in animal models, and whether ancestral proteins have lower immunogenicity than extant proteins in general.

Engineering from human antibody repertoires

We are in an exciting age of molecular serology, moving on from bulk measurements of humoral responses towards descriptions of full sequences and functions of individual antibodies in human repertoires[47-49]. These repertoire sets are a critical resource for understanding affinity

maturation, CDR development, and pairing between heavy and light variable chains (V_H–V_L). While there has not been a great deal of engineering work in the open literature exploiting these repertoire datasets, we expect that to change in the next few years. The most recent use of information from these repertoires was described from Adler et al.[50], who showed that incorporating native V_H–V_L pairing improved the quality and quantity of candidate anti-interleukin 21 receptor antibodies isolated from a yeast display screen. Anticipated use of these datasets include designing more efficient antibody libraries by incorporating information from CDR lengths and net charges, and position specific substitution patterns. To facilitate this goal, Sheng et al. constructed gene-specific substitution patterns for 69 common V genes[51], while Kovaltsuk et al. have collected nearly all repertoire datasets into a single data-base called the Observed Antibody Space[52] (antibodymap.org).

Perspective

The ability to write and read DNA at scale has transformed protein engineering into a big data

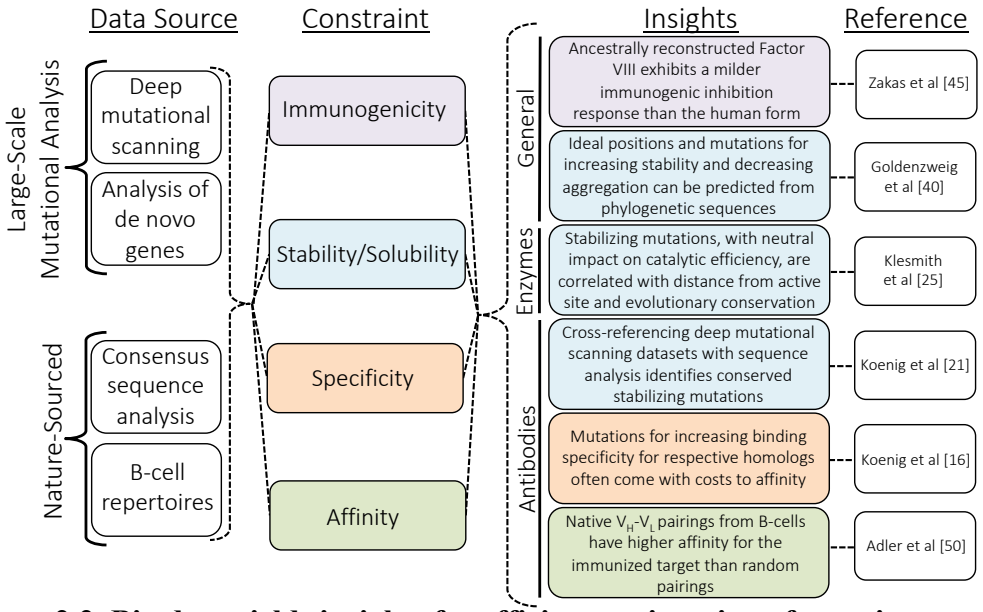

Figure 2.2: Big data yields insights for efficient engineering of protein therapeutics.

field. Here, we have discussed a diversity of data-rich methods for improving protein therapeutics, with each of these methods providing unique engineering insights (**Figure 2.2**). In the next few years, we anticipate further ways to leverage and combine data from multiple sources to improve protein engineering. One of the key emerging ideas is to enhance protein design by incorporating constraints from the evolutionary history of the protein. These constraints are mostly identification of conservation at a single position, but for human antibodies with millions of sequences we imagine more sophisticated algorithms that take into account correlation between residues within or across CDRs, bulk biophysical properties like CDR net charge, and sequence properties as a function of CDR loop length. We also anticipate a marriage of data analysis from a post-hoc view by combining nature-sourced data of existing protein sequences with the forward-evolutionary view of deep mutational scanning to observe the range of possible mutations for a given candidate therapeutic. This integration will allow a comprehensive look at 'what is' with 'what ought to be', enabling the protein engineer to design functional proteins on demand.

Acknowledgements

This work was supported by the National Science Foundation [Career Award #1254238 CBET to T.A.W] and Michigan State University [Johansen Crosby endowed chair to T.A.W.].

REFERENCES

REFERENCES

1. Hughes RA, Ellington AD: **Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology**. *Cold Spring Harb Perspect Biol* 2017, 9:a023812.

2. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A et al.: **Global analysis of protein folding using massively parallel design, synthesis, and testing**. *Science* 2017, 357:168-175.

3. Wrenbeck EE, Faber MS, Whitehead TA: **Deep sequencing methods for protein engineering and design**. *Curr Opin Struct Biol* 2017, 45:36-44.

4. Friedensohn S, Khan TA, Reddy ST: **Advanced methodologies in high-throughput sequencing of immune repertoires**. *Trends Biotechnol* 2017, 35:203-214.

5. Sauna ZE, Lagassé D, Pedras-Vasconcelos J, Golding B, Rosenberg AS: **Evaluating and mitigating the immunogenicity of therapeutic proteins**. *Trends Biotechnol* 2018, 36:1068-1084.

6. Rabia LA, Desai AA, Jhajj HS, Tessier PM**: Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility**. *Biochem Eng J* 2018, 137:365-374.

7. Lutz S, Williams E, Muthu P: **Engineering therapeutic enzymes**. *Directed Enzyme Evolution: Advances and Applications. Springer International Publishing;* 2017:17-67.

8. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat Methods* 2016, 13:928-930.

9. Sinha M, Jang YC, Oh J, Khong D, Wu EY, Manohar R, Miller C, Regalado SG, Francesco SL, Pancoast JR et al.: **Restoring systemic GDF11 levels reverses age-related dysfunction in mouse skeletal muscle**. *Science* 2014, 344:649-652.

10. Egerman MA, Cadena SM, Gilbert JA, Meyer A, Nelson HN, Swalley SE, Mallozzi C, Jacobi C, Jennings LJ, Clay I et al.: **GDF11 increases with age and inhibits skeletal muscle regeneration**. *Cell Metab* 2015, 22:164-174.

11. McPherron AC, Huynh TV, Lee SJ: **Redundancy of myostatin and growth/differentiation factor 11 function**. *BMC Dev Bio* 2009, 9:24-33.

12. Fredriksson R, Lagerström MC, Lundin LG, Schiöth HB: **The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints**. *Mol Pharmacol* 2003, 63:1256-1272.

13. Achen MG, Jeltsch M, Kukk E, Mäkinen T, Vitali A, Wilks AF, Alitalo K, Stacker SA: **Vascular endothelial growth factor D (VEGF-D) is a ligand for the tyrosine kinases VEGF receptor 2 (Flk1) and VEGF receptor 3 (Flt4)**. *Proc Natl Acad Sci U S A* 1998, 95:548-553.

14. Fredriksson L, Li H, Eriksson U: **The PDGF family: four gene products form five dimeric isoforms**. *Cytokine Growth Factor Rev* 2004, 15:197-204.

15. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, Mattos CD, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D: **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing**. *Nat Biotechnol* 2012, 30:543-548.

16. Koenig P, Sanowar S, Lee CV, Fuh G: **Tuning the specificity of a two-in-one fab against three angiogenic antigens by fully utilizing the information of deep mutational scanning**. *MAbs* 2017, 6:959-967.

17. Houlihan G, Gatti-Lafranconi P, Lowe D, Hollfelder F: **Directed evolution of anti-HER2 DARPins by SNAP display reveals stability/function trade-offs in the selection process**. *Protein Eng Des Sel* 2015, 28:269-279.

18. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y et al.: **Biophysical properties of the clinical-stage antibody landscape**. *Proc Natl Acad Sci U S A* 2017, 114:944-949.

19. Koday MT, Nelson J, Chevalier A, Koday M, Kalinoski H, Stewart L, Carter L, Nieusma T, Lee PS, Ward AB et al.: **A computationally designed hemagglutinin stem-binding protein provides in vivo protection from influenza independent of a host immune response**. *PLoS Pathog* 2016, 12:e1005409.

20. Julian MC, Li L, Garde S, Wilen R, Tessier PM: **Efficient affinity maturation of antibody variable domains requires co-selection of compensatory mutations to maintain thermodynamic stability**. *Sci Rep* 2017, 7:45259.

21. Koenig P, Lee CV, Walters BT, Janakiraman V, Stinson J, Patapoff TW, Fuh G: **Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding**. *Proc Natl Acad Sci U S A* 2017, 114:E486-E495.

22. Triplett TA, Garrison KC, Marshall N, Donkor M, Blazeck J, Lamb C, Qerqez A, Dekker JD, Tanno Y, Lu WC et al.: **Reversal of indoleamine 2,3-dioxygenase-mediated cancer immune suppression by systemic kynurenine depletion with a therapeutic enzyme**. *Nat Biotechnol* 2018, 36:758-764.

23. Knott SRV, Wagenblast E, Khan S, Kim SY, Soto M, Wagner M, Turgeon MO, Fish L, Erard N, Gable AL et al.: **Asparagine bioavailability governs metastasis in a model of breast cancer**. *Nature* 2018, 554:378-381.

24. Germain DP, Charrow J, Desnick RJ, Guffon N, Kempf J, Lachmann RH, Lemay R, Linthorst GE, Packman S, Scott CR et al.: **Ten-year outcome of enzyme replacement therapy with agalsidase beta in patients with Fabry disease**. *J Med Genet* 2015, 52:353-358.

25. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA: **Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning**. *Proc Natl Acad Sci U S A* 2017, 114:2265-2270.

26. Carlin DA, Caster RW, Wang X, Betzenderfer SA, Chen CX, Duong VM, Ryklansky CV, Alpekin A, Beaumont N, Kapoor H et al.: **Kinetic characterization of 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants**. *PLoS One* 2016, 11:e0147596.

27. Macdonald SS, Patel A, Larmour VLC, Morgan-Lang C, Hallam SJ, Mark BL, Withers SG: **Structural and mechanistic analysis of a b-glycoside phosphorylase identified by screening a metagenomic library**. *J Biol Chem* 2018, 293:3451-3467.

28. Servick K**: Genome writing project confronts technology hurdles**. *Science* 2017, 356:673-674.

29. Grilo AL, Mantalaris A: **The increasingly human and profitable monoclonal antibody market**. *Trends Biotechnol* 2018, 37:9-16 http://dx.doi.org/10.1016/j.tibtech.2018.05.014.

30. Lipinski CA: **Lead-and drug-like compounds: the rule-of-five revolution**. *Drug Discov Today Technol* 2004, 1:337-341.

31. Alam ME, Geng SB, Bender C, Ludwig SD, Linden L, Hoet R, Tessier PM: **Biophysical and sequence-based methods for identifying monovalent and bivalent antibodies with high colloidal stability**. *Mol Pharm* 2018, 15:150-163.

32. Datta-Mannan A, Thangaraju A, Leung D, Tang Y, Witcher DR, Lu J, Wroblewski VJ: **Balancing charge in the complementarity-determining regions of humanized mAbs without affecting pI reduces non-specific binding and improves the pharmacokinetics**. *MAbs* 2015, 7:483-493.

33. Chevalier A, Silva DA, Rocklin GJ, Hicks DR, Vergara R, Murapa P, Bernard SM, Zhang L, Lam KH, Yao G et al.: **Massively parallel de novo protein design for targeted therapeutics**. *Nature* 2017, 550:74-79.

34. Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S: **Multiplexed gene synthesis in emulsions for exploring protein functional landscapes**. *Science* 2018, 359:343-347.

35. Lehmann M, Pasamontes L, Lassen S, Wyss M: **The consensus concept for thermostability engineering of proteins**. *Biochim Biophys Acta* 2000, 1543:408-415.

36. Nguyen HA, Su Y, Lavie A: **Design and characterization of Erwinia Chrysanthemi L-asparaginase variants with diminished L-glutaminase activity**. *J Biol Chem* 2016, 291:17664-17676.

37. Vidya J, Sajitha S, Ushasree MV, Sindhu R, Binod P, Madhavan A, Pandey A: **Genetic and metabolic engineering approaches for the production and delivery of L-asparaginases: an overview**. *Bioresour Technol* 2017, 245:1775-1781.

38. Nguyen HA, Su Y, Zhang JY, Antanasijevic A, Caffrey M, Schalk AM, Liu L, Rondelli D, Oh A, Mahmud DL et al.: **A novel L-asparaginase with low L-glutaminase coactivity is highly efficacious against both T and B cell acute lymphoblastic leukemias in vivo**. *Cancer Res* 2018, 6:1549-1560.

39. Bednar D, Beerens K, Sebestova E, Bendl J, Khare S, Chaloupkova R, Prokop Z, Brezovsky J, Baker D, Damborsky J: **FireProt: energy-and evolution-based computational design of thermostable multiple-point mutants**. *PLoS Comput Biol* 2015, 11:e1004556.

40. Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J et al.: **Automated structure-and sequence-based design of proteins for high bacterial expression and stability**. *Mol Cell* 2016, 63:337-346.

41. Trudeau DL, Kaltenbach M, Tawfik DS: **On the potential origins of the high stability of reconstructed ancestral proteins**. *Mol Biol Evol* 2016, 33:2633-2641.

42. Lazarus RA, Scheiflinger F: **Mining ancient proteins for next-generation drugs**. *Nat Biotechnol* 2017, 35:28-29.

43. Risso VA, Sanchez-Ruiz JM, Ozkan SB: **Biotechnological and protein-engineering implications of ancestral protein resurrection**. *Curr Opin Struct Biol* 2018, 51:106-115.

44. Romero-Romero ML, Risso VA, Martinez-Rodriguez S, Ibarra-Molero B, Sanchez-Ruiz JM: **Engineering ancestral protein hyperstability**. *Biochem J* 2016, 473:3611-3620.

45. Zakas PM, Brown HC, Knight K, Meeks SL, Spencer HT, Gaucher EA, Doering CD: **Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction**. *Nat Biotechnol* 2017, 35:35-37.

46. for the adept™2 investigators: Lentz S, Ehrenforth S, Karim FA, Matsushita T, Weldingh KN, Windyga J, Mahlangu JN: **Recombinant factor VIIa analog in the management of hemophilia with inhibitors: results from a multicenter, randomized, controlled trial of vatreptacog alfa**. *J Thromb Haemost* 2014, 12:1244-1253.

47. Hwang JK, Wang C, Du Z, Meyers RM, Kepler TB, Neuberg D, Kwong PD, Mascola JR, Joyce MG, Bonsignori M et al.: **Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies**. *Proc Natl Acad Sci U S A* 2017, 114:8614-8619.

48. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, Georgiou G: **In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire**. *Nat Med* 2015, 21:86-91.

49. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ, Georgiou G: **Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires**. *Proc Natl Acad Sci U S A* 2016, 113:E2636- E2645.

50. Adler AS, Bedinger D, Adams MS, Asensio MA, Edgar RC, Leong R, Leong J, Mizrahi RA, Spindler MJ, Bandi SR et al.: **A natively paired antibody library yields drug leads with higher sensitivity and specificity than a randomly paired antibody library**. *MAbs* 2018, 10:431-443.

51. Sheng Z, Schramm CA, Kong R, NISC Comparative Sequencing Program, Mullikin JC, Mascola JR, Kwong PD, Shapiro L: **Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation**. *Front Immunol* 2017, 8:537.

52. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K**: Observed antibody space: a resource for data mining next generation sequencing antibody repertoires**. *J Immunol* 2018, 201:2502-2509.

CHAPTER 3

Impact of in vivo protein folding probability on local fitness landscapes

This chapter is adapted with permission from the publication "Impact of in vivo protein folding probability on local fitness landscapes" in *Molecular Biology and Evolution* by Matthew S. Faber, Emily E. Wrenbeck, Laura R. Azouz, Paul J. Steiner, and Timothy A. Whitehead.

Abstract

It is incompletely understood how biophysical properties like protein stability impact molecular evolution and epistasis. Epistasis is defined as specific when a mutation exclusively influences the phenotypic effect of another mutation, often at physically interacting residues. In contrast, nonspecific epistasis results when a mutation is influenced by a large number of non-local mutations. As most mutations are pleiotropic, the *in vivo* folding probability - governed by basal protein stability - is thought to determine activity-enhancing mutational tolerance, implying that nonspecific epistasis is dominant. However, evidence exists for both specific and nonspecific epistasis as the prevalent factor, with limited comprehensive datasets to support either claim. Here we use deep mutational scanning to probe how *in vivo* enzyme folding probability impacts local fitness landscapes. We computationally designed two different variants of the amidase AmiE with statistically indistinguishable catalytic efficiencies but lower probabilities of folding *in vivo* compared to wild-type. Local fitness landscapes show slight alterations among variants, with essentially the same global distribution of fitness effects. However, specific epistasis was predominant for the subset of mutations exhibiting positive sign epistasis. These mutations mapped to spatially distinct locations on AmiE near the initial mutation or proximal to the active site. Intriguingly, the majority of specific epistatic mutations were codon-dependent, with different synonymous codons resulting in fitness sign reversals. Together, these results offer a nuanced view of how protein folding probability impacts local fitness landscapes, and suggest that transcriptional-translational effects are as important as stability in determining evolutionary outcomes.

Introduction

Understanding the mechanisms of molecular evolution is important to molecular biology, virology, evolutionary biology, and protein engineering. Researchers interested in evolving natural proteins, designing proteins *de novo*, or understanding the extent of contingency on extant proteins must contend with the implicit evolutionary limitations set forth by nature. The challenge, then, is to understand what constrains protein evolution and by what mechanisms. How do these factors interact with one another to alter the frequency of mutations with increased fitness in a given environment, and how do they govern evolvability for new functions?

A particularly important component of evolution is epistasis, or the non-additive combination of mutations[1]. Epistasis impacts the rate of evolution and the spectrum of possible evolutionary pathways available to a protein[2]. Epistasis is said to be specific when a mutation exclusively influences the phenotypic effect of only a few select mutations, usually at physically interacting residues[3]. In contrast, epistasis is said to be nonspecific when a mutation impacts a global property like stability that can be rescued by large numbers of non-local mutations. Of the two classes, specific epistatic effects exert the greatest influence on the possible evolutionary outcomes[3]. This is the result of the precise and long-lasting amino acid constraints imposed by specific epistatic mutations, which decrease the evolutionary reversibility of a protein sequence in a given evolutionary trajectory. In turn, the delocalized and unconstrained effects resulting from non-specific epistasis often breakdown over evolutionary time, largely restricting its influence to short-term evolution. Non-specific epistasis can temporarily alter the mutational robustness of a protein by changing the permissivity to additional mutations at many more positions than does specific epistasis. By modulating the evolutionary trajectories available, epistatic phenomena exert immense influence on the short and long-term evolution of proteins[4].

What remains incompletely understood is how biophysical parameters like protein stability constrain epistasis. Protein stability as defined here is the cumulative balance of the thermodynamic stability, the folding rate, and the fidelity of the association process for oligomeric proteins; these parameters combine to determine the likelihood that an enzyme will assume its native state when expressed *in vivo*: the *in vivo* folding probability of a given protein[5,6]. For enzymes, fitness is often a function of flux through a pathway, which is a product of steady state enzyme concentration and specific velocity. The steady state enzyme concentration, in turn, is a function of the translation rate, the probability of the nascent peptide folding into the native state, and the protein degradation rate. Typical evolutionary models control for translation rate and protein degradation rate and then assume that (i.) thermodynamics of protein folding can be described by a 2-state model; and (ii.) this single Gibbs free energy term can account for the probability of folding. However, these assumptions fail for much of a typical proteome. Many proteins are oligomeric, multi-domain proteins have more complicated folding trajectories, and there is increasing evidence that formation of secondary structure and partial hydrophobic collapse before ribosomal release is important for on-target folding[7]. The in vivo folding probability encompasses all of these biophysical terms into the probability of reaching the folding state.

Analyses of the impacts of stability in evolution at the genomic[8], protein[9-11], and organismal[12] levels have uncovered a complex and dynamic equilibrium between stabilizing and destabilizing mutations. For enzymes in particular, previous studies have shown that missense mutations often act pleiotropically where catalytically enhancing mutations are, on average, moderately destabilizing[9,13,14]. Consequently, high basal stability can buffer catalytically beneficial but destabilizing mutations[15,16], allowing fixation. Deleterious destabilizing mutations

can be repaired by reversion mutations[17], or by specific and non-specific epistatic mutations that rescue stability[8,18]. These epistatic mutations are a central phenomenon in the stabilizing-destabilizing equilibrium, with significant consequences in long-term evolution[19,20]. It is uncertain whether specific or non-specific epistatic mutations are more likely to rescue a destabilized protein, with evidence existing for both arguments[17-21].

Deep mutational scanning experiments provide a wealth of mutational data that can be used to address questions in molecular evolution[22]. This technology comprises the use of large mutational libraries with selections coupled to deep sequencing to evaluate relative fitness of thousands of variants in a massively parallel fashion[14,23-25]. We previously used deep mutational scanning on the homohexameric aliphatic amidase AmiE from *Pseudomonas aeruginosa* to understand how local fitness landscapes, defined here as the set of all possible single-point amino acid substitutions from wild-type (WT), change with different substrates[25]. In this original study, AmiE was chosen as a model as it is stable in its genetic background and has a high probability of folding upon translation. To comprehensively assess how the initial probability of folding *in vivo* constrains mutational outcomes, we designed two variants of AmiE in which catalytic activity is unperturbed but the proteins have different *in vivo* folding probabilities. We then used deep mutational scanning to probe the local fitness landscapes of these variants. While we found moderate epistasis, local fitness landscapes are largely insensitive to the initial *in vivo* folding probability of the enzyme variant. In particular, the great majority of beneficial mutations were shared between all three starting points: WT AmiE, and the two disrupted single point-mutant enzymes. However, positive sign epistasis was present and was dominated by specific epistasis. Remarkably, we found that the sign of the fitness metric for many mutations depends on the codon used to encode the mutation, suggesting more complicated fitness landscapes than

predicted from intrinsic protein biophysics. Together, these results provide a nuanced view of how local fitness landscapes are perturbed under slightly different initial *in vivo* folding probabilities.

Results

The experimental pipeline used in this study is shown in **Figure 3.1A**. First, we designed variants of AmiE that possess wild-type catalytic activity but with a reduced probability of folding in vivo. Second, we developed selection conditions for the variants under which cell growth is proportional to enzyme activity using a growth selection with acetamide as the sole nitrogen source. Third, near-comprehensive single-site saturation mutant libraries for our variants were prepared[25] and growth selections performed. Fourth, pre- and post-selection populations were deep sequenced to extract mutant frequencies in the selected and reference populations. These frequencies were converted into a relative fitness metric (ζ_i) for each mutant i defined as

$$\zeta_i = log_2(\frac{\mu_i}{\mu_{REF}}) \quad (1)$$

where μ_i and μ_{REF} represent the specific growth rates in selection media for the mutant (μ_i) and unmutated AmiE variant (μ_{REF}), respectively. A relative fitness score above zero means that a strain harboring a given mutant has higher fitness than those carrying the unmutated variant. It is important to understand the limitations of deep mutational scanning experiments. While these experiments provide quantitative measurements of fitness relative to the respective backgrounds, deep mutational scanning cannot provide information on why a given mutation is beneficial or deleterious. Thus, specifying whether a mutation impacts stability, catalytic efficiency, on-path folding rate, etc. is left to reasoned speculation or further biophysical analysis.

AmiE variants with lower in vivo folding probabilities and wild-type catalytic efficiencies engineered

We first sought to identify mutations to AmiE that, under the selection conditions, would decrease the in vivo folding probability of the protein while maintaining wild-type catalytic efficiency. To identify such mutants, we chose to use a computational approach by modifying PROSS[26]. Briefly, PROSS designs a protein sequence that will have an improved probability of reaching the folded state in vivo relative to its input. This improved folding probability correlates with biophysical properties like improved protein stability, faster on-target folding rate, or reduced aggregation propensity. As our experimental objective is essentially the inverse problem, we modified the Rosetta FilterScan protocol undergirding PROSS and then selected point-mutations with higher energy scores relative to AmiE wild-type (WT) (**Figure 3.1B, Table A 1**). For each mutant these scores were then cross-referenced with experimental relative fitness scores previously determined for AmiE[25] to ensure that their relative fitness was below zero (**Table A 1**).

Of thirteen variants with 1-3 mutations from WT selected for experimental characterization, nine expressed as soluble proteins in E. coli BL21* (DE3). We purified a subset of these nine variants and assessed their catalytic efficiency with the substrate acetamide. While most mutants showed reduced enzymatic activity, both AmiE I38V and AmiE I122L showed statistically indistinguishable maximum turnover rates (kcat) and Michaelis constants (KM) compared with WT (**Figure 3.1C, Table A 2**). Furthermore, size exclusion chromatography showed no oligomeric differences between AmiE WT and AmiE I38V or AmiE I122L (**Figure**

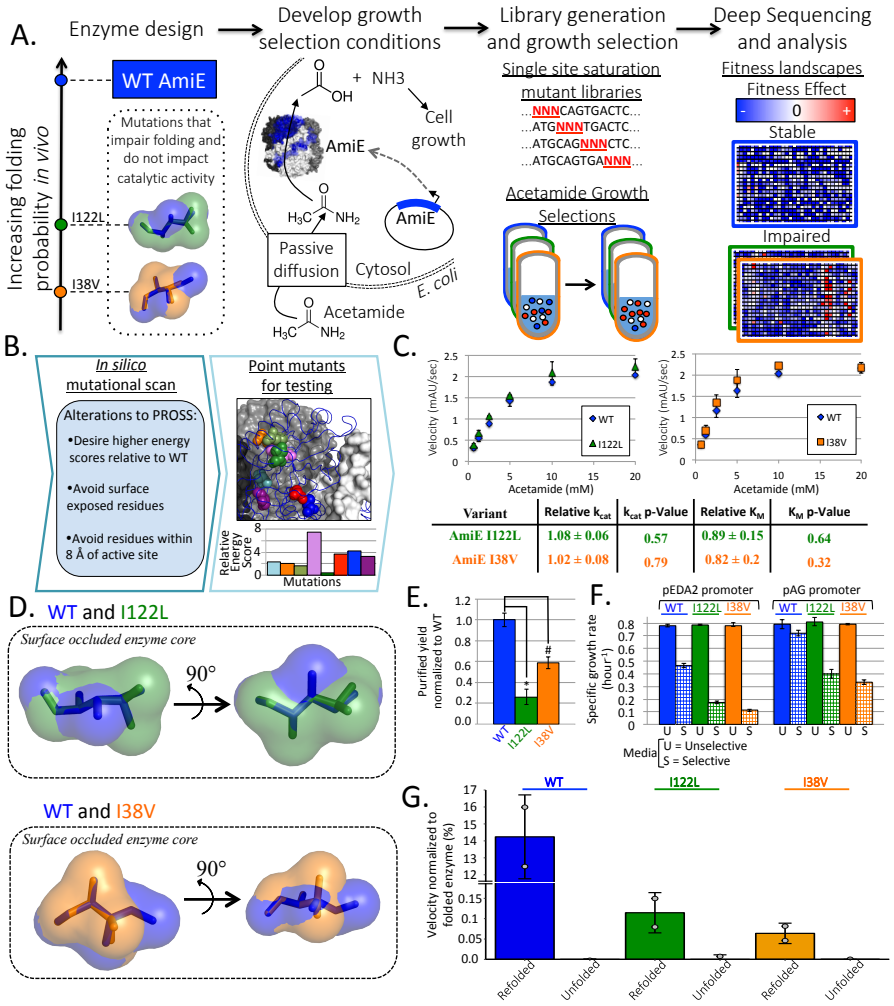

Figure 3.1: Design of deep mutational scanning experiment A. A graphical overview of this study. Two AmiE enzyme variants with single point mutants (I38V and I122L) with WT catalytic function and lower probabilities of folding *in vivo* were computationally designed and validated experimentally. Constitutive expression of each enzyme from a plasmid was tuned such that the growth rate of our bacterial growth selection strain in selection media was dependent on the expression of functional AmiE. Deep mutational scanning was performed on these variants and compared with WT AmiE. B-G. Design and validation of AmiE variants. B. A graphical representation of the computational enzyme design. C. Enzyme velocity as a function of acetamide and Michaelis-Menten parameters determined relative to WT AmiE. Error bars = 1 s.d., n = 2, p-values obtained using Student's t-test. D. Structural modeling of the cavities introduced into AmiE by designed mutations. E. Enzyme yield following *E. coli* auto-induction expression. Error bars = 1 s.d., n = 3, * = p-value = 0.0003, # = p-value = 0.002. F. Specific growth rates of strains in M9 (unselective) and in M9 with 10 mM acetamide as sole nitrogen source (selective) (pEDA2 - low expression, pAG - high expression). Error bars = 1 s.d., n ≥ 3. G. Comparison of enzyme reaction velocities at substrate saturation relative to a folded control. Grey dots represent biological replicates, Error bars = 1 s.d., n = 2.

A 1) in PBS at 30 μM, suggesting that the variants maintain the expected homohexameric quaternary structure. Finally, the secondary structure of the three AmiE proteins was analyzed using far-UV circular dichroism spectroscopy (**Figure A 2**), revealing indistinguishable spectra in the folded and unfolded states.

AmiE I38V removes a methyl group to open a small cavity in the core, while I122L modulates hydrophobic core packing in the monomer subunit (**Figure 3.1D, Figure A 3**). Both mutations are located in the hydrophobic core distal from the dimeric and homohexameric contacts necessary for quaternary assembly (**Figure A 3**). Based on Rosetta analysis, we predict that these mutations disrupt the core of AmiE resulting in thermodynamic destabilization of the native monomer. However, mutations in the stability cores of proteins can disrupt the hierarchy of folding[27] through the destabilization of folding intermediates[28], by limiting the intermediate states accessible during folding[29], and by decreasing the thermodynamic stability[30] of the monomeric subunits outside of the quaternary structure[31].

To distinguish among these possibilities, we attempted tryptophan fluorescence unfolding measurements using guanidinium-HCl (Gdn-HCl) as a denaturant to determine the effective thermodynamic stability. However, AmiE WT aggregated in moderate Gdn-HCl concentrations under most conditions (data not shown), and under conditions of no aggregation and complete unfolding no isosbestic point was recovered (**Figure A 4**). This lack of an isosbestic point indicates more complicated reversible folding at 4°C than simple 2-state models. We also performed thermal shift assays with the purified homohexameric enzymes in a series of dilutions (10, 5, 1, 0.5, 0.25 μM) (**Figure A 5**) to measure thermal stabilities. Two-state irreversible unfolding curves were obtained at 10 and 5 μM. Analysis of the melting curves reveals statistically indistinguishable melting temperatures between WT and variants (**Table A 3**). This

data suggests that all have a single transition from homohexamer into unfolded monomers. It is well established that oligomeric proteins are often more stable than in their natively folded monomeric or dimeric forms[31]. Thus it is likely that the thermal melt is measuring the stability of homohexameric assembly, which would be expected to be identical between WT and variants as neither mutation resides at an oligomeric interface. To identify AmiE concentrations for which the monomeric form is favored, we reasoned that hexameric dissociation would result in inactive enzyme[32], which could be measured colorimetrically using our established activity assay. Indeed, activity analysis of dilute solutions of AmiE WT at 500 nM showed larger decreases in activity than 900 nM over moderate incubation periods (**Figure A 6**). Unfortunately, usable signal was not detected at protein concentrations of less than 5 μM in the thermal shift assays (**Figure A 5**). To assess thermal denaturation at lower enzyme concentrations, we performed circular dichroism thermal melts using a protein concentration of 1 μM (**Figure A 7**). Under these conditions both AmiE I38V and AmiE I122L have modest but significantly lower melting temperatures than AmiE WT (p-value 0.01 for I38V and 0.03 for I122L; **Table A 4**).

While we were only able to establish moderate decreases in the thermal stabilities of the AmiE variants, complementary *in vitro* and *in vivo* experiments strongly support that both I122L and I38V variants have lower in vivo folding probabilities than WT in the general order: I38V<I122L<WT. All synonymous codons encoding the I38V and I122L mutations had fitness metric below zero (**Table A 1**), suggesting the loss of fitness is a result of changes at the protein level rather than effects resulting from the codon used[25]. When driven from the same T7 promoter under identical Studier auto-induction[33] protein expression conditions, both AmiE I38V and I122L have statistically significant lower purification yields of soluble protein than WT (**Figure 3.1E, Table A 2**). Furthermore, E. coli harboring the AmiE variants expressed from

the same plasmid – pEDA2[25] maintaining the same constitutive promoter, ribosome-binding site (RBS), and 5' untranslated region (5' UTR) – showed lower specific growth rates than WT when grown with acetamide as the sole nitrogen source (**Figure 3.1F, Table A 2**). These results suggest that the *in vivo* folding probability upon translation for these variants is lower than for WT. Finally, while denatured WT can refold into active enzyme at 14.2% yield, both I38V (0.06%) and I122L (0.11%) have vastly lower refolding yields (**Figure 3.1G, Table A 2**). These results together support a model where the I38V and I122L mutations result in a lower probability of correctly folding into active homohexameric enzyme in vivo.

Deep mutational scans for AmiE variants

Deep mutational scanning of these variants was performed using a previously developed growth selection[25] in media with 10 mM acetamide as the sole nitrogen source. These growth selections required tuning the constitutive amidase expression such that the specific growth rate of variant i expressed in E. coli MG1655 rph+ in the selection media relative to that in defined minimal media ($\mu s,i/\mu M9,i$) is 0.4-0.6. However, plasmid pEDA2 used for AmiE WT selections did not support high enough growth rates for the I38V and I122L variants (**Figure 3.1F, Table A 2**). Thus, we screened additional promoters for AmiE I38V and AmiE I122L while maintaining the same 5' UTR and RBS for all constructs in order to minimize potential variant-dependent mRNA effects on fitness (**Table A 5**). Plasmid pAG with a stronger constitutive promoter than pEDA2 supported a growth rate ratio of 0.49 ± 0.03 for AmiE I122L and 0.42 ± 0.02 for AmiE I38V (**Figure 3.1F, Table A 2**). By contrast, pAG AmiE WT had a nearly 2-fold higher growth rate ratio of 0.91 ± 0.04 (**Figure 3.1F, Table A 2**).

Next, we generated near comprehensive single-site saturation mutant libraries using nicking mutagenesis[34] (**full library statistics are shown in Table A 6** and **Table A 7**). For AmiE I38V mutations at residues 32-44 flanking the site of the disrupting mutation were not constructed, while for AmiE I122L mutations at residues 115-130 and 132 were not made. Plasmids expressing mutant enzyme libraries were electroporated into E. coli MG1655 rph+ under conditions minimizing double transformants. Then, strains harboring AmiE libraries underwent growth selections in replicate with initial population sizes of $>6x10^6$ cells for approximately 8 generations at 37°C. A biological replicate for AmiE WT covering residues 171-255 was also performed to compare with previous published results[25]. The pre- and post-selection populations were barcoded and deep sequenced. The resulting data was processed using PACT[35] to obtain the relevant fitness metrics for each mutant in the library. The depth of sequencing ranged from 155 to 300-fold coverage for the libraries (**Figure A 8**). In total, we recovered the relative fitness metrics for 93.7% and 91.8% of all possible non-synonymous mutants for AmiE I122L and AmiE I38V, respectively (**Table A 7**).

To estimate reproducibility, we compared the AmiE WT replicate selections performed here with data from an identical selection experiment performed in Wrenbeck et al.[25]. Correlation coefficients between mutation-specific fitness values in selections are ≥0.90 (**Figure A 9**), which is comparable to correlation between replicates performed for this work (AmiE I122L - 0.921; AmiE I38V - 0.952) (**Figure 3.2A**). Additionally, there was essentially no correlation between relative fitness and pre-selection frequency of a given mutant in the library, (WT AmiE – R = 0.011, AmiE I122L – R = 0.026, AmiE I38V – R = -0.0376) (**Figure A 10**) indicating that pre-selection read counts do not bias the fitness metrics obtained.

Distribution of beneficial fitness effects are largely insensitive to initial in vivo protein folding probability

The shape of the distribution of fitness effects (DFE) governs the local protein fitness landscape. Realizing that beneficial mutations are rare, the likelihood of finding beneficial mutations was predicted by Orr[36] to follow the Pareto family of distributions. Using the set of beneficial mutations – variants with relative fitness above wild-type under selective media – we were previously able to describe the shape of the DFE for beneficial mutations as exponential with high statistical power[25]. The new datasets allow us to ask directly whether the shape of DFE changes with respect to enzyme *in vivo* folding probability. Consistent with expectations, all variants have very similar distributions of fitness effects (**Figure 3.2B and Figure A 11**) with a tight range of total possible mutations that are beneficial. For all variants the Pareto family of functions also describes their distributions of beneficial fitness effects (**Table A 8**). Thus, given approximately the same relative fitness, the probability of finding rare beneficial mutations is independent of initial likelihood of native folding.

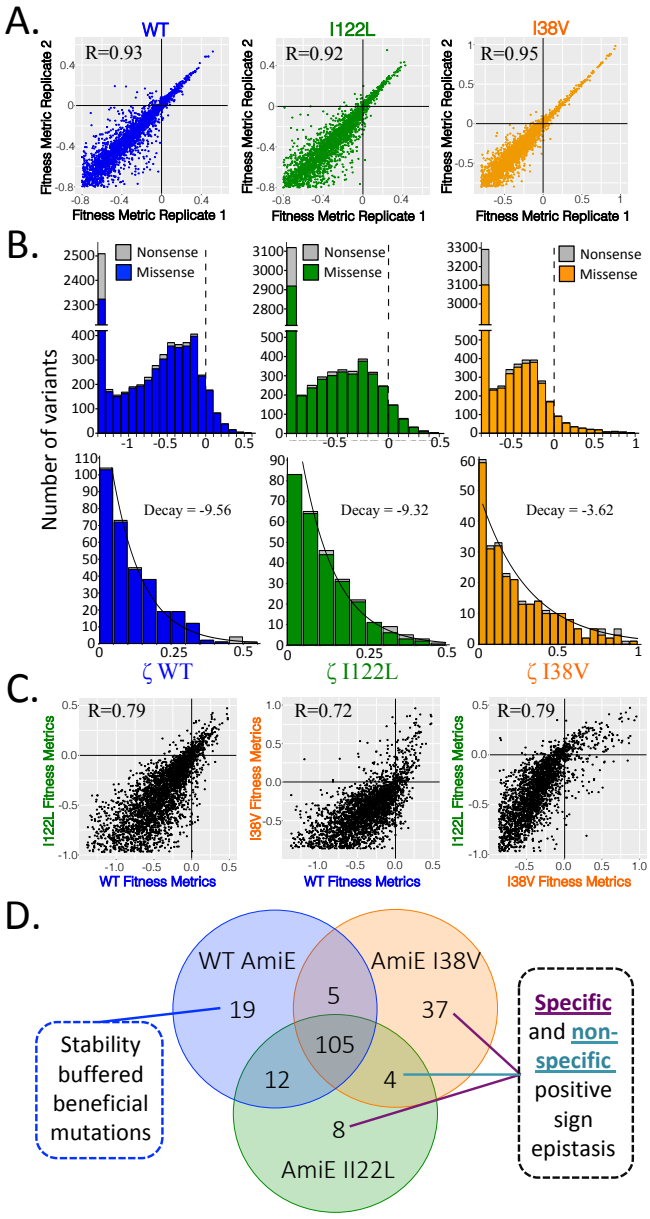

Figure 3.2: Local fitness landscapes are nearly insensitive to initial protein folding probability *in vivo*. A. Correlation between AmiE variant technical replicates. B. Distributions of fitness effects (DFE) of nonsense and missense mutations for the AmiE variants. Upper plots show full DFE, while lower plots include only beneficial mutants with best-fit exponential curve. C. Correlation of fitness between AmiE variants following the combination of replicate datasets and reprocessing using PACT. D. Venn diagram for all unique and shared beneficial mutations for the respective variants.

Moderate epistasis observed with decreasing enzyme in vivo folding probability

How does the local fitness landscape change in response to a single deleterious point-mutation that alters only the *in vivo* folding probability of an enzyme? If mutations were completely additive with the disrupting mutations, we would expect the comparison of the local fitness landscapes for the enzymes to have 1:1 correlations and approach the correlation coefficients found between replicates (R ~0.92). On the other hand, complete non-additivity of mutations would lead to minimal correlation. We were able to compare 2,813 mutations above the lower bound of relative fitness (45.4% of possible mutations) shared between the three datasets. Pearson's correlation analysis of the DFE finds that the WT local fitness landscape is reasonably correlated with that of the variants (WT vs. I38V R= 0.72, WT vs. I122L R = 0.79), and this correlation is similar to that between I122L vs. I38V (R= 0.79) (**Figure 3.2C**). Notably, these correlation coefficients are lower than for replicates. Furthermore, linear regression best fits show lower slopes between variants than within replicates (**Figure 3.2A** and **3.2C**).

Next, we tested for increased negative epistasis in our distribution of fitness effects. We predicted that the greater *in vivo* folding probability of AmiE WT provides a buffering effect, which could temper many deleterious mutations. To test this hypothesis, we generated empirical cumulative distribution functions (ECDF) for the deleterious mutations for each of the three enzymes (**Figure A 12 and Table A 9**). This analysis is limited to the range of deleterious mutations quantitatively captured in our experimental system[23]. Within this range, the application of the Kolmogorov-Smirnoff test failed to reject the null hypothesis that the ECDF of AmiE WT is below that of AmiE I38V (**Table A 9**). By the same test, we were unable to discriminate the ECDFs of AmiE WT and AmiE I122L (**Table A 9**). Therefore, AmiE I38V has a greater number of more deleterious mutations than AmiE WT. This indicates that the lower *in*

vivo folding probability of the AmiE I38V background increases the likelihood of negative epistatic effects when compared to the more resilient AmiE WT.

Precise measurements of negative and positive epistasis are complicated by the relatively narrow range of fitness our experimental system captures. However, we can determine the sign of fitness in our datasets with high precision. To evaluate the relative prevalence of sign epistasis, we defined any mutant as beneficial if $\zeta_i > 0$ for both replicates and if $\zeta_i > 0$ within a 95% confidence interval (see **Materials and Methods**). Conversely, we define a mutant as deleterious if $\zeta_i < 0$ for both replicates and if $\zeta_i < 0$ within a 95% confidence interval. We used these cutoffs to sort beneficial variants into the seven possible fitness bins (**Figure 3.2D**). Using a stricter requirement - that beneficial mutants are defined as those with a ≥10% increase in specific growth rate over the genetic background - leads to similar results (**Figure A 13**).

Most beneficial mutations in the WT background are shared

Previous studies found that stable proteins can buffer destabilizing mutations that are otherwise beneficial[37,38]. We are able to assess the extent of this phenomenon in our datasets. We find that 122/141 (86.5%) of beneficial mutations in the WT background are also beneficial in the I38V and/or the I122L genetic background (**Figure 3.2D**). Of these, six globally beneficial mutations (S9A, A28R, R89E, I165C, V201M, A234M) have been previously characterized biophysically[25] and are known to improve specific amidase flux under the selection conditions of 10 mM acetamide at 37°C. Conversely, only 19 of 141 beneficial mutations (13.5%) are specific in the WT background (**Figure 3.2D**). Therefore, beneficial mutations that are buffered in the stable background are present but in the minority.

The 19 WT-specific beneficial mutations map to two predominant locations: eleven (G291C, L297H, D311C, E320A, S325Y/T/C, 326Y/H, R336L, G341Y; **Figure 3.3A**) are at the extreme C-terminus that creates extensive homodimer contacts, while four (M72T, E74G, A78H, E82I; **Figure 3.3A**) are located on helix B and helix C distal to any oligomeric contacts in the homohexamer. The majority of the C-terminal mutations are adjacent to the homodimerization interface, which we speculate could lead to subtle structural rearrangements in the AmiE active site. Probable mechanisms behind the helix B/C mutations are more obscure as three of these mutations are at surface exposed positions over 10 Å away from any active site residue. We note that active site-induced effects may still be quite strong even this far away[14,25,39-41]. Regardless of the exact mechanisms, these findings indicate localized regions where small-scale mutational perturbations lead to increased fitness in a more stable genetic background.

To avoid information loss from the simple categorical separation of mutations into bins, normalized linear regression analysis was performed on correlation plots for the beneficial mutations shared in all backgrounds, in which the best-fit linear regression for the respective correlation plots was determined and normalized such that $Y = 1X + 0$ **(Figure A 14**). In these correlation plots the mutational effects are highly disperse and widely deviate from the predicted fitness metric function for non-epistatic mutational combinations. The regression normalized AmiE I38V **(Figure A 14A)** and AmiE I122L (**Figure A 14B)** datasets compared with WT both show decreased correlations (AmiE I38V Pearson's $R = 0.62$; AmiE I122L Pearson's $R = 0.66$) compared to within replicates (Pearson's $R > 0.92$). These distributions indicate that both classes of epistatic effects are present and imposing large impacts on this population of beneficial mutations.

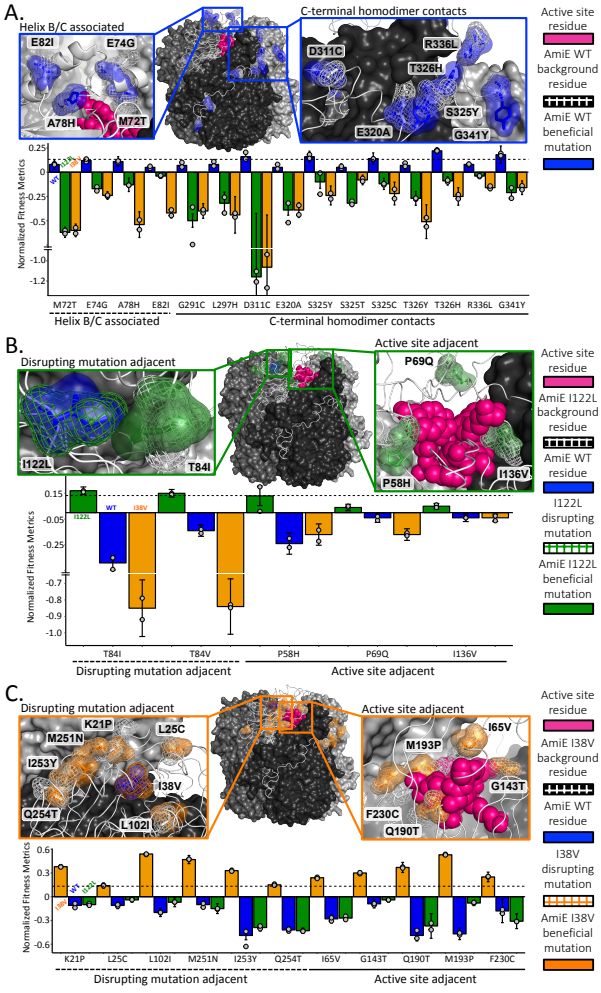

Figure 3.3: Positive sign epistatic mutations are spatially segregated and specific. A-C. Bar graphs of fitness metrics. Error bars represent 95% confidence intervals calculated from Poisson errors inherent in deep sequencing (Klesmith et al. 2015), while grey dots are fitness metric for each replicate. Horizontal dotted lines represent the cutoff value for mutations that increase the growth rate by ≥10%. Models show: trimer of dimers (wire + surface models), background residues for a respective enzyme (white mesh + sticks), the active site residues (magenta spheres), and where applicable the original mutation (green or orange mesh + sticks). In AmiE I122L and AmiE I38V the unique beneficial mutations tend to cluster around the disrupting mutations or near the active site. A. AmiE WT unique beneficial mutations are located at the C-terminal tail or in the B/C helices. B. Location of AmiE I122L unique beneficial mutations segregate to either positions adjacent to position 122 (T84I/V) or adjacent to the active site. C. Locations of a subset of AmiE I38V unique beneficial mutations. In B and C the blue transparent surface with sticks represents the residue in AmiE WT, and green or orange transparent surfaces with sticks represent respective unique beneficial mutations.

Positive sign epistasis is overwhelmingly specific

Our datasets allow direct comparisons between the prevalence of specific and nonspecific epistasis in a folding impaired background. Contrary to previous literature on different model proteins[2,15,38, 42-45], we find that specific reciprocal positive sign epistatic mutations dominate nonspecific mutations in the I38V and I122L backgrounds. In particular, we find 8 specific – unique beneficial – mutations for AmiE I122L and 37 specific mutations for AmiE I38V, compared with only 4 nonspecific – shared beneficial – mutations (**Figure 3.2D and 3.3B-C**). Analysis of the locations of the unique mutations in the structures of AmiE I122L and AmiE I38V suggest biophysical interpretations of their epistatic mechanism. For AmiE I122L the strongest specific mutations T84I/V directly contact 122L (**Figure 3.3B**), while similar mutations (K21P, L25C, L102I, M251N, I253Y, Q254T) occur on loops adjacent to I38V (**Figure 3.3C**). Other specific mutations line the enzyme active site. For AmiE I122L (**Figure 3.3B**) there are three (P58H, P69Q, I136V), while in AmiE I38V (**Figure 3.3C**) there are five (I65V, G143T, Q190T, M193P, F230C). Notably, since our datasets do not include immediately adjacent mutations for I38V and I122L, the extent of specific positive sign epistasis is probably underestimated.

A plurality of unique beneficial mutations is codon-dependent

Fitness conferred by a weak-link enzymes depends on intrinsic protein biophysics but also on mRNA sequence-dependent effects. Perhaps the best appreciated of these are synonymous mutations in the first ten codons of a polypeptide because they can substantially alter mRNA stability and access to the ribosome binding sites in bacteria[25,46]. Additionally, synonymous codons can differentially affect cotranslational folding[47] and impact fitness. Here

we mapped the variance between synonymous codons encoding beneficial missense mutations (**Figure 3.4A**). For all three variants, the majority of high variance codons occur in the first 10 codons, as expected (**Figure A 15A**). However, there were localized punctae of high variance at several downstream positions for both WT and I38V datasets corroborated in replicate measurements. While overall variance in beneficial synonymous codons is weakly or not statistically significant among datasets (**Figure A 15B**), contingency table analysis of synonymous codon fitness disparities for the unique and shared beneficial mutations finds correlation with unique beneficial mutations in the WT and I38V backgrounds (WT p-value = 2.5x10-8, I122L p-value = 0.053, I38V p-value = 0.00017; 2-tailed Fisher exact probability test)

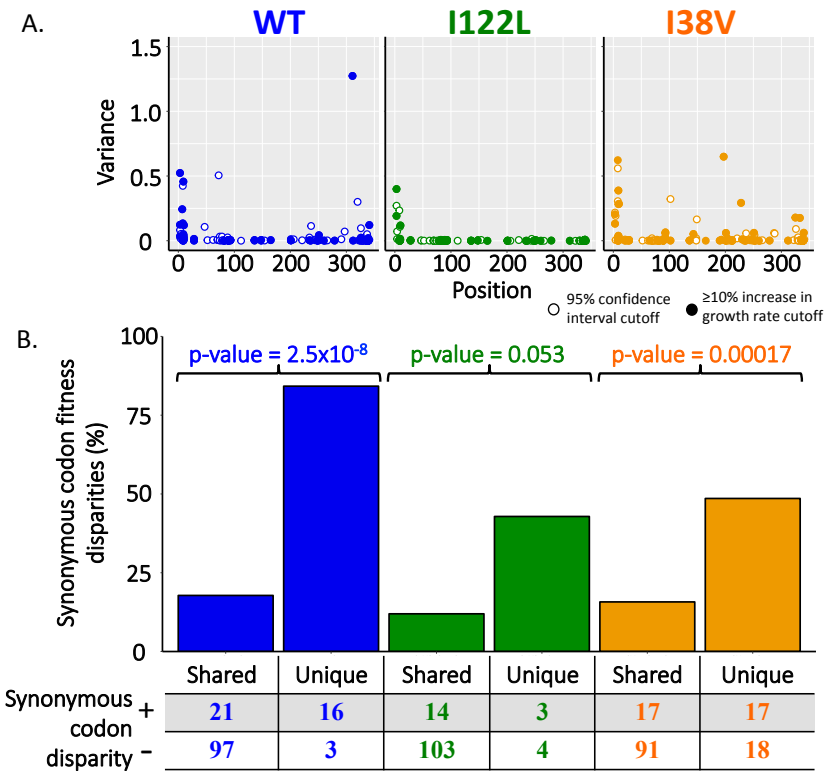

Figure 3.4: Unique beneficial mutations have high percentages of synonymous codon fitness disparities. A. Variance of fitness metrics for synonymous codons of beneficial mutations as a function of position in the primary sequence. B. Percentage of shared and unique beneficial mutations with synonymous codon fitness metric disparities. p-values reported are from contingency table analysis with 2-tailed Fisher exact probability test.

(**Figure 3.4B**). In fact, the vast majority of WT (84.2%) - and many of the I122L (42.8%) and I38V (48.6%) - unique beneficial mutations have synonymous codon fitness sign disparities (**Figure 3.4B**). This indicates that transcriptional-translational effects impose significant evolutionary constraints.

Discussion

In this study we used deep mutational scanning to analyze how the probability of attaining the folded, active state impacts local fitness landscapes. AmiE I38V and AmiE I122L were designed and validated to have identical catalytic parameters to WT but have a lower probability of folding under selection conditions. We found that the DFE for both variants was largely similar to the AmiE WT, and that most fitness-enhancing mutations are shared. However, there were two major surprises found when analyzing the set of mutations exhibiting positive sign epistasis.

First, we expected there to be a larger subset of beneficial mutations shared only between the I122L and I38V datasets, as current models of stability-induced epistasis posit that many nonspecific globally-distributed mutations can improve the probability of folding[3]. In contrast, we found that sign epistatic mutations were overwhelmingly specific for the I122L and I38V backgrounds. It is possible that beneficial non-specific sign epistatic mutations are the minority because of the alterations to in vivo expression that are required for the deep mutational scans. In the non-promoter tuned E. coli the impaired enzymes provide very weak cell growth. It is possible that non-specific epistatic effects may have arisen to be equivalent with, or dominant to, specific epistatic phenomenon if experiments could be performed without increasing the in vivo expression of the impaired enzymes. A limitation of our deep mutational scanning experiments is

that cell growth must be proportional to enzyme function and must operate within a window of growth rate ratios[23]. This requires increased in vivo expression of the impaired enzymes. By increasing the in vivo expression it is possible that the selective advantages that the non-specific mutations might possess could be dampened. Alternatively, the protein folding pathway for homohexameric AmiE in E. coli is potentially much more complicated than model systems of monomeric, single domain proteins that have built much of the current intuition about stability and epistasis. Therefore, the sparsity of non-specific beneficial mutations could be an artifact of our experimental system. These considerations also may explain the lack of a neutral peak for fitness expected from previous theoretical[48] and experimental[49] datasets on other proteins. More careful measurements on a wider array of oligomeric proteins should resolve this seeming contradiction.

As a second surprise, we found that unique beneficial mutations strongly depend on codon choice, as approximately 50% of sign epistatic mutations in the I38V background show sign disparities. We speculate that this unexpected result arises from the complicated co-translational folding in vivo of the homohexameric AmiE. Local, specific nonsynonymous mutations may recover on-target folding trajectories more efficiently than nonspecific, globally stabilizing mutations. Similarly, on-pathway folding kinetics may differ considerably between variants, which can be selectively modulated by codon choice[50,51]. As an alternative explanation, Kudla and colleagues recently report that synonymous codons can exert fitness effects through RNA toxicity itself[52] through an unknown mechanism.

While we were able to determine that both I122L and I38V mutations decrease the probability of active AmiE expression at 37°C, we were unable to measure the relative stability of the monomeric proteins. For the thermodynamic studies, AmiE aggregated under most

conditions, and where suitable conditions were found the lack of an isosbestic point hampered analysis. The thermal melts showed a single transition, most likely due to the hexameric dissociation. Lowering the AmiE concentration supported monomer formation at 25°C but yielded too weak a signal for analysis of monomer thermal stability. CD melts at low AmiE concentrations did show a modest but significant decrease in Tm in the variants relative to WT. Nevertheless, our results show that simple biophysical models currently used to model protein evolution are incomplete and that biophysical models may need to use kinetic models to account for the folding probability *in vivo*.

Materials and methods

Reagents

All antibiotics were purchased from GoldBio and all purchased enzymes were from New England Biolabs. All other chemicals were purchased from Sigma-Aldrich. Primers and mutagenic oligos were purchased from Integrated DNA Technologies and were designed using either Benchling (www.benchling.com) or the Agilent QuikChange Primer Design Program (www.agilent.com).

Computational design of folding impaired mutants

The FilterScan Rosetta script[26] was modified to predict mutations that would decrease thermodynamic stability without altering catalytic efficiency. The structural coordinates for AmiE[53] (PDB: 2UXY, 341 residues per monomer) were taken from the Protein Data Bank and prepped for use in Rosetta scripts through the 'clean_pdb_keep_ligand.py' script released with Rosetta 3[54]. The crystal structure data was refined through the "refine.xml" Rosetta scripts

(unaltered) from Goldenzweig et al.[26]. To avoid impacting catalytic efficiency residues within 8 Å of the active site were excluded from the FilterScan protocol. Additionally, surface residues were predicted and excluded from computational testing to avoid disturbing the native homohexameric state. The FilterScan script was modified to remove the input from the position specific scoring matrix (PSSM) evolutionary conservation term. Mutants with scores that predicted destabilization of the enzyme and were also shown to decrease relative fitness in a previous study[24] were selected for biophysical analysis.

Plasmid construction

The variants selected for biophysical analysis were constructed by mutating the AmiE WT sequence using the single mutation protocol of Nicking Mutagenesis[34]. The pEDA3 constitutive expression plasmid from Wrenbeck et al.[34] was used as the vector for the mutagenesis. Variants were subcloned from the pEDA3 background into protein expression plasmid pET-29b(+) (Novagen) or into the constitutive expression plasmid pEDA2[25] at *NdeI* and *XhoI* sites using classic restriction cloning. Plasmid pAG was constructed by mutating the -10 and -35 promoter regions of the pEDA3 plasmid using the multi-site nicking mutagenesis protocol from Wrenbeck et al.[34]. Following mutagenesis, the mutant promoter libraries were transformed into the *E. coli* growth selection strain MG1655 rph+ [F- λ-] (Coli Genetic Stock Center #7925, CGSC strain designation: BW30270). All AmiE variant DNA and protein sequences are listed in **Supplementary Notes A 1** and **2**.

Near comprehensive single-site mutant library construction

Near comprehensive mutant libraries for AmiE I38V and AmiE I122L were constructed using the comprehensive nicking mutagenesis protocol from Wrenbeck et al.[34]. The genes for both impaired enzymes were broken into the following tiles: tile 1 (residues 1-85), tile 2 (residues 86-170), tile 3 (residues 171-255), and tile 4 (residues 256-341) in pAG. For AmiE I38V 13 residues were excluded from mutagenesis (residues 32-44), while for AmiE I122L 17 residues were excluded from mutagenesis (residues 115-130 and 132). Nicking mutagenesis products were amplified and purified as in Klesmith et al.[24]. 10 ng each of the respective DNA libraries was transformed into *E. coli* MG1655 *rph*+ by electroporation performed with either a 1 mm electroporation cuvette at 1200 V (AmiE I122L; AmiE WT), or a 2 mm electroporation cuvette at 1600 V (AmiE I38V) using an Eppendorf Eporator. All experimental and control libraries were transformed into the selection strain with greater numbers than that required for theoretical complete library coverage (**Table A 6**). The transformation procedure for the selection strain was optimized to minimize double plasmid transformants as described in Kowalsky et al.[23]. -80°C freezer cell stocks of the libraries were prepared as detailed in Klesmith et al.[24].

Protein expression and purification

pET29(b) constructs harboring genes encoding AmiE variants were transformed into *E. coli* BL21*(DE3) cells (Invitrogen) and expressed using Studier auto-induction[33] at 22°C for 16-18 hours. Cultures were pelleted and frozen at -80°C. Proteins were purified from cell pellets by Ni-NTA affinity chromatography exactly as described in Klesmith et al.[24]. Purified enzymes were desalted into phosphate buffered saline (PBS; 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 2.7 mM

KCl, 137 mM NaCl, pH 7.4) using disposable PD-10 desalting columns (GE Healthcare), sterilized through a 0.22 μm syringe filter, and stored at 4°C until analysis. Purified enzymes showed as a single band by SDS-PAGE (**Figure A 16**). Purified enzyme solutions were quantified using the absorbance at 280 nm in 1x PBS using a published[55] theoretical A_{280} molar extinction coefficient of 56,980 M^{-1} cm^{-1}.

For quantitative comparison of purified product yields under the T7 promoter system in BL21* *E. coli*, slight alterations to the above induction scheme were used. For this analysis induction cultures were always started at an OD_{600} of 0.005 from 1 mL LB + 50 μg/mL kanamycin cultures grown at 37°C overnight. Next, 500 mL induction cultures were inoculated at an OD_{600} of 0.005 with the overnight cultures and grown at 37°C with shaking at 250x rpm for 6 hours, and following this growth step cultures were moved to 22°C and induced for ~17 hours. Induction cultures were then pelleted and the wet cell weights of the pellets recorded. Cultures were then purified, desalted, and quantified as described above.

Biophysical analysis of proteins

Analysis of the growth rates of the AmiE variants in the respective constitutive expression plasmids in MG1655 rph+ *E. coli* were performed exactly as in Wrenbeck et al.[25]. Assessment of the oligomeric state of the purified enzymes was performed using SEC-FPLC. Approximately 3 mL of 30 μM of the purified enzymes in PBS were run on an AKTA-FPLC system at 1 mL/min on an HiLoad 16/600 Superdex 200 column equilibrated with PBS. Enzyme kinetics (K_M and k_{cat}) was determined via phenol-alkaline hypochlorite end-point activity assays exactly as in Wrenbeck et al.[25]. Kinetic analysis of purified WT and folding impaired AmiE variants was performed within 6 days of purification.

To test for an isosbestic point the AmiE variants were denatured in guanidinium-HCl (GDN-HCl) and native tryptophan fluorescence was detected. Purified AmiE WT was diluted to 52 μM in increasing amounts of ice cold GDN-HCl (0 – 4 M) in PBS with 1 mM DTT and mixed gently. 200 μL of the denaturation mixtures were placed into opaque black 96 well plates and covered with optical film. Plates containing the samples were incubated at 4°C for 8 hours. Next, the native tryptophan fluorescence was measured in uncovered plates using an excitation of 290nm and emission of was detected over a range of wavelengths: 310 nm – 370 nm. This lack of an isosbestic point indicates an unfolding model that is more complex than a two-state model, presumably because of monomer folding and homohexamer association.

Thermal shift analysis was performed exactly as described in Wrenbeck et al.[25] with incubations for 2.5 hours at 25°C prior to addition of the dye and initiation of the thermal shift assays; the data was processed as in Huynh et al.[56]. In brief 10, 5, 2.5, 1, 0.5, and 0.25 μM purified AmiE in 1x PBS was incubated at 25°C for 2.5 hours. Next, the samples had 5 μL of 200x SYPRO-orange dye (Life Technologies) added to 45 μL of the diluted enzymes in 0.1 mL MicroAmp® 96-Well Reaction Plate (Life Technologies) and covered with MicroAmp™ optical film (Life Technologies). Thermal melt analysis was performed in a QuantStudio 6 Flex RT-PCR device (ThermoFisher). The melt ranged from 25°C to 98°C with 1°C change per minute and with a 2 minute incubation at the first and last temperatures. Thermal melt analysis with circular dichroism, and the far-UV spectral analysis of the folded and unfolded enzymes, was performed on a Chirascan plus spectrophotometer (Applied Photophysics). Purified AmiE was buffer exchanged into 10 mM phosphate buffer pH 7.5 using PD-10 desalting columns (GE Healthcare) approximately 24 hours prior to analysis and stored at 4°C. 1 μM samples were diluted in the 10 mM phosphate buffer and stored in a 25°C water bath for at least 12 hours prior

to analysis. Prior to initiating the thermal melt analysis circular dichroism spectra were obtained from 180 nm to 260 nm at either 15°C (10 μM samples) or at 25°C (1 μM samples) in a 0.5 mm cuvette with 0.5 seconds per-time-point. Thermal melt analysis of the 1 μM samples started at 25°C and ramped to 95°C at a rate of 1°C/min with 0.5°C steps and a tolerance of 0.2°C. Signal at a wavelength of 222 nm was measured every 24 seconds throughout the melt. After the samples had reached the maximum temperature, the far-UV spectra were measured as before but with the temperature set at 95°C for all samples. Following circular dichroism analysis the obtained spectra were adjusted for their respective buffer blanks and smoothed using the Savitsky-Golay filter. The Window Size of the Savitsky-Golay filter was set to 14 for smoothing all buffer blanks prior to adjusting the sample spectra, and to 6 when smoothing the adjusted sample spectra. Finally, the obtained data was converted to Mean Residue Ellipticities (millideg * cm2 * dmol-1) prior to curve fitting. Boltzmann curve fitting was performed to determine the apparent T_m for both thermal melt experiments using GraphPad Prism (www.graphpad.com).

To assess relative activity of the refolded enzymes, enzymes were first denatured in ice cold 3 M GDN-HCl in PBS supplemented with 1 mM DTT for 16 hours at 4°C at a final concentration of 50 μM. The solution was then diluted 50-fold into PBS with 1 mM DTT and 0.1% (w/v) BSA at 4°C in 96-well PCR plates that had been blocked with 1% (w/v) BSA in PBS for 1 hour at 37°C, resulting in a total protein concentration of 1 μM. Refolding mixtures were then incubated at 4°C for 5 minutes, and warmed to 37°C over 4.5 minutes. Samples were then held at 37°C for 20 minutes and then cooled to 4°C and held there until assaying. Immediately prior to assaying, samples were diluted in fresh ice-cold PBS with 1 mM DTT to ensure linearity in the activity assays. Enzymes were assayed using the phenol-alkaline hypochlorite end-point assay with 20 mM acetamide as the substrate. As a control, enzymes went through the same steps

as above except without initial GDN-HCl denaturation. The percent activity of the refolded enzyme was determined relative to the control sample. Two biological replicates were performed for all reactions. Refolding experiments were performed within 7 days of purification of the enzymes from the pellets.

Growth selections

Growth selections were performed exactly as in Wrenbeck et al.[25]. Briefly, starter cultures were grown overnight in the non-selective media (M9 minimal media: 47.6 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.54 mM NaCl, 18.68 mM NH_4Cl, 50 μg/mL carbenicillin, pH 7.0) and the following day the cells were washed in ice-cold M9 salt solution without ammonium chloride. Next, 3 mL of non-selective or selective media (M9 minimal media without ammonium chloride supplemented with 10 mM acetamide) was inoculated with the washed cells at an initial OD_{600} = 0.02 (~$6x10^6$ cells) in Hungate tubes. Cultures were grown at 37°C with shaking at 250x rpm for approximately 8 generations. Continuous exponential growth was ensured by harvesting cells after the first 4 generations and re-inoculating with 3 mL fresh media + antibiotic at OD_{600} = 0.02 prior to growth for the final 4 generations. Following 8 generations of growth the cells were stored and plasmid DNA extracted as in Klesmith et al.[24]. Unique selections started from the same unselected overnight culture were performed as replicates. To evaluate reproducibility between growth selections performed here and previous work performed on AmiE WT[25], a deep mutational scan of residues 171-255 in AmiE WT was performed in parallel to each growth selection performed in the present work.

Deep mutational scanning

AmiE variant DNA collected from the pre- and post-selection libraries was prepared for 300 BP paired end Illumina MiSeq sequencing as in Kowalsky et al.[23]. Primers used in the PCR reactions in preparation for Illumina sequencing are listed in (**Table A 10**). Sequencing of the variants was performed at the University of Illinois Chicago sequencing core. AmiE WT deep mutational scanning unprocessed sequencing results from Wrenbeck et al.[25] were downloaded from the SRA. Respective technical, and biological, replicates were processed independently. Regression analysis between the three starting points were performed by combining replicates within each variant. All data was processed using PACT[35] with the following changes from the default options entered into the configuration file: fast_filter_translate: qaverage = 20, and qlimit = 0; enrichment: ref_count_threshold = 5, sel_count_threshold = 0, strict_count_threshold = True. Normalized fitness metrics (ζ_i) were calculated by PACT as outlined in Kowalsky et al.[23]. To summarize, PACT calculates an enrichment ratio (ε_i) for mutations by assessing the pre- and post-selection counts of each mutant:

$$\varepsilon_i = \log_2\left(\frac{f_{fi}}{f_{oi}}\right) \quad (2)$$

Where f_{fi} is the frequency of mutant *i* in the post-selection population and f_{oi} is the frequency in the pre-selection population. The normalized fitness metric for each mutant i (ζ_i) was next calculated using the population-averaged number of doublings during selection (g_p) and the enrichment ratios of the mutant (ε_i) and the unmutated starting variant (ε_{ref}):

$$\zeta_i = \log_2\left(\frac{\left(\frac{\varepsilon_i}{g_p}\right)+1}{\left(\frac{\varepsilon_{ref}}{g_p}\right)+1}\right) \quad (3)$$

Lower bound fitness metrics - the cut-off below which fitness metrics cannot be discriminated from one another - were calculated as in Wrenbeck et al.[25] by using the median

read count for the pre-selection library and other statistics produced by PACT (**Figure A 8**). Lower fitness metrics were calculated to be: -1.38 for AmiE WT, -0.97 for AmiE I122L, and -0.86 for AmiE I38V.

AmiE I122L had seven outlier mutations removed because the difference in fitness metrics between replicates was greater than the 99.977% confidence intervals determined from sequencing depth of coverage (**Figure A 17**). No other mutants were removed from any of the datasets.

To account for global differences in fitness effects, we also generated normalized scatter plots for globally beneficial mutations. For each pairing of protein variants with AmiE WT (WT/I38V, WT/I122L), we performed a linear regression with the fitness of the first variant as the independent variable and the fitness of the second variant as the dependent variable. We then inverted the linear transformation obtained and applied it to the second fitness. Specifically, if the fitness of the second variant Y was modeled as $Y = mX + B$, then the normalized fitness Y_hat was computed as $(Y-B)/m$. Thus, after normalization, the least-squares regression associated with each pairwise plot has a slope of one and an intercept of zero.

Data availability

Raw sequencing reads have been deposited in the Sequencing Read Archive (SRA SAMN11258744 – SAMN11258771). The processed data sets are available in **Figures A 18 -20** and in **Tables A 11 – 20**. The AmiE I38V pAG and AmiE I122L pAG plasmids used in mutant library generation have been deposited in Addgene (Addgene ID: 129791 and 129792), while the AmiE WT base construct modified in Wrenbeck et al.[25] was previously deposited in Bienick et al.[55] (Addgene ID: 59837).

Acknowledgements

Thanks to Dr. J. Klesmith for his PACT troubleshooting help, E. Maurer and J. Hosten for their help with assorted tasks, and members of the Whitehead lab for providing feedback on ideas and figures. This work was supported by NSF CBET Career Award #1254238 to T.A.W.

REFERENCES

REFERENCES

1. Phillips PC: **Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems**. *Nat Rev Genetics* 2008, 9:855-867.

2. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS: **Robustness- epistasis link shapes the fitness landscape of a randomly drifting protein**. *Nature* 2006, 444:929–932.

3. Starr TN, Thornton JW: **Epistasis in protein evolution**. *Protein Science* 2016, 25:1204-1218.

4. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA: **Epistasis as the primary factor in molecular evolution**. *Nature* 2012, 490:535-538.

5. Baker D, Agard DA: **Kinetics versus thermodynamics in protein folding**. *Biochemistry* 1994, 33(24):7505-7509.

6. Shakhnovich EI: **Theoretical studies of protein-folding thermodynamics and kinetics**. *Curr Opin in Struc Biol* 1997, 7:29-40.

7. Goldenzweig A, Fleishman SJ: **Principles of protein stability and their application in computational design**. *Annu Rev Biochem* 2018, 87:105-129.

8. Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal Genomics, Davis EE, Sunyaev SR, Katsanis N: **Identification of cis-suppression of human disease mutations by comparative genomics**. *Nature* 2015, 524:225-229.

9. Tokuriki N, Stricher F, Serrano L, Tawfik DS: **How Protein Stability and New Functions Trade Off**. *PLoS Comp Biol* 2008, 4:e1000002.

10. Campbell E, Kaltenback M, Correy GJ, Carr PD, Porebski BT, Livingstone EK, Afriat-Jurnou L, Buckle AM, Weik M, Hollfelder F, Tokuriki N, Jackson CJ: **The role of protein dynamics in the evolution of new enzyme function**. *Nat Chem Biol* 2016, 12:944-950.

11. Kumar A, Natarajan C, Moriyama H, Witt CC, Weber RE, Fago A, Storz JF: **Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin**. *Mol Bio Evo* 2017, 34:1240-1251.

12. Serohijos AWR, Shakhnovich EI: **Contribution of Selection for Protein Folding Stability in Shaping the Patterns of Polymorphisms in Coding Regions**. *Mol Biol Evol* 2014, 31:165–176.

13. Tokuriki N, Jackson CJ, Afriat-Jurnou L, Wyganowski KT, Tang R, Tawfik DS: **Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme**. *Nat Comm* 2012, 3:1257.

14. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA: **Trade-offs between**

enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci U S A* 2017, 114:2265-2270.

15. Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability**. *Proc Natl Acad Sci U S A* 2006, 103:5869–5874.

16. Yu H, Dalby PA: **Exploiting correlated molecular-dynamics networks to counteract enzyme activity-stability trade-off**. *Proc Nat Aca Sci U S A* 2018, 115:E12192- E12200.

17. Ashenberg O, Gong LI, & Bloom JD: **Mutational effects on stability are largely conserved during protein evolution**. *Proc Nat Aca Sci U S A* 2013, 110:21071-21076.

18. Yu H, Dalby PA: **Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics**. *Proc Nat Aca Sci U S A* 2018, 115:E11043-11052.

19. Shah P, McCandlish DM, Plotkin JB: **Contingency and entrenchment in protein evolution under purifying selection**. *Proc Nat Aca Sci U S A* 2015, 112:E3226-E3235.

20. Dasmeh P, Serohijos AWR: **Estimating the contribution of folding stability to nonspecific epistasis in protein evolution**. *Proteins* 2018, 86:1242-1250.

21. Pollock DD, Thiltgen G, Goldstein RA: **Amino acid coevolution induces an evolutionary Stokes shift**. *Proc Nat Aca Sci U S A* 2012, 109:E1352-E1359.

22. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat Methods* 2014, 11:801–807.

23. Kowalsky CA, Klesmith JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA: **High-Resolution Sequence-Function Mapping of Full-Length Proteins**. *PLoS One* 2015, 10:e0118193.

24. Klesmith JR, Bacik JP, Michalczyk R, Whitehead TA: **Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli**. *ACS Synth Biol* 2015, 4:1235–1243.

25. Wrenbeck EE, Azouz LR, Whitehead TA: **Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded**. *Nat Comm* 2017, 8:15695.

26. Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J, Lieberman RL, Aharoni A, Silman I, Sussman JL, Tawfik DS, Fleishman SJ: **Automated Structure-and Sequence-Based Design of Proteins for High Bacterial Expression and Stability**. *Mol Cell* 2016, 63:337-346.

27. Raschke TM, Marqusee S: **The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions**. *Nature Struc Biol* 1997, 4:298-304.

28. Raschke TM, Kho J, Marqusee S: **Confirmation of the hierarchical folding of Rnase H: a protein engineering study**. *Nature Struc Biol* 1999, 6:825-831.

29. Karshikoff A, Nilsson L, Ladenstein R: **Rigidity versus flexibility: the dilemma of understanding protein thermal stability**. *FEBS* 2015, 282:3899-3917.

30. Robic S, Berger JM, Marqusee S: **Contributions of folding cores to the thermostabilities of two ribonucleases H**. *Protein Science* 2002, 11:381-389.

31. Scholl ZN, Yang W, Marszalek PE: **Direct observation of multimer stabilization in the mechanical unfolding pathway of a protein undergoing oligomerization**. *ACS nano* 2015, 9:1189-1197.

32. Cervoni L, Egistelli L, Mocan I, Giartosio A, Lascu I: **Quaternary structure of Dictyostelium discoideum nucleoside diphosphate kinase counteracts the tendency of monomers to form a molten globule**. *Biochemistry* 2003, 42:14599-14605.

33. Studier FW: **Protein production by auto-induction in high-density shaking cultures**. *Protein Exp Purif* 2005, 41:207–234.

34. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat Methods* 2016, 13:928-930.

35. Klesmith JR, Hackel BJ: **Improved mutation function prediction via PACT: Protein Analysis and Classifier Toolkit**. *Bioinformatics* 2018, doi/10.1093/bioinformatics/bty1042.

36. Orr HA: **The population genetics of beneficial mutations**. *Phil Trans R Soc B* 2010, 365:1195-1201.

37. Tokuriki N. Tawfik DS: **Stability effects of mutations and protein evolvability**. *Curr Opin Struct Biol* 2009, 19:596–604.

38. Gong LI, Suchard MA, Bloom JD: **Stability-mediate epistasis constrains the evolution of an influenze protein**. *eLife* 2013, 2:e00631.

39. Abriata LA, Palzkill T, Dal Peraro M: **How structural and physicochemical determinants shape sequence constraints in a functional enzyme**. *PloS One* 2015, 10:e0118684.

40. Jack BR, Meyer AG, Echave J, Wilke CO: **Functional sites induce long-range evolutionary constraints in enzymes**. *PLoS Biol* 2016,14:e1002452.

41. Mayorov A, Dal Peraro M, Abriata LA: **Active site-induced evolutionary constraints follow fold polarity principles in soluble globular enzymes**. *Mol Biol Evol* 2019, doi:10.1093/molbev/msz096.

42. Huang W, Palzkill T: **A natural polymorphism in beta-lacatamase is a global suppressor**. *Proc Nat Aca Sci U S A* 1997, 94:8801-8806.

43. Sideracki V, Huang W, Palzkill T, Gilbert HF: **A secondary drug resistance mutation of TEM-1 beta-lactamase that suppresses misfolding and aggregation**. *Proc Nat Aca Sci U S A* 2001, 98:283-288.

44. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH: **Thermodynamic prediction of protein neutrality**. *Proc Nat Aca Sci U S A* 2005,102:606-611.

45. Tokuriki N, Tawfik DS: **Chaperonin overexpression promotes genetic variation and enzyme evolution**. *Nature* 2009, 459:668–675.

46. Kristofich J, Morgenthaler AB, Kinney WR, Ebmeier, CC, Snyder DJ, Old WM, Cooper VS, Copley SD: **Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme**. *PLoS genetics* 2018, 14(8):e1007615.

47. Buhr F, Jha S, Thommen M, Mittlestaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA: **Synonymous codons direct co-translational folding towards different protein conformations**. *Mol Cell* 2016, 61:341-351.

48. Wylie CS, Shakhnovich EI: **A biophysical protein folding model accounts for most mutational fitness effects in viruses**. *Proc Nat Aca Sci U S A* 2011*,* 108:9916-9921.

49. Stiffler MA, Hekstra DR, Ranganathan R: **Evolvability as a function of purifying selection in TEM-1 β -lactamase**. *Cell* 2015, 160:882-892.

50. Weatheritt RJ, Babu MM: **Evolution. The hidden codes that shape protein evolution**. *Science* 2013, 342:1325-1326.

51. Firnberg E, Labonte JW, Gray JJ, Ostermeier M: **A comprehensive, high-resolution map of a gene's fitness landscape**. *Mol Biol Evol* 2014, 31:1581-1592.

52. Mittal R, Brindel J, Stephen JB, Kudla, G: **Codon usage influences fitness through RNA toxicity**. *Proc Nat Aca Sci U S A* 2018, 115(34):8639-8644.

53. Andrade J, Karmali A, Carrondo MA, Frazão C: **Structure of Amidase from Pseudomonas aeruginosa Showing a Trapped Acyl Transfer Reaction Intermediate State**. *J Biol Chem* 2007, 282:19598–19605.

54. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman KW, Renfrew PD, Smith CA, Sheffler W, et al.: **Rosetta3: An Object-Oriented Software Suite for Simulation and Design of Macromolecules**. *Methods in Enzymology* 2011, 487:545-574.

55. Bienick MS, Young KW, Klesmith JR, Detwiler EE, Tomek KJ, Whitehead TA: **The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate**. *PLoS One* 2014, 9:e109105.

56. Huynh K, Partch CL: **Analysis of protein stability and ligand interactions by thermal shift assay**. *Curr Pro in Protein Science* 2015, 79:28.9.1-28.9.14.

CHAPTER 4

Saturation mutagenesis genome engineering of infective ΦX174 bacteriophage via unamplified oligo pools and golden gate assembly

This work is adapted with permission from a publication that is in preparation by Matthew S. Faber, James T. Van Leuven, Martina M. Ederer, Holly H. Wichman Craig R. Miller, and Timothy A. Whitehead.

Abstract

Here we present the first steps in the preparation of a novel protocol for the construction of saturation single-site - and massive multi-site - mutant libraries of a bacteriophage. We segmented the ΦX174 genome into 15 non-toxic and non-replicative fragments compatible with golden gate assembly. We next used nicking mutagenesis with oligonucleotides prepared from unamplified oligo pools with individual segments as templates to prepare near-comprehensive single-site mutagenesis libraries of genes encoding the F capsid protein (421 amino acids scanned) and G spike protein (172 amino acids scanned). Libraries possessed greater than 99% of all 11,860 programmed mutations. Golden Gate cloning was then used to assemble the complete ΦX174 mutant genome libraries, and libraries transformed into *E. coli* C were infective. This protocol will expand reverse genetics experiments possible for studying viral evolution and, with some modifications, can be applied for engineering of therapeutically relevant bacteriophages with larger genomes.

Introduction

Predicting the tempo and trajectory of evolutionary change in the complex environments encountered by viruses and bacteria remains a challenge[1-3]. Understanding such changes are important for fundamental evolutionary studies as well as biotechnology applications like phage therapy for multi-drug resistant bacteria[4]. Two obstacles preventing better predictions are the oversimplification of the environment in experimental evolution studies compared to wild conditions and the vast number of possible mutational combinations that can occur, even in short adaptive walks and small genomes. Advances in DNA sequencing and synthesis have opened new ways to study microbial evolution and overcome some of these obstacles[5,6]. Unfortunately, suitable methods do not exist for generating comprehensive bacteriophage mutant libraries. This technology gap hinders those seeking to engineer phage for biotechnological applications and for those seeking a deeper understanding of how viruses evolve.

Methods for genetically engineering phages have recently been reviewed[7] and commonly involve homologous recombination and recombineering. While such approaches can be improved by clever incorporation of CRISPR-Cas systems[8-10], the overall modest efficiencies largely limits mutagenesis to the generation and single-site mutation of chimeric genomes[11,12], gene deletions[13], and limited multi-site mutagenesis[14]. By contrast, human viruses often have reverse genetics systems in place where comprehensive mutant libraries can be prepared for single genes[15] or regions within a gene[16,17] using replicative plasmids that encode whole viruses or viral components. Similarly, deep mutational scanning can be performed on plasmid-encoded phage proteins like the MS2 capsid protein[18].

Advances in the technologies for generating mutant libraries, synthesizing DNA, and for assembly of large DNA fragments[19] potentially allow for the facile construction and assembly of

user-defined mutagenesis of long nucleic acids. Nicking mutagenesis (NM) can be used to construct comprehensive single-site or other user-defined mutant libraries[20] using plasmid dsDNA as a template. In NM an oligo encodes the desired mutations by mismatch with the parental template. Recently, on-chip ink-jet printed oligo pools have been integrated into NM[21], so one can now construct heterogeneous libraries of oligos that contain tens to hundreds of thousands of high fidelity, unique sequences[21,22] with a low per base pair cost.

We wondered whether combining the advantages of replicative plasmids with advances in DNA synthesis and mutagenesis would allow us to create large user-defined libraries of bacteriophages. As a model bacteriophage system we selected ΦX174 because it has been extensively studied[23-26], has high resolution X-ray crystallography structures of the capsid and spike proteins[27,28], and has a very small genome[29] of 5386 nucleotides. These attractive features have led to proof-of-principle demonstrations from other groups for synthetic genome assembly from oligonucleotides[30] and complete refactoring of the phage[31]. This method combines Golden Gate assembly, nicking mutagenesis, and oligo pool technology to construct near comprehensive single-site and expansive multisite mutant libraries for the genes encoding the entire capsid F protein and spike G protein of the bacteriophage ΦX174. To our knowledge this is the first time near comprehensive single site mutant libraries of full capsid and spike proteins have been generated for an infective bacteriophage.

Results

Our strategy for generating large user-defined mutagenesis libraries of bacteriophage ΦX174 is shown in **Figure 4.1**. The circular 5386 nucleotide genome compactly encodes 11 genes where the F gene encodes a capsid and G encodes the spike protein (**Figure 4.1A**). The

genome was segmented onto 15 separate plasmids, including 3 plasmids for the F gene and 2 for the G gene (**Figure 4.1B**). User-defined comprehensive mutations on F and G were encoded using nicking mutagenesis with a single unamplified oligo pool containing all 11,860 mutagenic oligonucleotides (**Figure 4.1C**). ΦX174 mutant genomes are to be reconstituted from individual plasmids using Golden Gate cloning and transformed into a host *E. coli* strain (**Figure 4.1D**). Harvested phages will contain single or double non-synonymous mutations in the F and G genes depending on the mutagenized segments used for Golden Gate cloning (**Figure 4.1E**).

We first sought a reverse genetics system for ΦX174 wherein the virus chromosome was segmented and encoded on individual plasmids. Our method of construction closely followed that for the assembly of human coronavirus[32] NL63, where each plasmid contained unique BsmB1 type IIS restriction endonuclease sites flanking unique five nucleotide overlaps of wild-type (WT) ΦX174 sequence. This architecture allows for faithful assembly of the complete genome via Golden Gate cloning. However, as plasmids are replicated in *E. coli*, which ΦX174 naturally infects, the ΦX174 chromosome was encoded into15 separate nontoxic plasmids where genes were separated

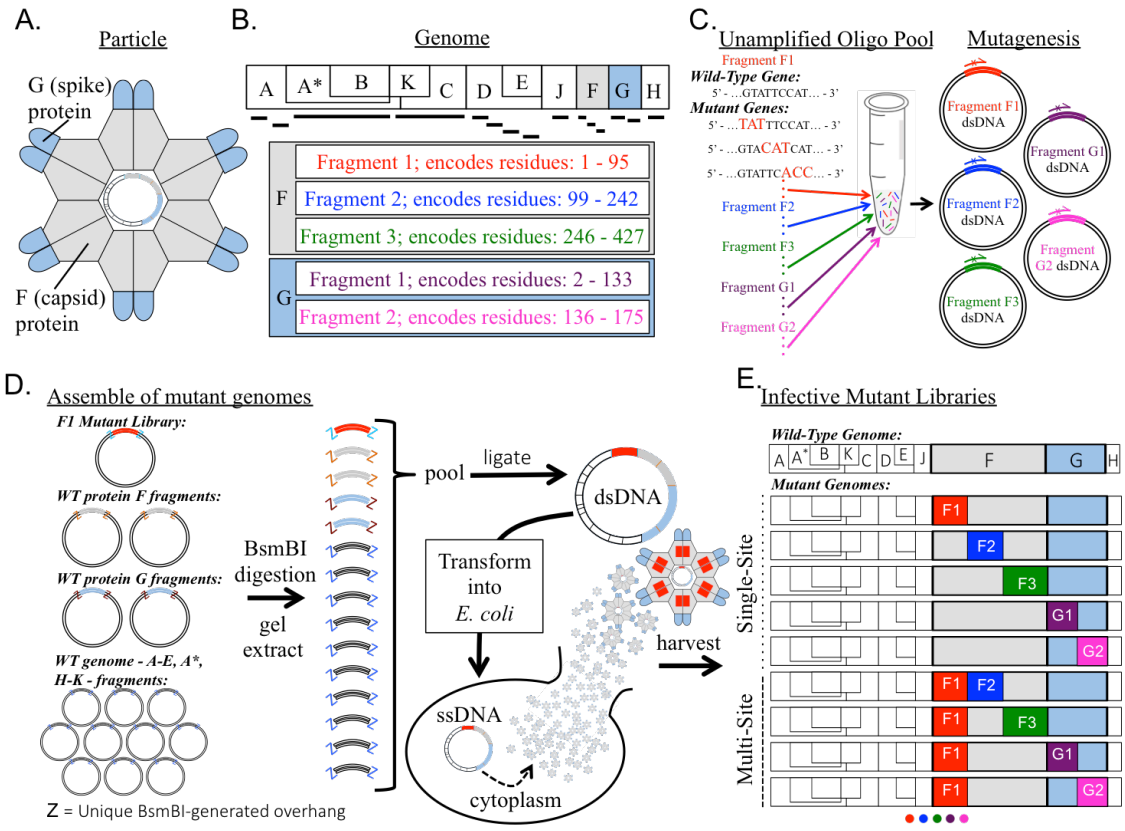

Figure 4.1: ΦX174 mutant library assembly. A. Anatomy of the viral external surface. B. A linear schematic of the circular ΦX174 ssDNA genome. Black lines delineate portions of the virus encoded on each of the 15 replicative plasmids. C. An unamplified oligo pool, containing all user-defined mutagenic oligonucleotides for F and G genes, was created and stored as a single mixture. This primer set was used in nicking mutagenesis for the generation of single-site saturation mutant libraries of genes encoding F and G proteins. D. Golden Gate cloning was used to assemble the mutant genomes. The dsDNA genome was transformed into *E. coli* and allowed a single burst phase. The resulting phage particles were collected and sequenced using SMRT sequencing. E. Linear schematics of the single site mutant libraries and some of the possible multi-site mutant libraries.

from their promoters and larger genes were segmented (full sequences for all plasmid inserts are given in **Note B 4.1**). ΦX174 phage could be reconstituted by digesting all plasmids with BsmB1, ligating inserts overnight, and transforming into electrocompetent *E. coli* C cells. Following incubation, phage plaques were tabulated. Sequencing of the genome showed that the recombinant phage encoded the intended sequence.

The F and G genes were targeted for saturation mutagenesis. We chose to introduce mutations by nicking mutagenesis, which requires the presence of a unique BbvCI nicking site on the dsDNA plasmid. The F gene plasmid F3 encoding residues 246-427 of the F gene product contained a unique BbvCI sequence, while the remaining four plasmids (F1, F2, G1, G2) required introduction of BbvCI nicking sites in the vector backbone. The presence in each plasmid of the unique BbvCI nicking site was verified with the successful generation of circular ssDNA from the modified plasmids by BbvCI.Nt and exonuclease digestion (**Figure B 1**).

In nicking mutagenesis, desired mutations are encoded through libraries of mutagenic oligonucleotides, which can be sourced from unamplified oligo pools[21]. We custom synthesized a single oligo pool containing 11,840 oligos encoding nearly every non-synonymous single mutation in the F and G genes. Application of nicking mutagenesis using this same oligo pool for different plasmids (F1, F2, F3, G1, G2) resulted in at least 14-fold excess transformants required for 99.9% theoretical coverage of the desired library (**Table B 1**). The diversity of the mutant libraries was validated using deep sequencing on an Illumina MiSeq platform and all libraries were found to have >99% coverage of all possible single mutations (full library statistics are given in **Table 4.1**).

Table 4.1: Mutant library NGS statistics. Summary table of the of the libraries prior to viral genome assembly.

Fragment	F1	F2		F3		G1		G2
Tile Number	1	1	2	1	2	1	2	1
Residues	1-95	99-157	158-242	246-336	337-427	2-53	55-133	136-175
Sequencing reads post quality filter	1163959	681935	1714460	800377	757177	747800	969275	813221
Fold oversampling of codon combinations	441.1	499.8	851	364.8	345.1	632.8	517.6	821.9
Percent of reads with:								
No nonsynonymous mutations	23.0%	63.3%	52.3%	54.0%	55.0%	68.3%	48.8%	25.2%
One nonsynonymous mutation	52.6%	27.5%	36.3%	33.1%	32.1%	24.1%	39.8%	59.7%
Multiple nonsynonymous mutation	24.4%	9.2%	11.3%	12.9%	12.8%	7.6%	11.4%	15.1%
Coverage of possible single nonsynonymous mutations	100%	100%	99.8%	100%	99.9%	99.6%	100%	100%

ϕX174 genomes were assembled by Golden Gate cloning with 1 plasmid encoding a mutant library combined with the remaining 14 WT plasmids. All five ϕX174 libraries prepared

contained active phage. Golden Gate reaction products were transformed in *E. coli* C cells with the number of plaques resulting from each reaction tabulated (**Table 4.2**). 50-60% of all assembled genomes contained at least one mutation as estimated by Sanger sequencing of at least 30 plaques. In contrast, for the control where all plasmids in the Golden Gate reaction were WT, 0% of the phage contained mutations in the encoded regions. The percent theoretical coverage ranged from 9-94% (**Table 4.2**). The pooled G1 and G2 libraries were the most successful but only reach 94% the required number of transformants for near complete coverage. The least efficient, F2, provided only 9% of the needed number of transformants.

Table 4.2: Mutant genome assembly statistics. A summary table of the number of possible mutant viruses, percent theoretical coverage, the number of viable viruses recovered from one transformation, and the number of mutant and wild type genomes recovered.

Fragment	F1	F2	F3	G1	G2	WT
Number of possible mutants	1,900	2,880	3,640	3,440		NA
Percent theoretical coverage	45	9	55	94		NA
Number of viable transformants	2,300	460	5,840	19,200		6,400
Number of plaques sequenced	34	25	30	60		30
Number of WT plaques	17	11	15	30		30

Discussion

Here we present the first steps in the development of a novel method for generating comprehensive single-site saturation – and massive multi-site – virulent mutant libraries of the spike and capsid proteins of the bacteriophage ΦX174. This method uses unamplified oligo pools, nicking scanning mutagenesis, and Golden Gate cloning. We are fortunate that our model virus has a small genome, under 6,000 nts, which we predict will enable us to assemble the 15

non-toxic fragments with sufficient transformants for near-complete coverage of the user-defined mutations. However, most biotech-relevant phages have larger genomes, and extension of this method to other viruses would encounter the following technical challenges. First, increasing genome sizes decrease transformational efficiencies, and larger genomes have increased susceptibility to DNA shearing[33]. Both size-dependent effects can disrupt the integrity and coverage of the phage libraries[34]. Second, our method includes restrictions on the DNA sequences that can be used. Golden Gate cloning requires a genome without unique Type IIS restriction sequences, while nicking mutagenesis requires that there be only one orientation of the BbvCI nicking site within the template DNA fragment. Third, ligating 15 fragments is close to the upper limit of Golden Gate cloning. Additionally, increasing the size of individual fragments is difficult as the presented method depends on replicating plasmids in *E. coli*, and larger fragments are more likely to be toxic *in vivo*.

Based on the above considerations, adapting this method to larger viruses will require modifications for the genome assembly steps. There are a variety of methods available for attempting to improve genome library assembly and amplification. For improving genome assembly, we speculate that a combination of hierarchical assembly[35] with other yeast based assembly methods[36] will allow for large viral genomes to be efficiently assembled. Finally, we expect the Tx.Tl cell free expression system[37] will be able to produce the viral libraries with less bias than *in vivo* amplification in *E. coli*. In summary, we have presented a complete method for efficient deep mutational scanning of the bacteriophage ΦX174. We anticipate this method will find utility in fundamental molecular evolution studies as well as translate to potential medicinal applications.

Materials and methods

Reagents

All purchased enzymes and DNA purification kits were from New England Biolabs, antibiotics were purchased from GoldBio, other chemicals were purchased from Sigma-Aldrich. Individual primers were purchased from Integrated DNA Technologies.

Segmentation of the ΦX174 genome

A phage assembly platform for ΦX174 was devised following Donaldson et al.[32]. The ΦX174 chromosome was divided into 15 genomic fragments designed to avoid host cell toxicity by separating genes from their promoters and breaking large genes into multiple segments (**Note B 4.1**). Each segment is flanked by unique five nucleotide overlaps of WT ΦX174 sequence so that they can be amplified from the ancestral ΦX174 using PCR primers designed to incorporate terminal BsmB1 restriction sites. Amplicons were cloned into pCR2.1 using the Invitrogen TOPO TA cloning system (Life Technologies, Grand Island, NY).

Introduction of nicking site

The gene fragments – F1, F2, G1, G2 – in the pCR2.1-TOPO plasmid (**Notes B 4.1 and 4.2**) had the BbvCI nicking site introduced via overhang PCR, type I restriction enzyme cutting, and ligation. First, PCR was performed with overhang primers (**Table B 2**) to introduce the BbvCI site. Standard Phusion Polymerase HF reaction conditions were used with 4 ng of template DNA, cycling is follows: 98°C - 1 min, 25x cycles of: 98°C – 10 seconds, 67°C 15 seconds, 72°C 2.5 minutes, followed by 72°C for 10 minutes. PCR products were run on a 1% agarose gel stained with SYBR™ safe stain (Invitrogen) and the DNA bands at ~4600 bp were

extracted using a Monarch® DNA Gel Extraction Kit. Next, 1 µg of the PCR product was digested with KpnI (20 U) in NEB Buffer 1.1 at 37°C for two hours. Digested DNA was then clean and concentrated with a Monarch® PCR and DNA Clean and Concentrate Kit and eluted into 20 µL nuclease free H_2O. One microliter of the purified and digested DNA was ligated using T4 DNA ligase at ~25°C for 1 hour in standard conditions in a 20 µL reaction. Five microliters of the ligation reaction were transformed into chemically competent XL1-Blue *E. coli* via standard protocols. Cells were plated on LB agar containing 100 µg/mL carbenicillin and 50 µg/mL kanamycin and grown at 37°C for ~16 hours. Cells were picked from transformation plates and grown in 50 mL TB with 100 µg/mL carbenicillin and 50 µg/mL kanamycin for ~12 hours, cells were pelleted, and DNA purified using compact midi-preps.

Comprehensive single site mutant library construction

Comprehensive mutant libraries were generated using nicking mutagenesis (NM) as in Wrenbeck et al.[20] with modifications for using oligo pool mutagenic primers as noted in Medina et al.[21]. A single oligo pool encoding for all possible single missense and nonsense substitutions in F and G was designed using the custom python scripts from Medina et al.[21] and custom synthesized by Agilent (full sequences of all oligos are given in the associated file "**VMA_supplemental_data.csv**" in the published text). Oligo pools were designed with 20-24 bases of gene overlap flanking the mutated codon. Codons were chosen based on *E. coli* codon usage frequency. This oligo pool was used directly in NM without further amplification and using 2 mg of the relevant golden gate plasmid as a template. Libraries were transformed into high efficiency electrocompetent XL1-Blue *E. coli* (Agilent cat #: 200228) using 1 mm electroporation cuvettes at 1200V. Cells were plated on large bioassay plates (245mm x 245mm

x 25mm, Sigma-Aldrich) containing LB agar + 100 μg/mL carbenicillin and 50 μg/mL kanamycin and grown at 37°C for ~16 hours. Plates were scraped in 10 mL plain LB, broken into ~1.2 mL aliquots, pelleted, and stored -80°C. DNA was purified from 1x aliquot using a Monarch® Plasmid Mini-Prep Kit.

Illumina sequencing prep and analysis

Purified library DNA was prepared for deep sequencing as in Kowalsky et al.[38] using the primers listed in **Table B 3** and gene tiling as specified in **Table 4.1**. DNA was Illumina sequenced on a Mi-Seq platform with 250 BP paired end reads. The University of Colorado BioFrontiers Sequencing Core performed the Illumina sequencing. Data was processed using PACT[39] to determine the library coverage with the following changes to the default options in the configuration file: fast_filter_translate: qaverage = 20, qlimit = 0; enrichment: ref_count_threshold = 5, sel_count_threshold = 0, strict_count_threshold = True. The heatmaps of counts can be found in **Figure B 2 - 9**.

Assembly of mutant genomes

We pooled plasmid DNA containing all 15 of the phage DNA fragments in equimolar amounts and digested them with BsmB1 (Fermentas Fast Digest, Life Technologies, Grand Island, NY) for 30 minutes to 1 hour at 37°C. The digested plasmids were subjected to agarose gel electrophoresis for 10 to 15 minutes using a 1.2% agarose gel to separate the vector from the inserts. The inserts were excised from the gel, purified using the GeneJET gel extraction kit (Fermentas), ligated overnight at 14°C with T4 DNA ligase (Promega Corporation, Madison, WI), and transformed by electroporation into 100 μl competent *E. coli* C cells. The

transformation mix was resuspended with 1 ml of ΦLB and either plated immediately or incubated for about 20 minutes at 37°C to allow for one viral burst. The ΦLB was added to 3 ml of ΦLB top agar and plated onto a ΦLB agar plate. After four to five hours of incubation at 37°C, recombinant phage plaques were visible and plates were removed from the incubator. To verify that the recombinant phage encoded the intended sequence, we picked about 30 plaques for each intended mutational target and Sanger sequenced the entire targeted gene. We also sequenced the F and G genes for about 30 wild type plaques to assure that no mutations were naturally accumulating. Briefly, individual plaques were picked with sterile toothpicks and placed in 200 uL ΦLB and gently swirled. 1 uL of this mix was used to PCR amplify approximately ½ of the ΦX174 genome using ΦX-0F (5'-GAGTTTTATCGCTTCCATG-3') and ΦX-2953R (5'-CCGCCAGCAATAGCACC-3') primers. Internal sequencing primers ΦX-979F (5'-CGGCCCCTTACTTGAGG-3') and ΦX-1500R (5'-TTGAGATGGCAGCAACGG-3') were used to sequence gene F. ΦX-2953R was used to sequence gene G. PCR cleanups and sequencing was done at Eurofins Genomics. PCR reaction conditions were; 5 uL10X Taq buffer, 2.5 uL 10 uM ΦX-0F primer, 10 uM ΦX-2953F primer, 0.8 uL 12.5 uM dNTPs, 0.5 uL Taq polymerase (NEB #M0273), 1 uL template, 37.7 uL H_20. Thermocycling conditions were 1 cycle at 95°C for 2 min, 30 cycles at 95°C for 15 sec, 52°C for 30 sec, 68°C for 2 min, 1 cycle at 68°C for 5min.

REFERENCES

REFERENCES

1. Holmes EC: **What can we predict about viral evolution and emergence?**. *Curr Opin in Vir* 2013, 3:180-184.

2. Neher RA: **Genetic draft, selective interference, and population genetics of rapid adaptation**. *Annu Rev Ecol Evol Syst* 2013, 44:195-215.

3. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R: **Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures**. *Nature Gen* 2014, 46:82-87.

4. Dedrick RM, Guerrero-Bustamante CA, Garlena RA, Russell DA, Ford K, Harris K, Gilmour KC, Soothill J, Jacobs-Sera D, Schooley RT, Hatfull GF, Spencer H: **Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant *Mycobacterium abscessus***. *Nature Medicine.* 2019, 25:730-733.

5. Barrick JE, Lenski RE: **Genome dynamics during experimental evolution**. *Nat Rev Genetics* 2013, 14:827-839.

6. Fowler DM, Fields S: **Deep mutational scanning: a new type of protein science**. *Nat Methods* 2014, 11:801-807.

7. Pires DP, Cleto S, Sillankorva S, Azeredo J, Lu TK: **Genetically engineered phages: a review of advances over the last decade**. *Microbiol and Mol Biol Rev* 2016, 80:523-543.

8. Kiro R, Shitrit D, Qimron U: **Efficient engineering of a bacteriophage genome using the type I-E CRISPR-Cas system**. *RNA Biol* 2014, 11:42-44.

9. Lemay ML, Tremblay DM, Moineau S: **Genome engineering of virulent lactococcal phages using CRIPSR-Cas9**. *ACS Synth Biol* 2017, 6:1351-1358.

10. Schilling T, Dietrich S, Hoppert M, Hertel R: **A CRISPR-Cas9-based toolkit for fast and precise in vivo genetic engineering of Bacillus subtilis phages**. *Viruses* 2018, 10:241.

11. Doore SM, Fane BA: **The kinetic and thermodynamic aftermath of horizontal gene transfer governs evolutionary recovery**. *Mol. Biol. Evo.* 2015, 32:2571-2584.

12. Doore SM, Schweers NJ, Fane BA: **Elevating fitness after a horizontal gene exchange in bacteriophage ΦX174**. *Virology* 2017, 501:25-34.

13. Uchiyama A, Heiman P, Fane BA: **N-terminal deletions of the ΦX174 external scaffolding protein affect the timing and fidelity of assembly**. *Virology* 2009, 386:303-309.

14. Stockdale SR, Collins B, Silvia S, Douillard FP, Mahony J, Cambillau C, van Sinderen D: **Structure and assembly of TP901-1 Virion unveiled by mutagenesis**. *PLOS ONE* 2015,

10:e0131676.

15. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, Bloom JD: **Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants**. *Proc. Natl. Aca. Sci. USA* 2018, 115:E8276-E8285.

16. Doub MB, Lee JM, Bloom JD: **How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin**. *Nature Comm.* 2018, 9:1386.

17. Wu NC, Qi H: **Application of deep mutational scanning in hepatitis C virus**. *Met. in Mol. Biol.* 2019, 1911:183-190.

18. Hartman EC, Jakobson CM, Favor AH, Lobba MJ, Álvarez-Benedicto E, Francis MB, Tullman-Ercek D: **Quantitative characterization of all single amino acid variants of a viral capsid based drug delivery vehicle**. *Nature Comm.* 2018, 9:1385.

19. van Dolleweerd CJ, Kessans SA, Van de Bittner KC, Bustamante LY, Bundela R, Scott B, Nicholson MJ, Parker EJ: **MIDAS: a modular DNA assembly system for synthetic biology**. *ACS Synth Biol* 2018, 7:1018-1029.

20. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat. Methods* 2016, 13:928-930.

21. Medina-Cucurella A, Steiner PJ, Faber MS, Beltrán J, Borelli A, Kirby MB, Cutler SR, Whitehead TA: **User-defined single pot mutagenesis using unpurified oligo pools**. *Prot. Eng. Des. Sel.* 2019, TBD:TBD.

22. Kosuri S, Church GM: **Large-scale *de novo* DNA synthesis: technologies and applications**. *Nat. Methods* 2014, 11:499-507.

23. Aoyama A, Hamatake RK, Hayashi M: **Morphogenesis of ΦX174: *in vitro* synthesis of infectious phage from purified viral components**. *Proc. Natl. Aca. Sci. USA* 1981, 12:7285-7289.

24. Hafenstein S, Fane BA: **ΦX174 genome-capsid interactions influence the biophysical properties of the virion: evidence for a scaffolding-like function for the genome during the final stages of morphogenesis**. *J. of Viro.* 2002, 76:5350-5356.

25. Bernal RA, Hafenstein S, Esmeralda R, Fane BA, Rossmann MG: **The ΦX174 protein J mediates DNA packaging and viral attachment to host cells**. *J. Mol. Biol.* 2004, 337:1109-1122.

26. Rokyta DR, Burch CL, Caudle SB, Wichman HA: **Horizontal gene transfer and the evolution of microvirid coliphage genomes**. *J. of Bacter.* 2006, 188:1134-1142.

27. McKenna R, Xia D, Willingmann R, Ilag LL, Krishnaswamy S, Rossmann MG, Olson NH, Baker TS, Incardona NL: **Atomic structure of single-stranded DNA bacteriophage ΦX174 and its functional implications**. *Nature* 1992, 355:137-143.

28. Dokland T, McKenna R, Ilag LL, Bowman BR, Incardona NL, Fane BA, Rossmann MG: **Structure of a viral procapsid with molecular scaffolding**. *Nature* 1997, 389:308-313.

29. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchinson CA 3rd, Slocombe PM, Smith M: **The nucleotide sequence of bacteriophage phiX174**. *J Mol Biol* 1978, 125:225-246.

30. Smith HO, Hutchinson III CA, Pfannkoch C, Venter JC: **Generating a synthetic genome by whole genome assembly: ΦX174 bacteriophage from synthetic oligonucleotides**. *Proc. Natl. Aca. Sci. USA* 2003, 100:15440-15445.

31. Jaschke PR, Lieberman EK, Rodriguez J, Sierra A, Endy D: **A fully decompressed synthetic bacteriophage ΦX174 genome assembled and archived in yeast**. *Virology* 2012, 434:278-284.

32. Donaldson EF, Yount B, Sims AC, Burkett S, Pickles RJ, Baric RS: **Systematic Assembly of a Full-Length Infectious Clone of Human Coronavirus NL63**. *J. Virol.* 2008, 82:11948–11957.

33. Rogers SO, Bendich AJ: **Extraction of DNA from plant tissues**. *Plant Mole Biol Manual* 1988, A6:1-10.

34. Edwards RA, Rohwer F: **Viral metagenomics**. *Nature Reviews Micro* 2005, 3:504-509.

35. Lee, ME, DeLoache WC, Cervantes B, Dueber JE: **A highly characterized yeast toolkit for modular, multipart assembly**. *ACS Synth Biol* 2015, 4:975-986.

36. Ando H, Lemire S, Pires DP, Lu TK: **Engineering modular viral scaffolds for targeted bacterial population editing**. *Cell Systems* 2015, 1:187-196.

37. Garamella J, Marshall R, Rustad M, Noireaux V: **The all *E. coli* Toolbox 2.0: a platform for cell-free synthetic biology**. *ACS Synth Biol* 2016, 5:344-355.

38. Kowalsky CA, Klesmith, JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA: **High-Resolution Sequence-Function Mapping of Full-Length Proteins**. *PLoS One* 2015, 10:e0118193.

39. Klesmith JR, Hackel BJ: **Improved mutation function prediction via PACT: Protein Analysis and Classifier Toolkit**. *Bioinformatics* 2018, doi/10.1093/bioinformatics/bty1042.

CHAPTER 5

Conclusions and perspectives

This thesis focuses on the generation and analysis of large mutational datasets for understanding molecular evolution. Deep mutational scanning (DMS) is a technique that combines saturation mutagenesis libraries with high throughput selection and deep sequencing to assess the impact of nearly all single-point mutations on protein function[1]. Here, DMS is used to study how folding probability constrains the molecular evolution of enzymes. This thesis also expands the research questions DMS experiments can address through a novel method for generating mutant genome libraries of virulent phages. This chapter provides a summary of the work performed and a discussion of the broader impacts of this PhD thesis.

DMS experiments produce large datasets that quantitatively describe the impacts of single mutations[2-5]. In **Chapter 2**, we reviewed the use of DMS datasets - as well as datasets from several other experimental and evolutionary sources – in the forward engineering of protein therapeutics. We describe the use of these datasets in the identification of mutations that can tune the specificity, affinity, and solubility of a target protein. As more comprehensive studies of mutational impacts are performed, we expect the integration of the resultant data to provide better predictions of which mutations will, and will-not, work toward a specific goal. Eventually, we expect guidelines like Lipinski's rule of five for small molecule drugs[6] to emerge for protein therapeutics. Indeed, work aiming to establish these guidelines are already underway[7].

In **Chapter 3**, I demonstrated how DMS can be used to understand how a single biophysical parameter of an enzyme constrains its molecular evolution. In this study I performed near-comprehensive single-mutation analysis on two unique single-point mutants of the aliphatic amidase AmiE from *P. aeruginosa*. The point-mutants have decreased *in vivo* folding probabilities and statistically indistinguishable catalytic activities compared to the starting enzyme. In a previous study[5] the starting enzyme had undergone DMS with the same selection

conditions applied in this study, allowing us to compare the obtained datasets. Because the only biophysical parameter modulated is the *in vivo* folding probability, we are able to assess how only this parameter alters the types of mutations that can arise. Additionally, this study provides a near comprehensive analysis of how single-point mutations combine, providing a detailed view of epistasis, the non-additive combination of mutational effects[8]. The novelty of this study is found in the aforementioned details and in the fact that this study is the most comprehensive analysis of how a single initial biophysical parameter impacts mutational outcomes to date.

Comparing the datasets obtained for our two disrupted enzymes with that of the initial starting enzyme provides intriguing insights into molecular evolution. First, in all genetic backgrounds the distributions of beneficial fitness effects are described by the General Pareto distribution. Thus, the likelihood of uncovering a beneficial mutation is insensitive to initial *in vivo* folding probability in our experiments. Interestingly, most beneficial mutations are shared, and only 19 mutations are likely to require an increased *in vivo* folding probability to buffer stability penalties.

Epistasis was found in both the beneficial and deleterious populations of mutations following decreases in the *in vivo* folding probabilities. Interestingly, in the population of double mutants that are deleterious individually and beneficial when combined, phenotype-specific interactions dominated those that are phenotype-independent. This suggests that mutations rescuing *in vivo* folding probability are more likely to be in direct contact with the disrupting mutation, and less likely to alter the global properties of the enzyme. Additionally, beneficial mutations with synonymous codon fitness disparities - opposite mutational outcomes for synonymous codons - were correlated with background specific beneficial mutations. This

correlation reveals the importance of the mRNA sequence selected, and how it significantly restricts mutational outcomes for identical peptides.

Taken together, these results indicate that models of protein stability and evolution must include all aspects of the protein life cycle, from transcription to degradation, to fully understand the impact of a single mutation. It is uncertain if these findings are unique to this experimental system, or if they are descriptive of protein evolution in general. Additional studies similar to this one, using other proteins and with different biophysical parameters altered, will improve our ability to predict mutational outcomes.

The final goal of this thesis was methodological. In **Chapter 4** we expanded the experimental range of DMS by generating comprehensive single-site saturation, and massive multi-site, mutant genome libraries of the bacteriophage ϕX174. The mutant genomes were generated by performing nicking scanning mutagenesis[9] on non-toxic truncations of the genome and assembling complete viral genomes with Golden Gate cloning. These mutant genomes will allow researchers to probe how host specificity is encoded into the primary sequence, how intra- and inter-molecular epistasis is involved in viral evolution and micro-compartment formation, and expand our understanding of the molecular evolution of viruses.

In conclusion, this thesis provides significant insights into molecular evolution and how to study it. All aspects of this thesis involve the generation and/or analysis of large mutational datasets with the goal of understanding how proteins change over evolutionary time. In the coming decades the mutational data obtained from more DMS experiments, and from other experimental and archival sources, will provide valuable insights into molecular evolution. I expect this information will be combined and integrated into a new generation of predictive algorithms, which will significantly improve the protein engineering process.

REFERENCES

REFERENCES

1. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science.** *Nat Methods* 2014, 11:801–807.

2. Kowalsky CA, Klesmith JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA: **High-Resolution Sequence-Function Mapping of Full-Length Proteins**. *PLoS One* 2015, 10:e0118193.

3. Klesmith JR, Bacik JP, Michalczyk R, Whitehead TA: **Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli.** *ACS Synth Biol* 2015, 4:1235–1243.

4. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA: **Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning.** *Proc Natl Acad Sci U S A* 2017, 114:2265-2270.

5. Wrenbeck EE, Azouz LR, Whitehead TA: **Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded**. *Nat Comm* 2017, 8:15695.

6. Lipinski CA: **Lead-and drug-like compounds: the rule-of-five revolution**. *Drug Discov Today Technol* 2004, 1:337-341.

7. Rabia LA, Desai AA, Jhajj HS, Tessier PM: **Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility**. *Biochem Eng J* 2018, 137:365-374.

8. Starr TN, Thornton JW: **Epistasis in protein evolution**. *Protein Science* 2016, 25:1204-1218.

9. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA: **Plasmid-based one-pot saturation mutagenesis**. *Nat Methods* 2016, 13:928-930.

APPENDICES

APPENDIX A

Chapter 3 supporting information

Note A 1: Amino acid sequences of the AmiE variants. Mutations are highlighted in red.

>AmiE_WT

MRHGDISSSNDTVGVAVVNYKMPRLHTAAEVLDNARKIAEMIVGMKQGLPGMDLVVF
PEYSLQGIMYDPAEMMETAVAIPGEETEIFSRACRKANVWGVFSLTGERHEEHPRKAPY
NTLVLIDNNGEIVQKYRKIIPWCPIEGWYPGGQTYVSEGPKGMKISLIICDDGNYPEIWRD
CAMKGAELIVRCQGYMYPAKDQQVMMAKAMAWANNCYVAVANAAGFDGVYSYFG
HSAIIGFDGRTLGECGEEEMGIQYAQLSLSQIRDARANDQSQNHLFKILHRGYSGLQASG
DGDRGLAECPFEFYRTWVTDAEKARENVERLTRSTTGVAQCPVGRLPYEGLEHHHHHH

>AmiE_I122L

MRHGDISSSNDTVGVAVVNYKMPRLHTAAEVLDNARKIAEMIVGMKQGLPGMDLVVF
PEYSLQGIMYDPAEMMETAVAIPGEETEIFSRACRKANVWGVFSLTGERHEEHPRKAPY
NTLVLLDNNGEIVQKYRKIIPWCPIEGWYPGGQTYVSEGPKGMKISLIICDDGNYPEIWR
DCAMKGAELIVRCQGYMYPAKDQQVMMAKAMAWANNCYVAVANAAGFDGVYSYF
GHSAIIGFDGRTLGECGEEEMGIQYAQLSLSQIRDARANDQSQNHLFKILHRGYSGLQAS
GDGDRGLAECPFEFYRTWVTDAEKARENVERLTRSTTGVAQCPVGRLPYEGLEHHHHH
H

>AmiE_I38V

MRHGDISSSNDTVGVAVVNYKMPRLHTAAEVLDNARKVAEMIVGMKQGLPGMDLVV
FPEYSLQGIMYDPAEMMETAVAIPGEETEIFSRACRKANVWGVFSLTGERHEEHPRKAP
YNTLVLIDNNGEIVQKYRKIIPWCPIEGWYPGGQTYVSEGPKGMKISLIICDDGNYPEIWR
DCAMKGAELIVRCQGYMYPAKDQQVMMAKAMAWANNCYVAVANAAGFDGVYSYF
GHSAIIGFDGRTLGECGEEEMGIQYAQLSLSQIRDARANDQSQNHLFKILHRGYSGLQAS
GDGDRGLAECPFEFYRTWVTDAEKARENVERLTRSTTGVAQCPVGRLPYEGLEHHHHH
H

Note A 2: DNA sequences of the AmiE variants. Codons mutated are underlined and bold, while mutations are highlighted in red.

>AmiE_WT

ATGAGACATGGCGATATTAGCTCGTCAAATGATACCGTAGGCGTAGCCGTGGTGAA
TTACAAGATGCCGCGTTTACATACTGCTGCTGAAGTCCTGGATAATGCCCGCAAAAT
TGCGGAAATGATCGTTGGTATGAAGCAAGGTCTGCCGGGCATGGATCTGGTTGTGTT
TCCTGAATATTCTTTACAGGGTATTATGTACGACCCTGCTGAAATGATGGAAACAGC
CGTGGCGATTCCAGGCGAAGAAACGGAAATCTTTAGCCGTGCTTGTAGAAAAGCAA
ATGTTTGGGGTGTGTTCTCCCTGACCGGCGAACGTCATGAAGAACACCCTAGAAAGG
CACCATACAACACTCTGGTCTTGATCGATAACAACGGTGAAATCGTACAAAAGTAC
AGAAAGATCATCCCATGGTGTCCGATTGAAGGCTGGTATCCAGGTGGCCAGACATA
CGTCTCTGAAGGTCCGAAAGGCATGAAGATCTCATTAATTATCTGCGATGACGGTAA
TTATCCGGAAATTTGGAGAGATTGTGCCATGAAGGGTGCGGAATTGATCGTTCGCTG
CCAAGGCTATATGTACCCTGCTAAAGACCAACAAGTTATGATGGCTAAGGCAATGG
CCTGGGCGAATAACTGTTATGTCGCTGTAGCAAACGCTGCAGGTTTTGATGGCGTTT
ATAGCTACTTCGGTCATAGTGCCATTATCGGTTTTGACGGCCGTACTCTGGGTGAAT
GCGGCGAAGAAGAAATGGGCATTCAATACGCGCAGTTGTCTCTGTCACAAATCCGC
GATGCCCGTGCGAATGACCAAAGTCAGAACCATTTGTTTAAAATCTTGCACAGAGGT
TACTCCGGTTTGCAGGCTTCGGGCGATGGCGACCGTGGTCTGGCAGAATGCCCATTT
GAATTCTACCGTACCTGGGTTACTGATGCTGAAAAGGCAAGAGAAAACGTGGAACG
CCTGACTCGCTCCACAACAGGTGTCGCCCAATGCCCAGTAGGTCGTCTGCCGTATGA
AGGCCTCGAGCACCACCACCACCACCAC

>AmiE_I122L

ATGAGACATGGCGATATTAGCTCGTCAAATGATACCGTAGGCGTAGCCGTGGTGAA
TTACAAGATGCCGCGTTTACATACTGCTGCTGAAGTCCTGGATAATGCCCGCAAAAT
TGCGGAAATGATCGTTGGTATGAAGCAAGGTCTGCCGGGCATGGATCTGGTTGTGTT
TCCTGAATATTCTTTACAGGGTATTATGTACGACCCTGCTGAAATGATGGAAACAGC
CGTGGCGATTCCAGGCGAAGAAACGGAAATCTTTAGCCGTGCTTGTAGAAAAGCAA
ATGTTTGGGGTGTGTTCTCCCTGACCGGCGAACGTCATGAAGAACACCCTAGAAAGG
CACCATACAACACTCTGGTCTTG**CTC**GATAACAACGGTGAAATCGTACAAAAGTAC
AGAAAGATCATCCCATGGTGTCCGATTGAAGGCTGGTATCCAGGTGGCCAGACATA
CGTCTCTGAAGGTCCGAAAGGCATGAAGATCTCATTAATTATCTGCGATGACGGTAA
TTATCCGGAAATTTGGAGAGATTGTGCCATGAAGGGTGCGGAATTGATCGTTCGCTG
CCAAGGCTATATGTACCCTGCTAAAGACCAACAAGTTATGATGGCTAAGGCAATGG
CCTGGGCGAATAACTGTTATGTCGCTGTAGCAAACGCTGCAGGTTTTGATGGCGTTT
ATAGCTACTTCGGTCATAGTGCCATTATCGGTTTTGACGGCCGTACTCTGGGTGAAT
GCGGCGAAGAAGAAATGGGCATTCAATACGCGCAGTTGTCTCTGTCACAAATCCGC
GATGCCCGTGCGAATGACCAAAGTCAGAACCATTTGTTTAAAATCTTGCACAGAGGT
TACTCCGGTTTGCAGGCTTCGGGCGATGGCGACCGTGGTCTGGCAGAATGCCCATTT
GAATTCTACCGTACCTGGGTTACTGATGCTGAAAAGGCAAGAGAAAACGTGGAACG

CCTGACTCGCTCCACAACAGGTGTCGCCCAATGCCCAGTAGGTCGTCTGCCGTATGA
AGGCCTCGAGCACCACCACCACCACCAC

>AmiE_I38V

ATGAGACATGGCGATATTAGCTCGTCAAATGATACCGTAGGCGTAGCCGTGGTGAA
TTACAAGATGCCGCGTTTACATACTGCTGCTGAAGTCCTGGATAATGCCCGCAAA**GT**
TGCGGAAATGATCGTTGGTATGAAGCAAGGTCTGCCGGGCATGGATCTGGTTGTGTT
TCCTGAATATTCTTTACAGGGTATTATGTACGACCCTGCTGAAATGATGGAAACAGC
CGTGGCGATTCCAGGCGAAGAAACGGAAATCTTTAGCCGTGCTTGTAGAAAAGCAA
ATGTTTGGGGTGTGTTCTCCCTGACCGGCGAACGTCATGAAGAACACCCTAGAAAGG
CACCATACAACACTCTGGTCTTGATCGATAACAACGGTGAAATCGTACAAAAGTAC
AGAAAGATCATCCCATGGTGTCCGATTGAAGGCTGGTATCCAGGTGGCCAGACATA
CGTCTCTGAAGGTCCGAAAGGCATGAAGATCTCATTAATTATCTGCGATGACGGTAA
TTATCCGGAAATTTGGAGAGATTGTGCCATGAAGGGTGCGGAATTGATCGTTCGCTG
CCAAGGCTATATGTACCCTGCTAAAGACCAACAAGTTATGATGGCTAAGGCAATGG
CCTGGGCGAATAACTGTTATGTCGCTGTAGCAAACGCTGCAGGTTTTGATGGCGTTT
ATAGCTACTTCGGTCATAGTGCCATTATCGGTTTTGACGGCCGTACTCTGGGTGAAT
GCGGCGAAGAAGAAATGGGCATTCAATACGCGCAGTTGTCTCTGTCACAAATCCGC
GATGCCCGTGCGAATGACCAAAGTCAGAACCATTTGTTTAAAATCTTGCACAGAGGT
TACTCCGGTTTGCAGGCTTCGGGCGATGGCGACCGTGGTCTGGCAGAATGCCCATTT
GAATTCTACCGTACCTGGGTTACTGATGCTGAAAAGGCAAGAGAAAACGTGGAACG
CCTGACTCGCTCCACAACAGGTGTCGCCCAATGCCCAGTAGGTCGTCTGCCGTATGA
AGGCCTCGAGCACCACCACCACCACCAC

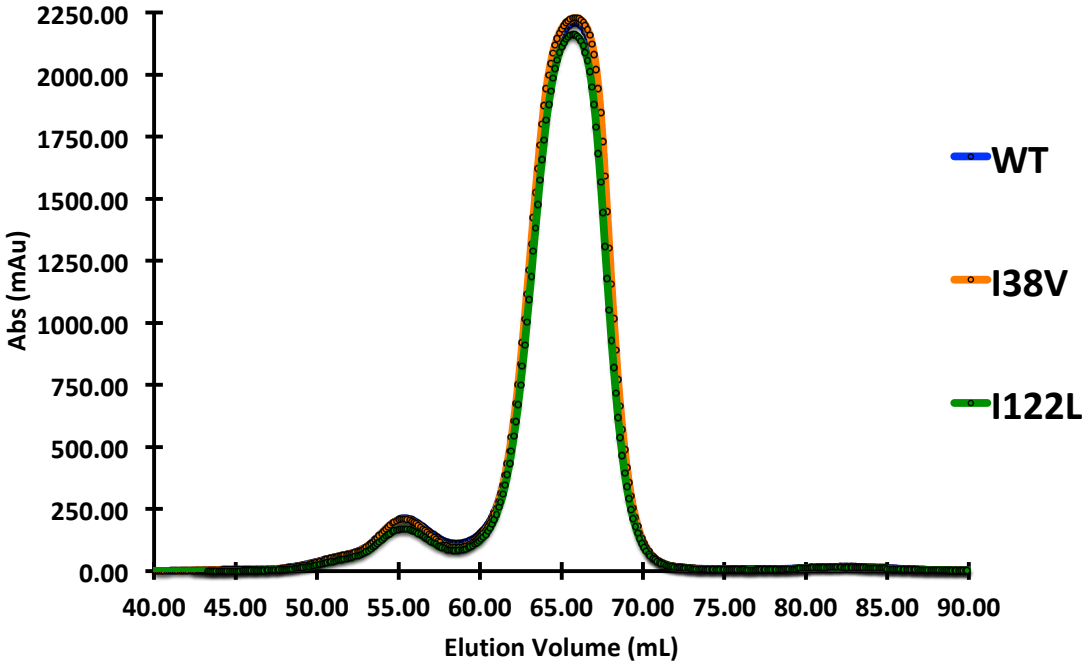

Figure A 1: Chromatogram of purified AmiE variants. SEC-FPLC chromatograms of AmiE proteins run on a HiLoad 16/600 Superdex 200 column at 1 mL/min in PBS as the mobile phase. No gross differences in oligomeric state were determined for the proteins.

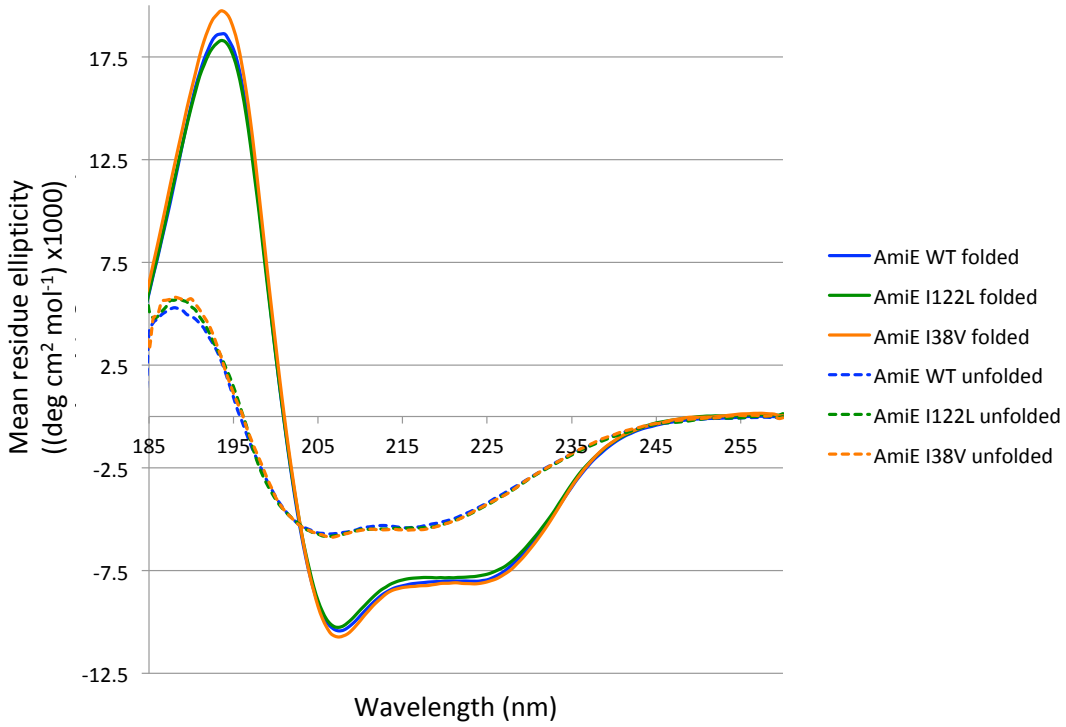

Figure A 2: Representative far-UV spectra of the folded and unfolded AmiE variants. Far-UV spectra of folded and unfolded 10 μM AmiE WT, 10 μM AmiE I122L, 10 μM AmiE I38V at 15°C (folded) and 90°C (unfolded) in 10 mM phosphate buffer at pH 7.5.

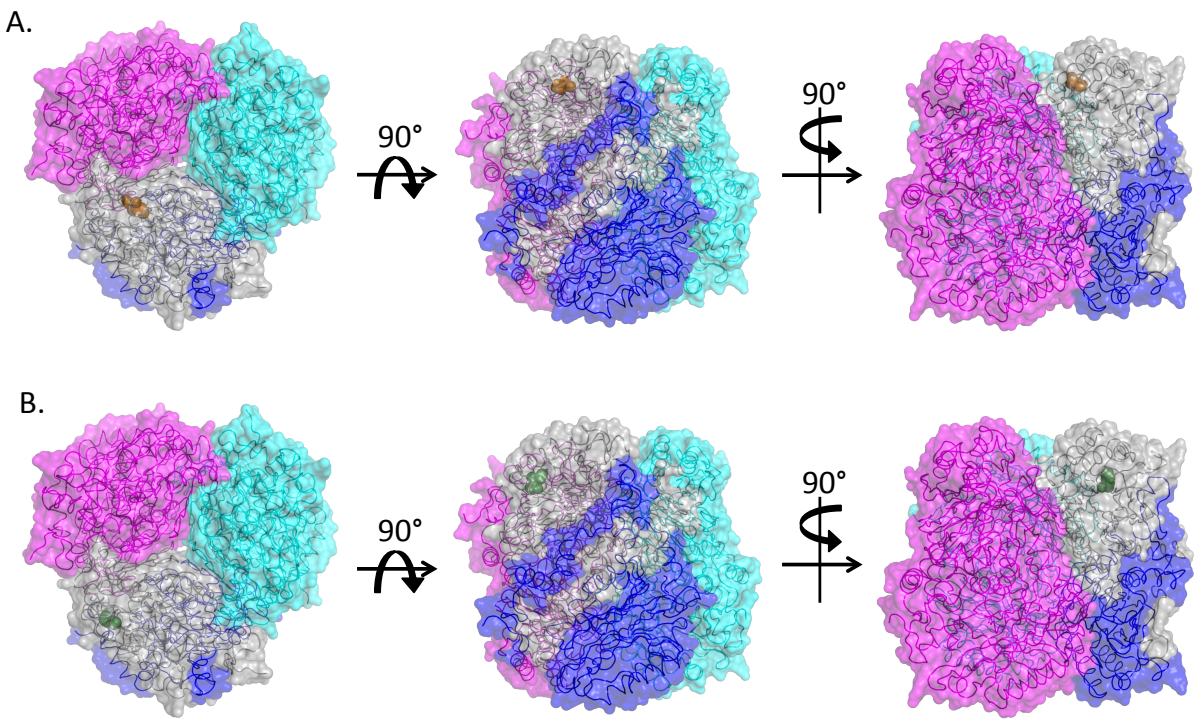

Figure A 3: Locations of I38V and I122L mutations in AmiE quaternary structure. A. Three perspectives on the location of the disrupting I38V mutation, the residue I38 is shown as orange spheres. B. Three perspectives on the location of the disrupting I122L mutation, the residue I122 is shown as green spheres. Both mutations are in the monomer core and are not be predicted to impact quaternary structure or assembly.

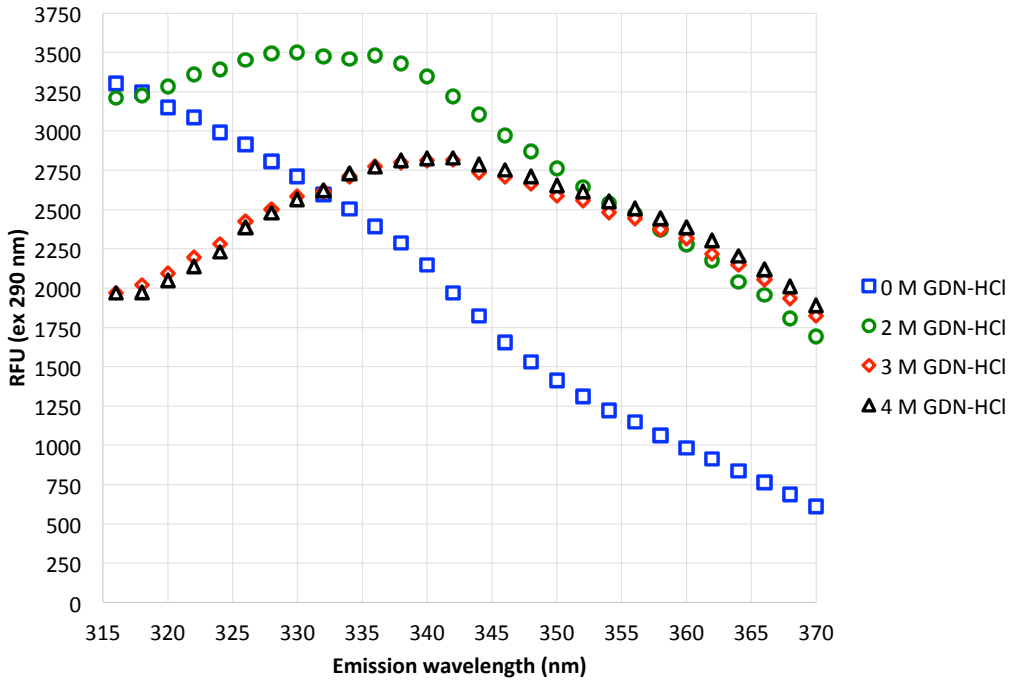

Figure A 4: AmiE WT tryptophan emission spectra as a function of GDN-HCl concentration. 52 μM Protein was incubated in PBS with the indicated concentration of GDN-HCl at 4oC for 8 hr before fluorescence measurements. The excitation wavelength was 290 nm, while emission was detected at 315-370 nm. The lack of an isosbestic point between spectra indicates more complicated unfolding than a two-state model.

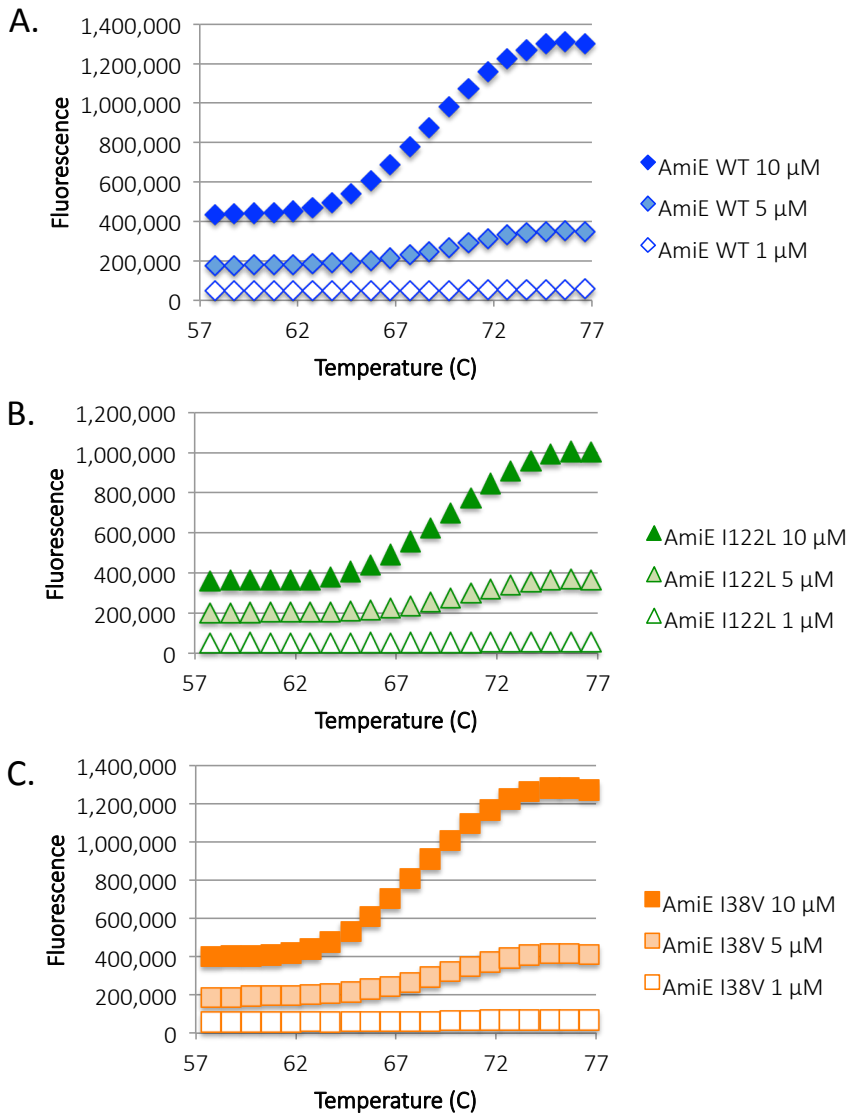

Figure A 5: Thermal shift analysis of the purified AmiE variants. A. Melt curves for 10, 5, and 1 μM AmiE WT. B. Melt curves for 10, 5, and 1 μM AmiE I122L. C. Melt curves for 10, 5, and 1 μM AmiE I38V. All experiments were performed with biological (n=2) and technical (n=3) replicates, and representative curves are shown.

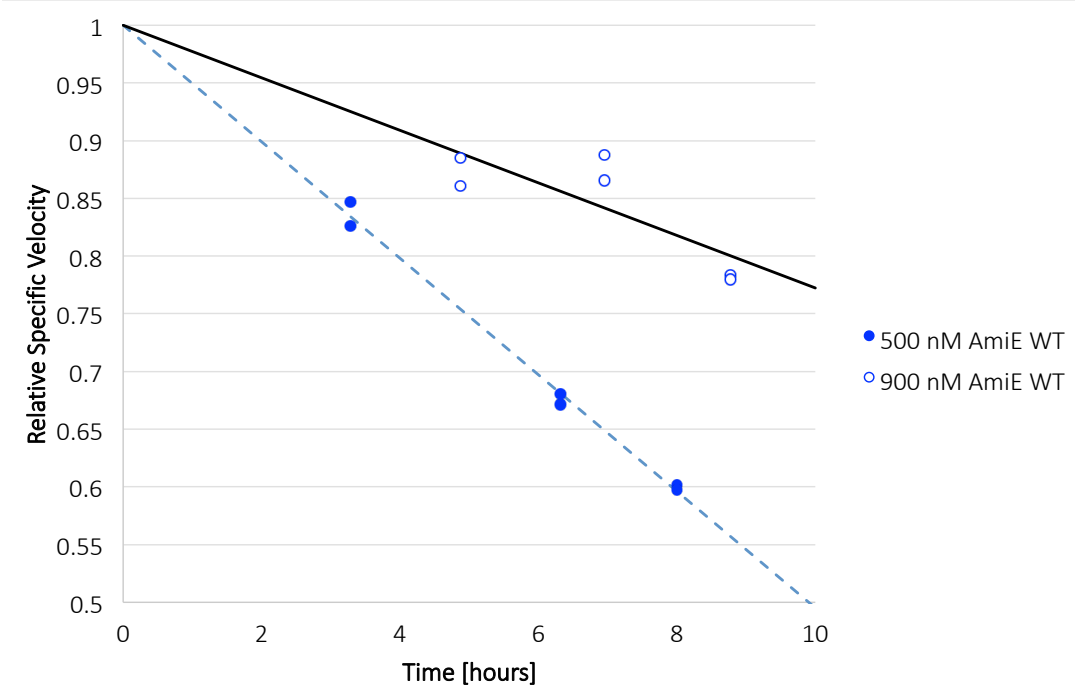

Figure A 6: Activity loss of AmiE WT following dilution. Comparison of the relative specific AmiE reaction velocity in a saturating amount of 20 mM acetamide following incubations at dilute concentration in 1x PBS over a time course. Dotted and solid lines represent the best fit linear regressions for the respective data sets (n=6).

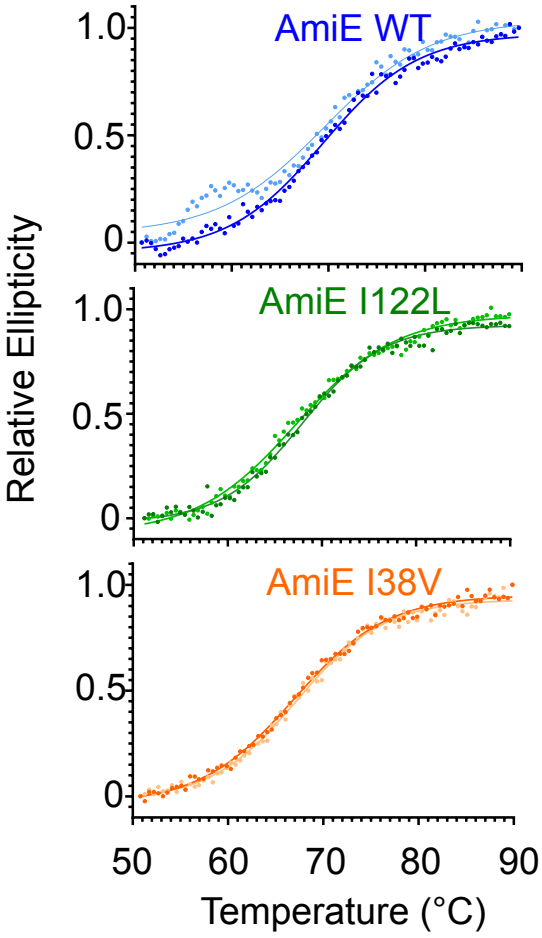

Figure A 7: Thermal denaturation monitored by far-UV circular dichroism. Melt curves for AmiE WT (blue), AmiE I122L (green), and AmiE I38V (orange). All experiments were performed by scanning λ_{222nm} at 1 μM protein concentration with 2 biological replicates (different colors on each panel represent a replicate experiment). Relative ellipticity ranges from completely folded state (0) to unfolded (1).

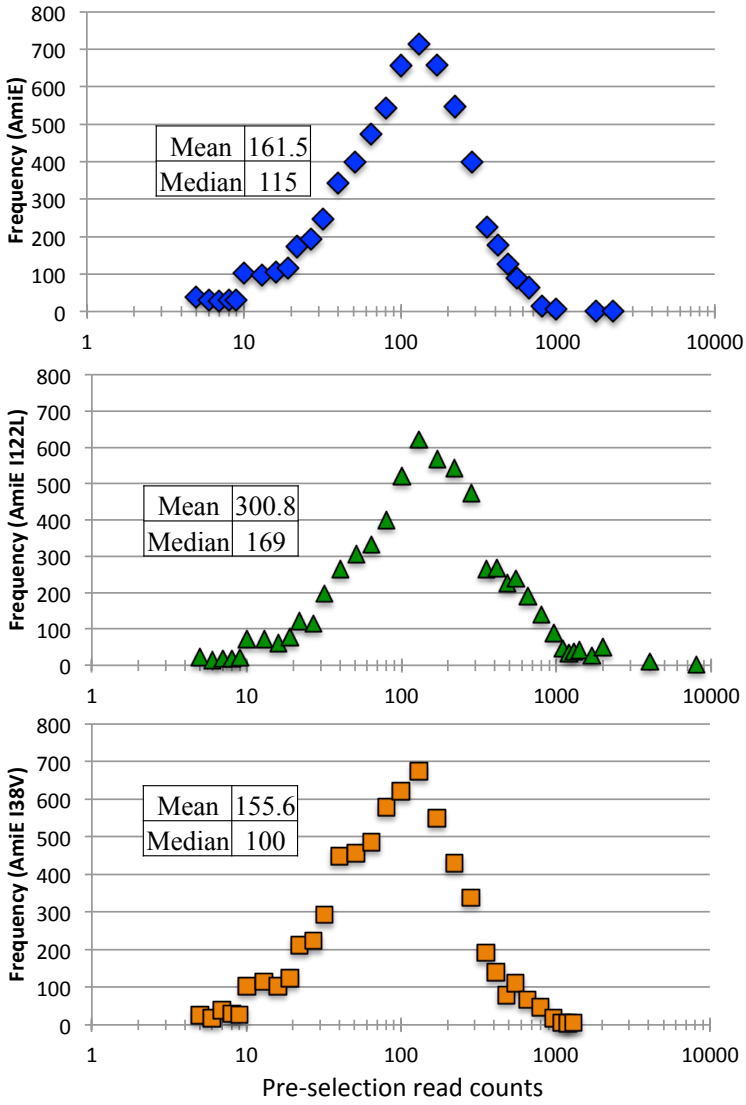

Figure A 8: Frequency distribution of pre-selection read counts for AmiE libraries. Mean and median pre-selection read counts for AmiE variants are reported on the plots.

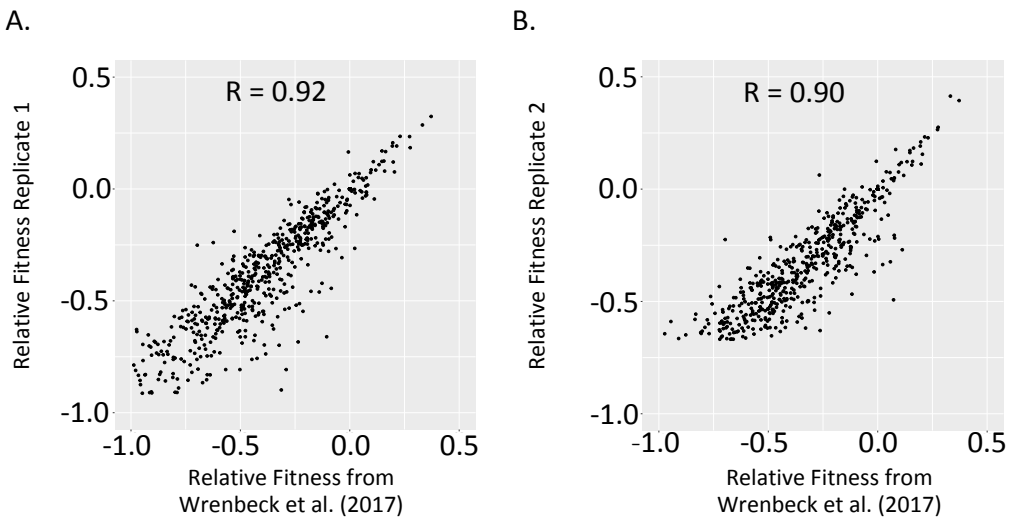

Figure A 9: Deep mutational scanning replicates for AmiE WT compared with previous literature. Comparison of relative fitness metrics for a AmiE WT mutational library covering residues 171-255 to those from same mutants reported originally in Wrenbeck et al (2017). These replicate deep mutational scans were performed in parallel to AmiE variant growth selections as internal controls. Pearson's correlation coefficients reported on plots. A. control for AmiE I122L growth selection. B. control for AmiE I38V growth selection.

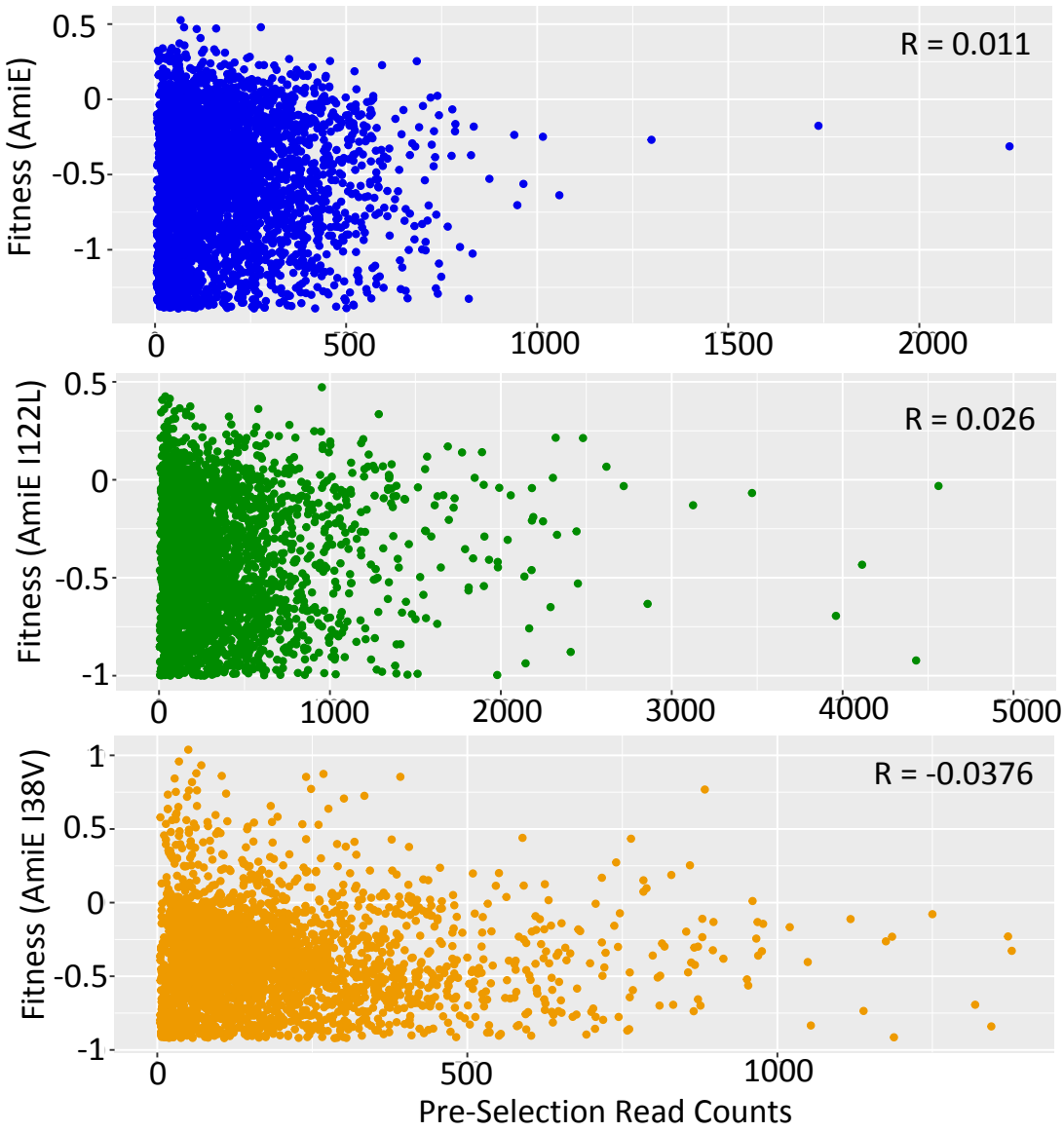

Figure A 10: Relative fitness metrics for AmiE proteins as a function of pre-selection read counts. Absolute Pearson's correlation coefficients reported on plots are below 0.04 in all cases, showing that less than 0.2% of the variance can be explained by initial frequency of a given mutant in the library.

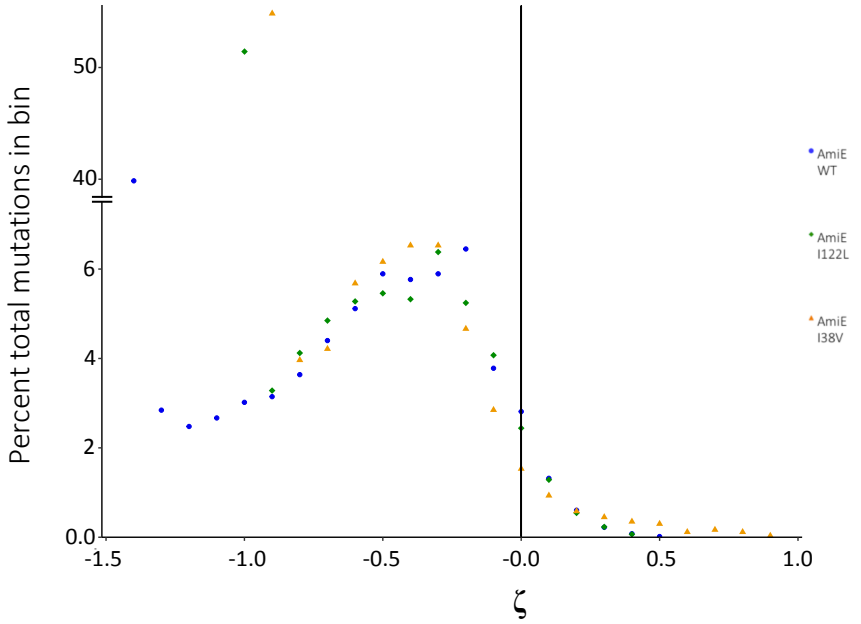

Figure A 11: Normalized distribution of fitness effects for WT, I122L, and I38V backgrounds.

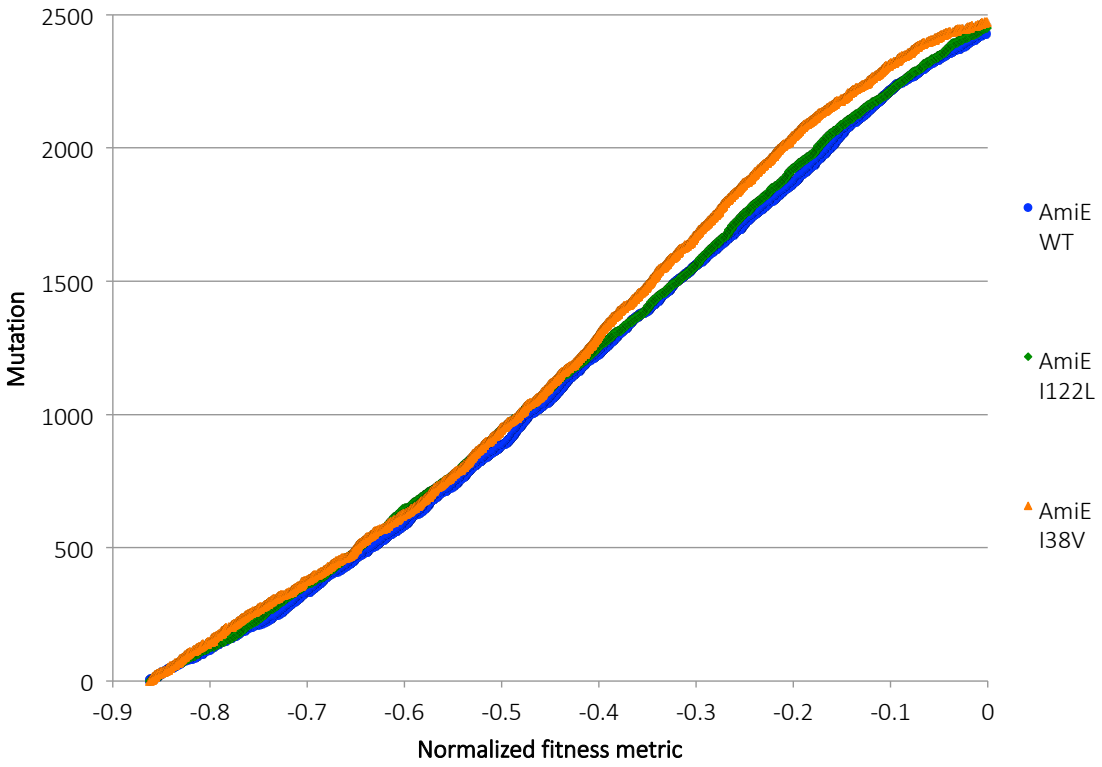

Figure A 12: Analysis of proportions of deleterious mutations. Cumulative distribution functions for the deleterious mutations for each enzyme variant are plotted above the lower bounds of experimental measurements.

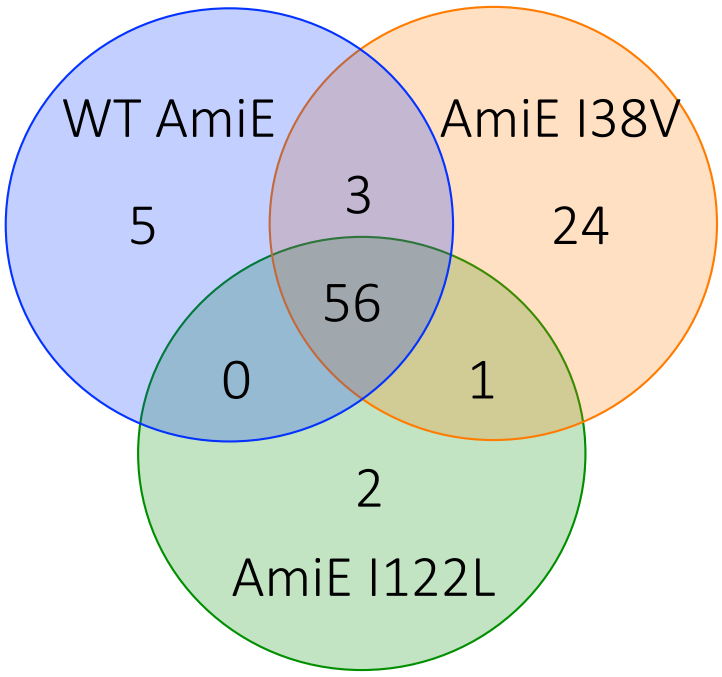

Figure A 13: Venn diagram of the shared and unique beneficial mutations using the strict cutoff. Mutations had to improve the growth rate by ≥10% to be classified as a beneficial mutation ($\zeta_i \geq 0.138$).

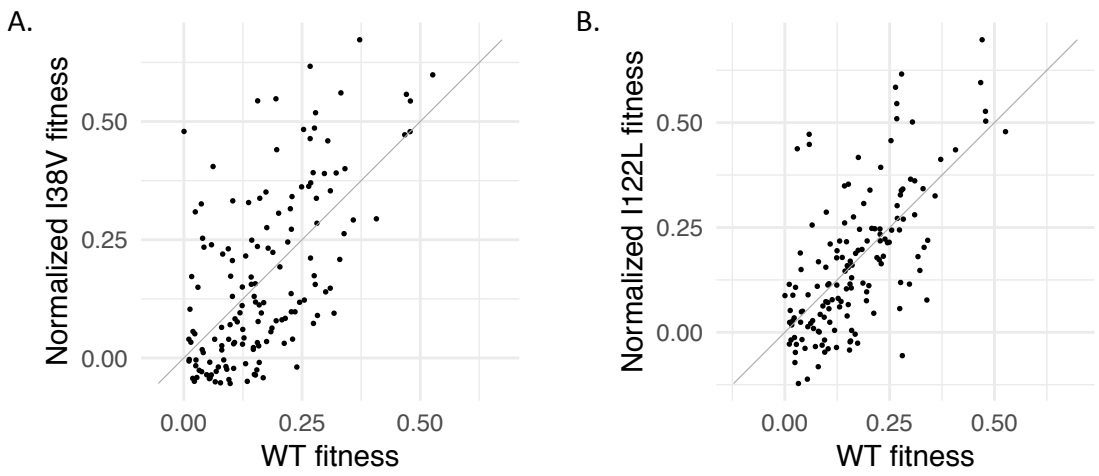

Figure A 14: Correlation analysis of linear regression normalized shared beneficial mutations. Comparison of the linear regression normalized fitness metrics for the beneficial mutations shared by all enzymes; solid lines represent the linear function Y = 1X + 0. A. Comparison of linear regression normalized AmiE I38V fitness metrics with AmiE WT. B. Comparison of linear regression normalized AmiE I122L fitness metrics with AmiE WT.

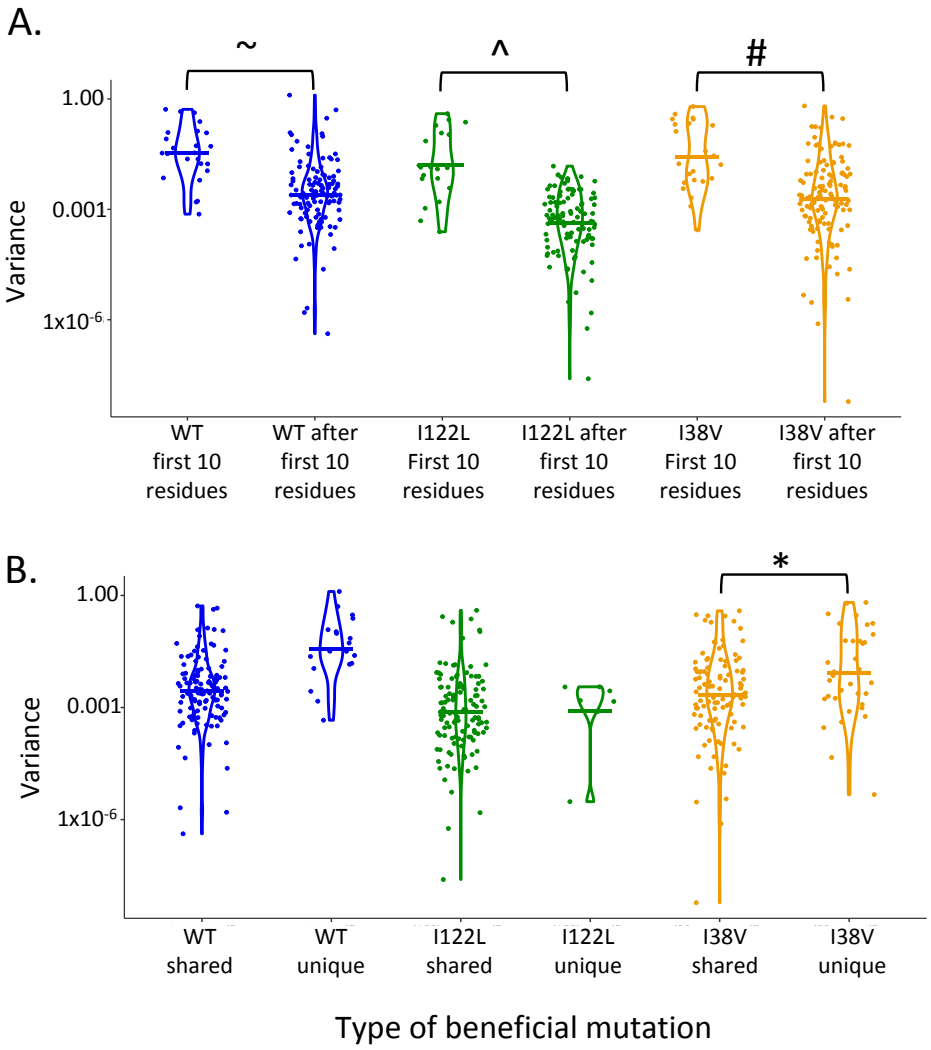

Figure A 15: Dot plots of fitness effect synonymous codon variances for beneficial mutations. A. Dot plots of the fitness effect synonymous codon variances for either the beneficial mutations located in the first 10 residues, or all other beneficial mutations after the first 10 residues (95% confidence interval cutoffs). The distribution of the variances is represented by the violin plot overlay and the colored marker lines represent the mean variance for the population, in all three enzymes the first 10 residues have significantly high variance than those outside of the window: ~ = WT p-value = 0.02, ^ = I122L p-value = 0.01, # = I38V p-value = 0.011 from students t-test. B. Dot plots of the fitness effect synonymous codon variances for the shared and unique beneficial mutations for the respective enzymes (95% confidence interval cutoffs) The distribution of the variances is represented by the violin plot overlay and the colored marker lines represent the mean variance for the population. I38V has a significantly high variance in the unique beneficial mutations: * = I38V p-value = 0.049 from Students t-test.

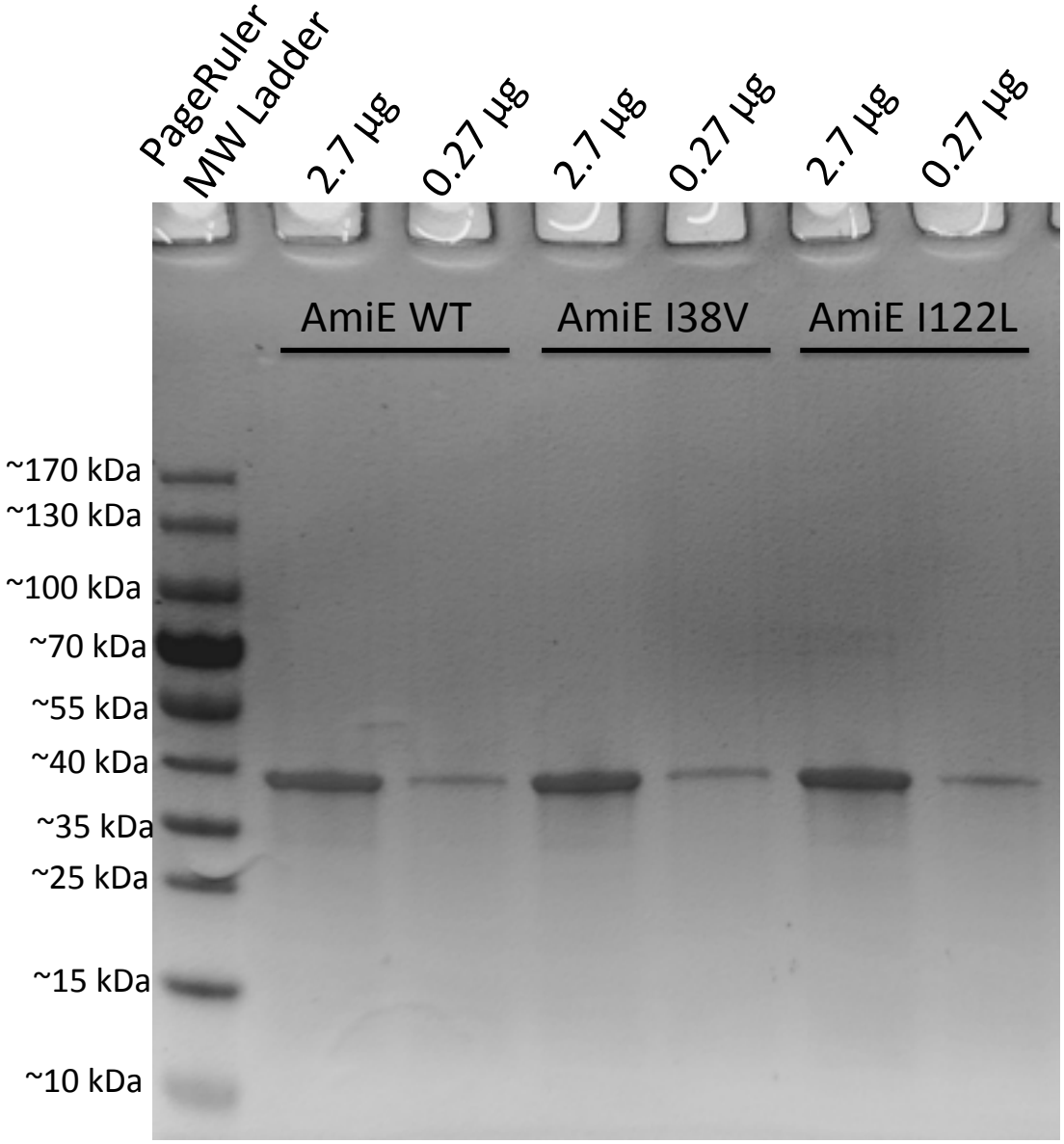

Figure A 16: SDS-PAGE analysis of the purity of the purified AmiE variants. Samples were denatured in SDS-PAGE loading buffer (Laemmli's buffer supplemented to 1.5% β-mercaptoethanol at 1x) at 98° for 10 minutes. Samples run on 4-20% Mini-PROTEAN® TGX™ precast gels (Bio-Rad) at 120 V for 1 hour and were washed and stained with SimplyBlue SafeStain (Thermo-Fisher) as described by the manufacturer. Molecular weight ruler used is the PageRuler Prestained Protein Ladder (10-180 kDa, Thermo-Fisher).

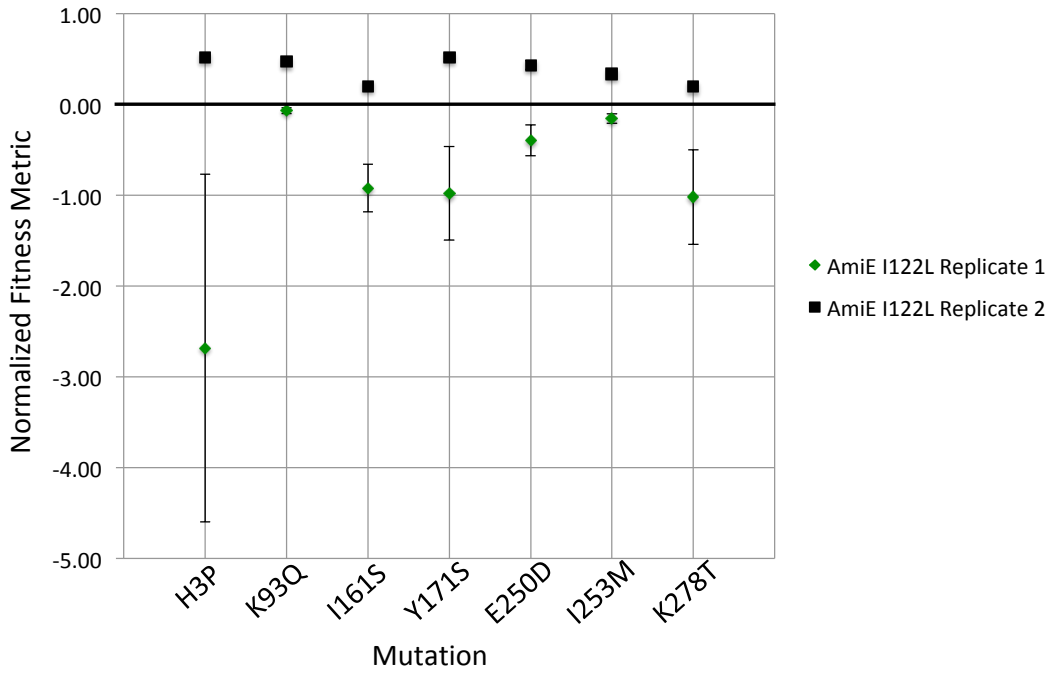

Figure A 17: Comparison of AmiE I122L outlier mutation technical replicates. Mutations removed from analysis have their fitness metrics for the respective technical replicates shown, error bars represent the 99.977% confidence intervals (3.5x σ) for the fitness metrics. This indicates that errors such as these should occur less than once per dataset for the size of our experiment.

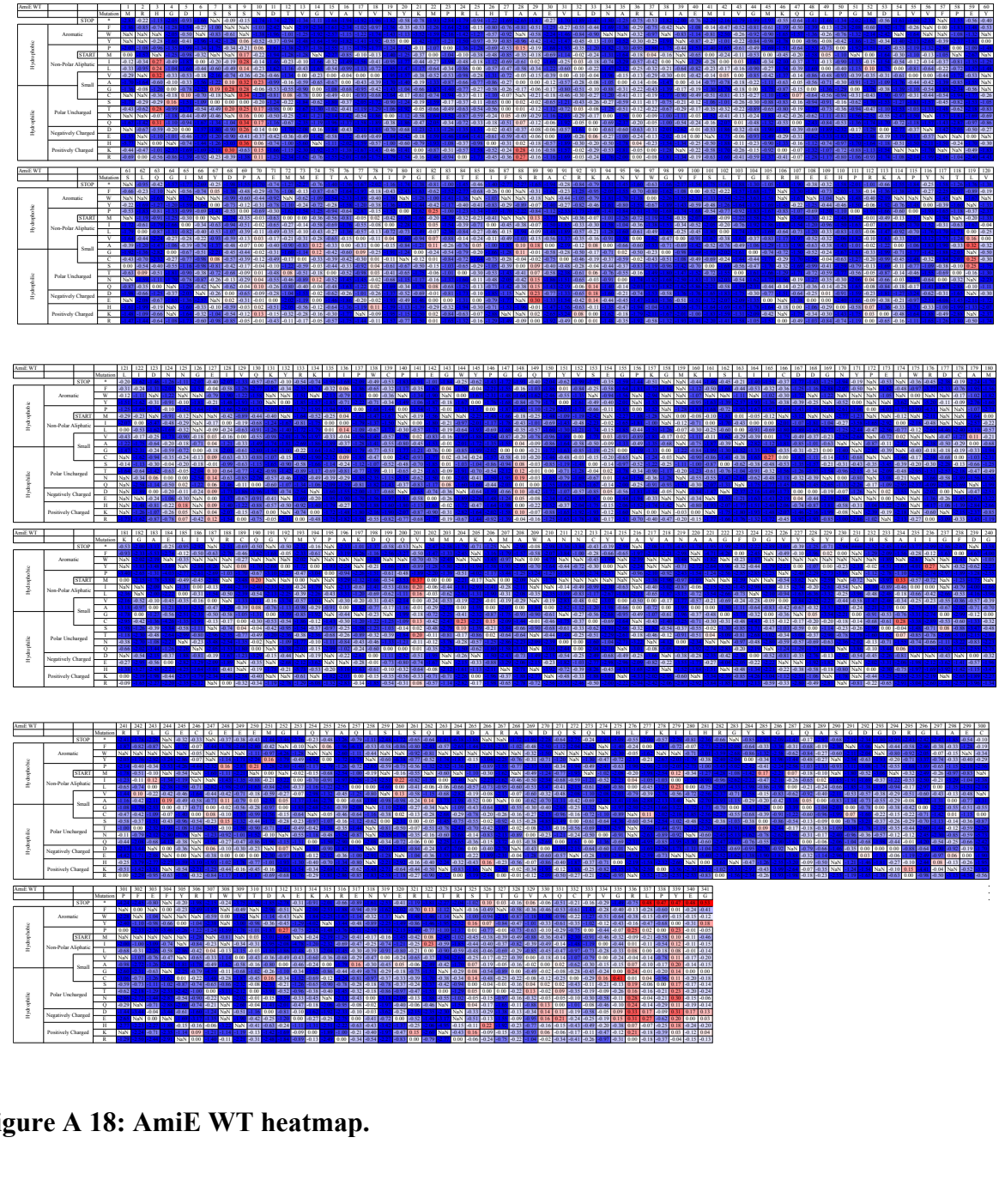

Figure A 18: AmiE WT heatmap.

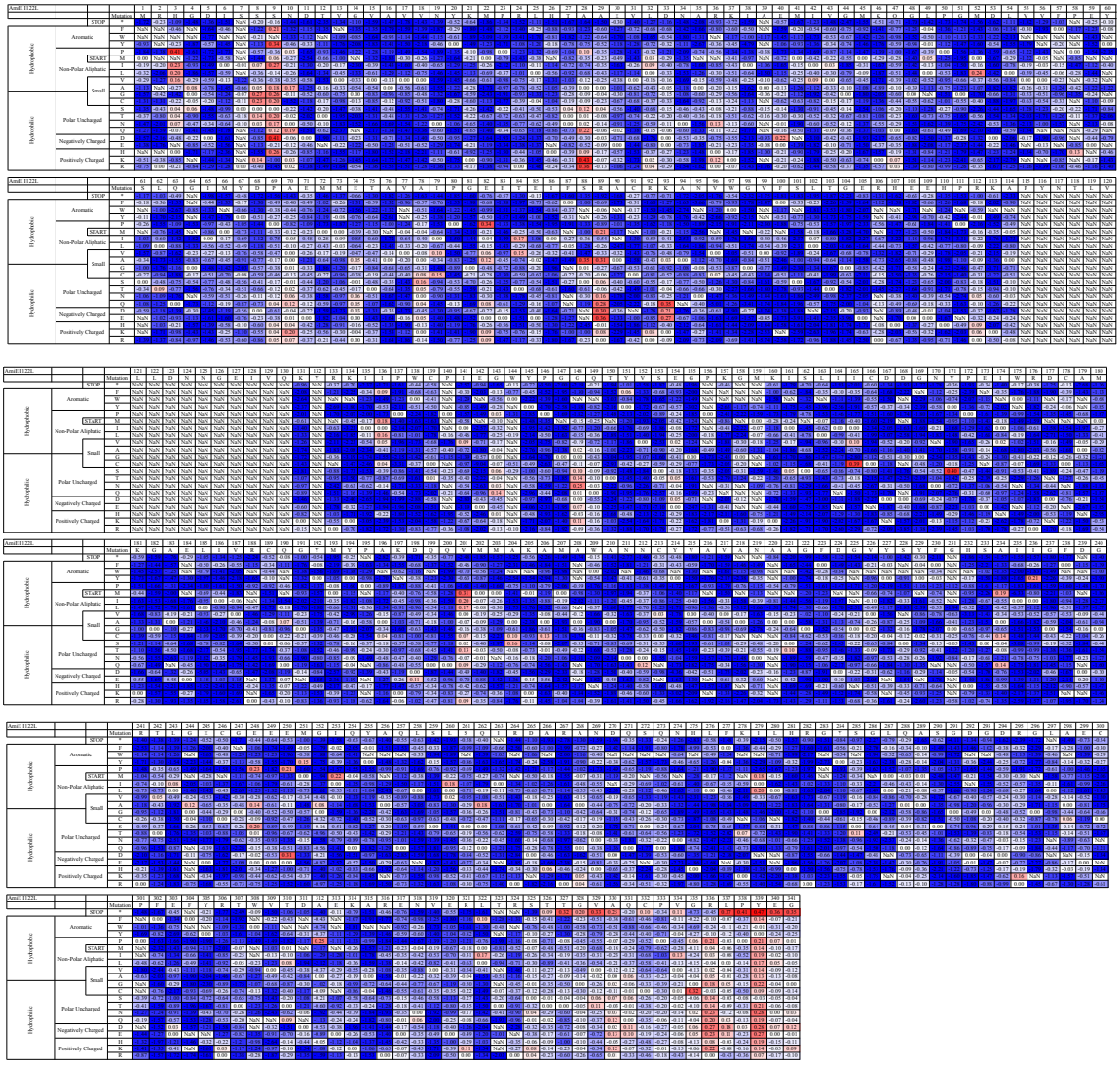

Figure A 19: AmiE I122L heatmap.

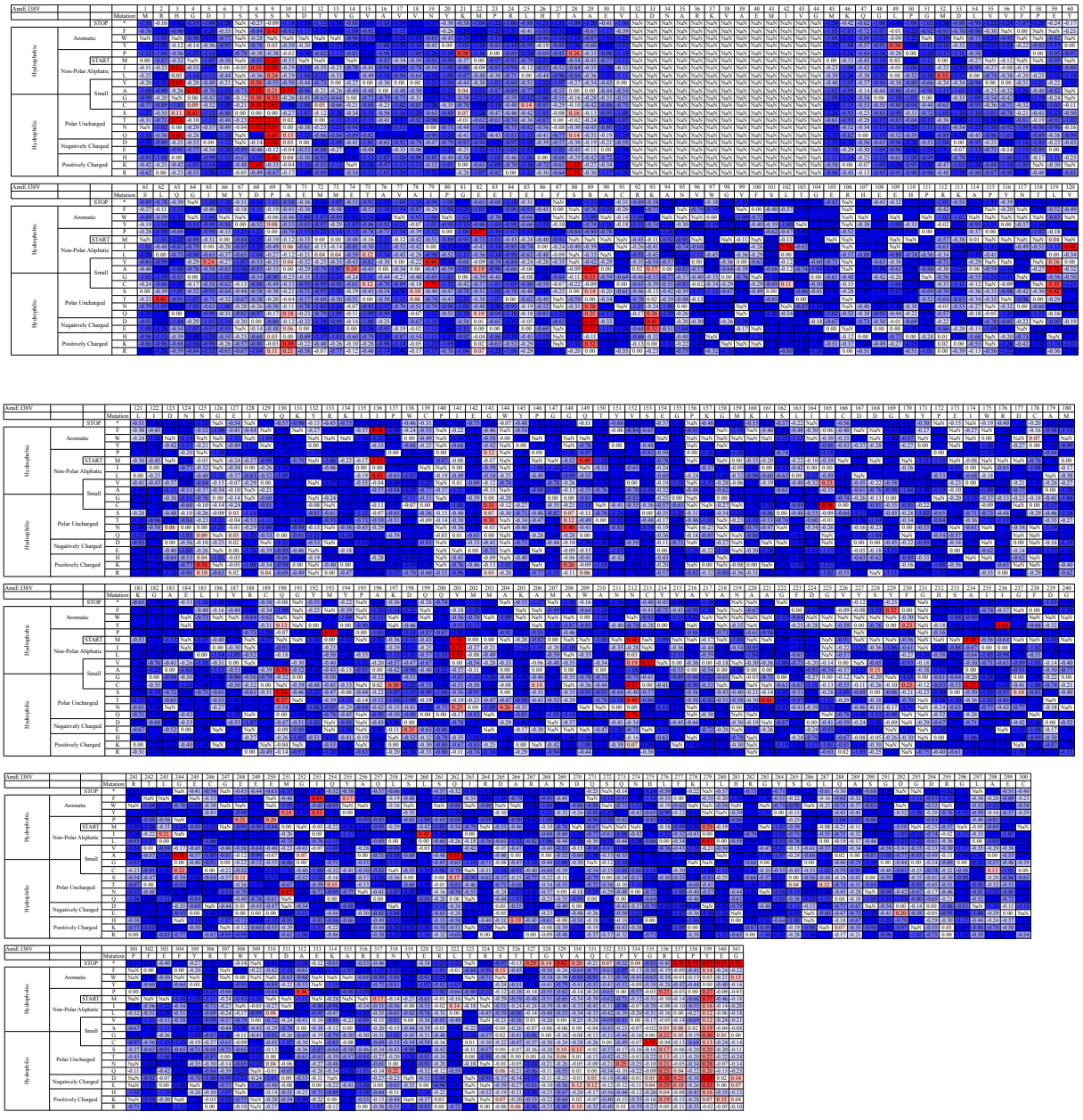

Figure A 20: AmiE I38V heatmap.

Table A 1: Relative fitness for I38V and I22L synonymous codons in the AmiE WT background. Green highlight indicates the codons used for the deep mutational scans presented in this work. Data is from Wrenbeck et al.[24].

	Relative Fitness			Relative Fitness
I38V codon (GTT)	-0.17		I122L codon (TTA)	-0.89
I38V codon (GTG)	-0.43		I122L codon (TTG)	-0.69
I38V codon (GTC)	-1.17		I122L codon (CTT)	-0.46
I38V codon (GTA)	-1.26		I122L codon (CTC)	-0.31
I38V average	-0.35		I122L codon (CTA)	-0.54
			I122L codon (CTG)	-0.69
			I122L average	-0.53

Table A 2: Biophysical analysis of the AmiE variants. Error values reported are 1 s.d. except in the kinetic analysis where 95% confidence interval is given.

	AmiE Variant		
	WT	**I122L**	**I38V**
Rosetta FilterScan score^	0.00	0.40	2.02
Relative k_{cat} *	1.00 ± 0.06	1.08 ± 0.06	1.02 ± 0.08
k_{cat} p-value#	1.00	0.57	0.79
Relative K_M *	1.00 ± 0.12	0.89 ± 0.15	0.82 ± 0.20
K_M p-value#	1.00	0.64	0.32
Relative protein yield₸	1 ± 0.06	0.26 ± 0.04	0.59 ± 0.08
Refolded enzyme velocity normalized to folded enzyme (%)	14.2 ± 2.47	0.114 ± 0.0494	0.064 ± 0.0250
Growth Rate in M9 (hr^{-1}) - pEDA2 plasmid	0.78 ± 0.01	0.79 ± 0.01	0.78 ± 0.02
Growth Rate in Selection Media (hr^{-1}) - pEDA2 plasmid	0.47 ± 0.02	0.18 ± 0.01	0.11 ± 0.01
Growth Rate in M9 (hr^{-1}) - pAG plasmid	0.79 ± 0.04	0.81 ± 0.03	0.79 ± 0.01
Growth Rate in Selection Media (hr^{-1}) - pAG plasmid	0.72 ± 0.02	0.40 ± 0.03	0.33 ± 0.02

^ >0 represents a less favorable score to the starting wild-type structure (i.e. decreased likelihood folding)
* Relative to AmiE WT
# Calculated for variant compared to AmiE WT
₸ Expressed via Studiers autoinduction

Table A 3: Thermal shift analysis data and statistics. Summary of thermal melting temperatures (T_m) determined using Boltzmann curve fit of a 2-state transition. Students t-test p-values are reported for tests comparing AmiE WT with the I38V and I122L variants (N = 6 from two biological replicates of each enzyme; ± indicates 1 s.d.).

Variant	Assaying Concentration (μM)	T_m (°C)	p-value	Assaying Concentration (μM)	T_m (°C)	p-value	Assaying Concentration (μM)	T_m (°C)	p-value
AmiE WT	1	N/A	-	5	69.6 ± 0.3	-	10	68.8 ± 0.2	-
AmiE I122L	1	N/A	N/A	5	69.2 ± 1.3	0.42	10	68.7 ± 1	0.74
AmiE I38V	1	N/A	N/A	5	69.3 ± 0.4	0.2	10	68.5 ± 0.4	0.16

Table A 4: Circular dichroism analysis of thermal denaturation statistics. Summary of apparent thermal melting temperatures (T_m) determined using Boltzmann curve fit of a 2-state folded-unfolded transition. Student's t-test p-values are reported for tests comparing AmiE WT with the I38V and I122L variants (N = 2 from respective biological replicates of each enzyme).

Variant	Average T_m (^oC)	Replicate 1	Replicate 2	p-value relative to AmiE WT
AmiE WT	69.5	69.5	69.6	-
AmiE I122L	67.5	67.9	67.2	0.03
AmiE I38V	67.1	66.9	67.4	0.01

Table A 5: DNA sequences of transcriptional elements in plasmids used for deep mutational scans. Differences between sequences are shown in red, the RBS in the 5' UTR is underlined.

	-35 Promoter	-10 Promoter	5' UTR
pEDA2 – AmiE	TTCCCG	TAATAT	TCAGGGAGACCACAACGG TTTCCCTCTACAAATAATT TTGTTTAACTTTCTAGAAA TAATTTTGTTTAACTT TAAGAAGTTTTTATACAT
pAG – AmiE I38V	TTGGCG	TACAAT	
pAG – AmiE I122L	TTGGCG	TACAAT	

Table A 6: Summary of MG1655 rph+ transformants obtained during selection strain mutant library preparation. 5 x 10^4 transformants corresponds to >99% theoretical library coverage.

Selection Strain SSM Library Preparation	Transformants Obtained
AmiE I122L pAG Tile 1	15 x 10^5
AmiE I122L pAG Tile 2	19x 10^5
AmiE I122L pAG Tile 3	7.5 x 10^5
AmiE I122L pAG Tile 4	23 x 10^5
AmiE I38V pAG Tile 1	6.9 x 10^5
AmiE I38V pAG Tile 2	5.8 x 10^5
AmiE I38V pAG Tile 3	7.5 x 10^5
AmiE I38V pAG Tile 4	7.8 x 10^5
WT amiE pEDA2 Tile 3	13 x 10^5

Table A 7: Summary of mutational library statistics.

Screen	Growth				
Enzyme	AmiE WT				
Tile Number	1	2	3	4	Cumulative
Residues	1-85	86-170	171-255	256-341	1-341
Population	Growth Selected	Growth Selected	Growth Selected	Growth Selected	Growth Selected
Number of mutated codons	85	85	85	86	341
Reference sequencing reads post quality filter	448689	393052	561875	511337	1914953
Selected sequencing reads post quality filter	1454211	3016045	2101231	1691121	8262608
Fold oversampling of codon combinations					Average
Reference total	81	71.6	101.8	91.6	86.5
Selected total	256.6	544	378.1	292.1	367.7
Reference non-synonymous	50.8	38.3	54.5	55.8	49.9
Selected non-synonymous	79.8	95.7	57.4	95	82.0
Percent of reads with:					Average
No nonsynonymous mutations	37.6	46.9	46.7	39.4	42.7
One nonsynonymous mutation	60.6	52.2	51.9	59.1	56.0
Multiple nonsynonymous mutation	1.8	0.9	1.4	1.4	1.4
Coverage of possible single nonsynonymous mutations	95.2	93.9	89	94.7	93.2
Screen	Growth				
Enzyme	AmiE I122L				
Tile Number	1	2	3	4	Cumulative
Residues	1-85	86-114, 131, 133-170	171-255	256-341	1-114,131,133-341
Population	Growth Selected	Growth Selected	Growth Selected	Growth Selected	Growth Selected
Number of mutated codons	85	68	85	86	324.0
Reference sequencing reads post quality filter	674718	593530	897746	1214190	3380184
Selected sequencing reads post quality filter	1286349	1002404	1115963	931995	4336711
Fold oversampling of codon combinations					Average
Reference total	118.7	99	146.8	210	143.6
Selected total	221.8	176.5	198.1	156	188.1
Reference non-synonymous	76.1	61.5	90.1	137	91.2
Selected non-synonymous	71.6	35.8	37.8	63.6	52.2
Reference population percent of reads with:					Average
No nonsynonymous mutations	35.4	35.3	35.2	34	35.0
One nonsynonymous mutation	60.4	55.4	53.7	61.2	57.7
Multiple nonsynonymous mutation	4.3	9.3	11.1	4.8	7.4
Coverage of possible single nonsynonymous mutations	94.5	92.5	93	94.8	93.7
Screen	Growth				
Enzyme	AmiE I38V				
Tile Number	1	2	3	4	Cumulative
Residues	1-31, 45-85	86-170	171-255	256-341	1-31, 45-341
Population	Growth Selected	Growth Selected	Growth Selected	Growth Selected	Growth Selected
Number of mutated codons	72	85	85	86	328
Reference sequencing reads post quality filter	550823	500238	525635	490442	2067138
Selected sequencing reads post quality filter	1373021	1060307	991922	1129295	4554545
Fold oversampling of codon combinations					Average
Reference total	100.7	91.5	96.1	88.6	94.2
Selected total	236.5	189	173.1	185	195.9
Reference non-synonymous	50.3	45.7	48.3	46	47.6
Selected non-synonymous	96.4	42.6	69.8	81	72.5
Percent of reads with:					Average
No nonsynonymous mutations	50.5	50.5	50.3	48.6	50.0
One nonsynonymous mutation	48.9	48.9	49.2	50.8	49.5
Multiple nonsynonymous mutation	0.6	0.5	0.6	0.6	0.6
Coverage of possible single nonsynonymous mutations	96.7	88.6	87.9	93.9	91.8

Table A 8: Goodness-of-fit test statistics for distribution fittings. For all enzymes the determined p-values indicate a failure to reject the null hypotheses that they can be fit by a general Pareto distribution with a negative shape parameter. All calculations were performed as in Wrenbeck et al.[24].

General Pareto distribution	AmiE WT	AmiE I122L	AmiE I38V
bootstrap test p-value	0.16	0.29	0.10
shape parameter	-0.28	-0.33	-0.32

Table A 9: Deleterious empirical cumulative distribution function (ECDF) analysis statistics. A table of the two-sample Kolmogorov-Smirnoff test results for comparing the deleterious mutation ECDFs for the three enzymes.

Pairing (x-y)	K-S Test	D	p-value	Null hypothesis	Alternative hypothesis	AmiE WT (N)	AmiE variant (N)
AmiE WT – AmiE I38V	two sided	0.058	0.0005	is equal to	not equal	2428	2472
AmiE WT – AmiE I38V	greater	0.002	0.99	x not greater than y	the CDF of x lies above that of y	2428	2472
AmiE WT – AmiE I38V	less	0.058	0.0002	x not less than y	the CDF of x lies below that of y	2428	2472
AmiE WT – AmiE I122L	two sided	0.024	0.48	is equal to	not equal	2428	2450
AmiE WT – AmiE I122L	greater	0.010	0.77	x not greater than y	the CDF of x lies above that of y	2428	2450
AmiE WT – AmiE I122L	less	0.024	0.24	x not less than y	the CDF of x lies below that of y	2428	2450

Table A 10: Inner and outer primers for PCR reactions for Illumina sequencing. Red indicates overhang regions for attaching Illumina adapter primers (inner PCR primers) or overhangs for attaching to inner PCR product (outer PCR primers), black is the overlap region in the gene or the barcode, blue is the Illumina adapter.

Inner PCR primers	Sequence (5' to 3')
Fwd_Tile_1	gttcagagttctacagtccgacgatcttaactttaagaagtttttatacat
Fwd_Tile_2	gttcagagttctacagtccgacgatcggcgaagaaacggaa
Fwd_Tile_3	gttcagagttctacagtccgacgatcctgcgatgacggtaat
Fwd_Tile_4	gttcagagttctacagtccgacgatcaagaaatgggcattcaatac
Rev_Tile_1	ccttggcacccgagaattccaaagcacggctaaagat
Rev_Tile_2	ccttggcacccgagaattccactctccaaatttccggata
Rev_Tile_3	ccttggcacccgagaattccacagagacaactgcgc
Rev_Tile_4	ccttggcacccgagaattccatggtggtgctcgag
Illumina outer primer adapter	aatgatacggcgaccaccgagatctacacgttcagagttctacagtccga
Illumina outer PCR adapters and barcodes	
RPI37 (AmiE WT unselected, Tile 1)	caagcagaagacggcatacgagatATTCCGgtgactggagttccttggcaccc gagaattcca
RPI22 (AmiE WT unselected, Tile 2)	caagcagaagacggcatacgagatCGTACGgtgactggagttccttggcaccc gagaattcca
RPI39 (AmiE WT unselected, Tile 3)	caagcagaagacggcatacgagatGTATAGgtgactggagttccttggcaccc gagaattcca
RPI40 (AmiE WT unselected, Tile 4)	caagcagaagacggcatacgagatTCTGAGgtgactggagttccttggcaccc gagaattcca
RPI41 (AmiE WT selected, Tile 1 replicate 1)	caagcagaagacggcatacgagatGTCGTCgtgactggagttccttggcaccc gagaattcca
RPI38 (AmiE WT selected, Tile 1 replicate 2)	caagcagaagacggcatacgagatAGCTAGgtgactggagttccttggcaccc gagaattcca
RPI33 (AmiE WT selected, Tile 2 replicate 1)	caagcagaagacggcatacgagatCGCCTGgtgactggagttccttggcaccc gagaattcca
RPI34 (AmiE WT selected, Tile 2 replicate 2)	caagcagaagacggcatacgagatGCCATGgtgactggagttccttggcaccc gagaattcca
RPI43 (AmiE WT selected, Tile 3 replicate 1)	caagcagaagacggcatacgagatGCTGTAgtgactggagttccttggcaccc gagaattcca
RPI40 (AmiE WT selected, Tile 3 replicate 2)	caagcagaagacggcatacgagatTCTGAGgtgactggagttccttggcaccc gagaattcca
RPI44 (AmiE WT selected, Tile 4 replicate 1)	caagcagaagacggcatacgagatATTATAgtgactggagttccttggcaccc gagaattcca
RPI41 (AmiE WT selected, Tile 4 replicate 2)	caagcagaagacggcatacgagatGTCGTCgtgactggagttccttggcaccc gagaattcca
RPI20 (AmiE I38V unselected, Tile 1)	caagcagaagacggcatacgagatGGCCACgtgactggagttccttggcaccc gagaattcca
RPI21 (AmiE I38V unselected, Tile 2)	caagcagaagacggcatacgagatCGAAACgtgactggagttccttggcaccc gagaattcca
RPI27 (AmiE I38V unselected, Tile 3)	caagcagaagacggcatacgagatAGGAATgtgactggagttccttggcaccc gagaattcca

Table A 10 (cont'd)

RPI34 (AmiE I38V unselected, Tile 4)	caagcagaagacggcatacgagatGCCATGgtgactggagttccttggcaccc gagaattcca
RPI37 (AmiE I38V selected, Tile 1 replicate 1)	caagcagaagacggcatacgagatATTCCGgtgactggagttccttggcaccc gagaattcca
RPI38 (AmiE I38V selected, Tile 1 replicate 2)	caagcagaagacggcatacgagatAGCTAGgtgactggagttccttggcaccc gagaattcca
RPI16 (AmiE I38V selected, Tile 2 replicate 1)	caagcagaagacggcatacgagatGGACGGgtgactggagttccttggcaccc gagaattcca
RPI14 (AmiE I38V selected, Tile 2 replicate 2)	caagcagaagacggcatacgagatGGAACTgtgactggagttccttggcaccc gagaattcca
RPI48 (AmiE I38V selected, Tile 3 replicate 1)	caagcagaagacggcatacgagatTGCCGAgtgactggagttccttggcaccc gagaattcca
RPI23 (AmiE I38V selected, Tile 3 replicate 2)	caagcagaagacggcatacgagatCCACTCgtgactggagttccttggcaccc gagaattcca
RPI32 (AmiE I38V selected, Tile 4 replicate 1)	caagcagaagacggcatacgagatTGAGTGgtgactggagttccttggcaccc gagaattcca
RPI44 (AmiE I38V selected, Tile 4 replicate 2)	caagcagaagacggcatacgagatATTATAgtgactggagttccttggcaccc gagaattcca
RPI19 (AmiE WT control, AmiE I38V selection, unselected, Tile 3)	caagcagaagacggcatacgagatTTTCACgtgactggagttccttggcaccc gagaattcca
RPI26 (AmiE WT control, AmiE I38V selection, selected, Tile 3)	caagcagaagacggcatacgagatGCTCATgtgactggagttccttggcaccc gagaattcca
RPI33 (AmiE I122L unselected, Tile 1, sequencing runs 1 and 2)	caagcagaagacggcatacgagatCGCCTGgtgactggagttccttggcaccc gagaattcca
RPI46 (AmiE I122L unselected, Tile 2, sequencing runs 1 and 2)	caagcagaagacggcatacgagatTCGGGAgtgactggagttccttggcaccc gagaattcca
RPI40 (AmiE I122L unselected, Tile 3, sequencing runs 1and 2)	caagcagaagacggcatacgagatTCTGAGgtgactggagttccttggcaccc gagaattcca
RPI29 (AmiE I122L unselected, Tile 4, sequencing runs 1 and 2)	caagcagaagacggcatacgagatTAGTTGgtgactggagttccttggcaccc gagaattcca
RPI17 (AmiE I122L selected, Tile 1 replicate 1)	caagcagaagacggcatacgagatCTCTACgtgactggagttccttggcaccc gagaattcca
RPI32 (AmiE I122L selected, Tile 1 replicate 2)	caagcagaagacggcatacgagatTGAGTGgtgactggagttccttggcaccc gagaattcca
RPI12 (AmiE I122L selected, Tile 2 replicate 1)	caagcagaagacggcatacgagatTACAAGgtgactggagttccttggcaccc gagaattcca
RPI42 (AmiE I122L selected, Tile 2 replicate 2)	caagcagaagacggcatacgagatCGATTAgtgactggagttccttggcaccc gagaattcca
RPI25 (AmiE I122L selected, Tile 3 replicate 1)	caagcagaagacggcatacgagatATCAGTgtgactggagttccttggcaccc gagaattcca
RPI46 (AmiE I122L selected, Tile 3 replicate 2)	caagcagaagacggcatacgagatTCGGGAgtgactggagttccttggcaccc gagaattcca
RPI31 (AmiE I122L selected, Tile 4 replicate 1)	caagcagaagacggcatacgagatATCGTGgtgactggagttccttggcaccc gagaattcca
RPI15 (AmiE I122L selected, Tile 4 replicate 2)	caagcagaagacggcatacgagatTGACATgtgactggagttccttggcaccc gagaattcca
RPI7 (AmiE WT control, AmiE I122L selection, unselected, Tile 3)	caagcagaagacggcatacgagatGATCTGgtgactggagttccttggcaccc gagaattcca
RPI10 (AmiE WT control, AmiE I122L selection, selected, Tile 3)	caagcagaagacggcatacgagatAAGCTAgtgactggagttccttggcaccc gagaattcca

Table A 11: Beneficial mutations shared by all enzymes.

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
9T	0.25	0.24	0.26	0.20	0.19	0.20	0.59	0.60	0.59
9Q	0.34	0.34	0.34	0.12	0.11	0.12	0.45	0.46	0.44
9N	0.16	0.21	0.11	0.17	0.14	0.18	0.56	0.56	0.56
9I	0.28	0.26	0.29	0.27	0.25	0.28	0.49	0.49	0.48
9H	0.36	0.36	0.36	0.26	0.24	0.28	0.49	0.49	0.50
9G	0.28	0.25	0.29	0.26	0.25	0.27	0.33	0.31	0.34
9D	0.26	0.28	0.25	0.41	0.38	0.43	0.60	0.59	0.60
9C	0.34	0.34	0.34	0.20	0.19	0.21	0.65	0.65	0.65
9A	0.32	0.31	0.32	0.18	0.19	0.16	0.21	0.22	0.21
93E	0.14	0.14	0.15	0.27	0.26	0.29	0.32	0.32	0.32
93D	0.18	0.17	0.19	0.21	0.20	0.23	0.41	0.40	0.42
93A	0.08	0.07	0.09	0.03	0.03	0.02	0.17	0.16	0.18
8T	0.20	0.20	0.20	0.14	0.15	0.13	0.52	0.52	0.51
8K	0.30	0.30	0.30	0.14	0.14	0.14	0.64	0.63	0.64
8G	0.28	0.28	0.28	0.27	0.26	0.27	0.30	0.31	0.29
8A	0.10	0.12	0.09	0.05	0.04	0.06	0.55	0.55	0.55
89S	0.09	0.07	0.11	0.06	0.08	0.03	0.14	0.12	0.15
89Q	0.15	0.14	0.16	0.28	0.28	0.27	0.25	0.25	0.25
89N	0.15	0.14	0.16	0.16	0.16	0.17	0.30	0.29	0.31
89E	0.30	0.29	0.31	0.36	0.36	0.36	0.73	0.73	0.74
89D	0.23	0.21	0.24	0.30	0.30	0.30	0.57	0.57	0.55
89A	0.18	0.17	0.18	0.31	0.32	0.30	0.47	0.46	0.49
82Q	0.08	0.07	0.09	0.08	0.08	0.07	0.10	0.09	0.11
82P	0.25	0.25	0.26	0.34	0.33	0.34	0.77	0.77	0.77
82A	0.11	0.10	0.12	0.12	0.13	0.11	0.19	0.18	0.19
79V	0.04	0.05	0.03	0.10	0.11	0.09	0.44	0.44	0.44
78S	0.04	0.02	0.05	0.16	0.16	0.16	0.10	0.11	0.10
74A	0.12	0.13	0.12	0.05	0.05	0.05	0.24	0.22	0.25
70Q	0.10	0.08	0.11	0.12	0.11	0.12	0.18	0.17	0.19
70K	0.13	0.13	0.13	0.20	0.20	0.19	0.39	0.38	0.39
62T	0.09	0.10	0.09	0.09	0.11	0.08	0.41	0.41	0.41
52L	0.10	0.09	0.11	0.24	0.23	0.24	0.33	0.33	0.33
4S	0.16	0.14	0.17	0.06	0.08	0.04	0.41	0.42	0.41
3V	0.32	0.32	0.33	0.16	0.16	0.15	0.64	0.63	0.64
3L	0.24	0.24	0.24	0.20	0.20	0.20	0.05	0.03	0.07
3I	0.27	0.25	0.28	0.23	0.21	0.24	0.61	0.61	0.60
341D	0.13	0.11	0.14	0.12	0.10	0.13	0.14	0.13	0.14
339V	0.11	0.10	0.11	0.14	0.13	0.15	0.12	0.11	0.13
339T	0.23	0.22	0.23	0.21	0.19	0.22	0.22	0.22	0.21

Table A 11 (cont'd)

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
339Q	0.11	0.10	0.11	0.19	0.19	0.20	0.20	0.21	0.18
339P	0.23	0.22	0.23	0.21	0.21	0.22	0.27	0.28	0.27
339N	0.30	0.28	0.31	0.28	0.26	0.30	0.28	0.29	0.27
339M	0.10	0.09	0.11	0.14	0.13	0.14	0.37	0.36	0.38
339L	0.08	0.08	0.08	0.17	0.17	0.17	0.12	0.11	0.13
339I	0.12	0.13	0.12	0.19	0.18	0.19	0.16	0.15	0.18
339H	0.18	0.16	0.20	0.19	0.17	0.20	0.16	0.15	0.16
339G	0.14	0.12	0.15	0.22	0.21	0.23	0.30	0.30	0.30
339E	0.20	0.18	0.21	0.27	0.27	0.27	0.35	0.36	0.35
339D	0.31	0.29	0.32	0.28	0.28	0.28	0.58	0.58	0.59
339C	0.11	0.08	0.12	0.09	0.08	0.09	0.13	0.13	0.12
339A	0.20	0.19	0.20	0.13	0.12	0.13	0.19	0.19	0.19
337E	0.27	0.26	0.28	0.11	0.09	0.12	0.18	0.20	0.16
337D	0.17	0.16	0.17	0.18	0.18	0.18	0.25	0.25	0.24
336T	0.16	0.15	0.16	0.14	0.13	0.15	0.13	0.12	0.13
336S	0.19	0.17	0.19	0.14	0.14	0.14	0.17	0.17	0.16
336Q	0.24	0.25	0.24	0.20	0.19	0.20	0.25	0.24	0.26
336P	0.25	0.26	0.25	0.21	0.20	0.22	0.25	0.26	0.25
336N	0.28	0.27	0.29	0.23	0.23	0.23	0.21	0.20	0.21
336K	0.21	0.18	0.22	0.22	0.21	0.22	0.19	0.17	0.22
336H	0.07	0.06	0.07	0.08	0.07	0.09	0.04	0.03	0.04
336G	0.24	0.24	0.23	0.18	0.16	0.19	0.22	0.21	0.23
336E	0.31	0.30	0.31	0.23	0.23	0.24	0.29	0.27	0.31
336D	0.33	0.33	0.33	0.27	0.26	0.28	0.38	0.37	0.38
336A	0.07	0.07	0.07	0.05	0.05	0.06	0.05	0.06	0.05
335D	0.09	0.09	0.09	0.06	0.06	0.05	0.05	0.05	0.05
335C	0.41	0.38	0.42	0.32	0.35	0.30	0.50	0.50	0.50
331E	0.21	0.18	0.23	0.10	0.10	0.10	0.12	0.13	0.12
330S	0.02	0.03	0.02	0.07	0.07	0.06	0.15	0.14	0.16
330E	0.16	0.18	0.14	0.13	0.11	0.16	0.12	0.09	0.14
329T	0.13	0.13	0.13	0.11	0.13	0.09	0.06	0.06	0.06
329S	0.04	0.05	0.04	0.06	0.06	0.07	0.10	0.09	0.10
322I	0.23	0.21	0.24	0.17	0.16	0.18	0.14	0.14	0.13
312P	0.27	0.25	0.28	0.25	0.27	0.23	0.38	0.38	0.38
28R	0.27	0.27	0.27	0.36	0.36	0.37	0.74	0.74	0.74
28P	0.15	0.16	0.14	0.10	0.10	0.09	0.26	0.25	0.27
28K	0.28	0.26	0.29	0.43	0.42	0.43	0.82	0.81	0.82
279M	0.12	0.10	0.14	0.18	0.15	0.19	0.29	0.30	0.28
279L	0.23	0.21	0.24	0.20	0.20	0.20	0.47	0.46	0.47

Table A 11 (cont'd)

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
266H	0.16	0.14	0.17	0.06	0.07	0.05	0.19	0.20	0.18
262A	0.14	0.11	0.15	0.18	0.17	0.18	0.55	0.54	0.55
260I	0.22	0.19	0.23	0.18	0.19	0.16	0.43	0.43	0.43
252A	0.05	0.02	0.07	0.08	0.08	0.09	0.07	0.07	0.06
251Y	0.16	0.20	0.14	0.15	0.15	0.15	0.24	0.20	0.27
250P	0.21	0.21	0.22	0.21	0.21	0.22	0.20	0.20	0.20
248S	0.15	0.14	0.15	0.20	0.20	0.19	0.11	0.11	0.11
248P	0.16	0.17	0.16	0.23	0.23	0.23	0.21	0.22	0.21
244A	0.19	0.21	0.19	0.12	0.12	0.12	0.86	0.86	0.86
243I	0.12	0.09	0.13	0.08	0.07	0.09	0.21	0.21	0.22
236Y	0.27	0.27	0.28	0.21	0.21	0.22	0.64	0.64	0.64
234M	0.33	0.34	0.33	0.19	0.19	0.19	0.88	0.88	0.88
234C	0.28	0.27	0.28	0.14	0.14	0.14	0.77	0.77	0.78
221T	0.04	0.04	0.04	0.10	0.11	0.10	0.41	0.42	0.41
206C	0.15	0.15	0.14	0.13	0.11	0.15	0.10	0.11	0.10
201M	0.37	0.37	0.37	0.31	0.31	0.31	1.04	1.04	1.04
201L	0.16	0.16	0.15	0.17	0.17	0.18	0.85	0.85	0.86
201I	0.20	0.20	0.19	0.20	0.19	0.21	0.71	0.70	0.71
165C	0.27	0.26	0.27	0.39	0.38	0.39	0.96	0.93	0.98
148T	0.12	0.12	0.13	0.14	0.14	0.14	0.12	0.12	0.13
148S	0.08	0.07	0.09	0.10	0.10	0.10	0.07	0.07	0.08
148N	0.19	0.17	0.20	0.25	0.24	0.26	0.40	0.39	0.41
148K	0.10	0.08	0.12	0.11	0.12	0.10	0.26	0.23	0.30
136L	0.14	0.14	0.15	0.16	0.16	0.16	0.43	0.44	0.43
136F	0.06	0.06	0.06	0.09	0.08	0.09	0.66	0.65	0.67
10Q	0.17	0.18	0.17	0.19	0.17	0.20	0.13	0.12	0.13
10A	0.23	0.22	0.23	0.17	0.17	0.17	0.53	0.53	0.53

Table A 12: Beneficial mutations shared by only AmiE WT and AmiE I122L.

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
74N	0.12	0.11	0.12	0.06	0.06	0.06	-0.07	-0.05	-0.08
82V	0.07	0.07	0.07	0.06	0.06	0.06	-0.09	-0.10	-0.08
88A	0.10	0.09	0.10	0.35	0.35	0.35	-0.09	-0.08	-0.10
92K	0.08	0.07	0.09	0.08	0.07	0.08	-0.13	-0.13	-0.13
112N	0.04	0.03	0.05	0.05	0.05	0.05	-0.05	-0.05	-0.06
136C	0.09	0.09	0.09	0.04	0.04	0.05	-0.11	-0.16	-0.06
154D	0.05	0.04	0.05	0.05	0.06	0.03	-0.11	-0.11	-0.10
201C	0.13	0.14	0.13	0.07	0.06	0.09	-0.51	-0.58	-0.45
201R	0.08	0.09	0.07	0.09	0.08	0.10	-0.11	-0.10	-0.13
201T	0.20	0.20	0.20	0.13	0.12	0.14	-0.19	-0.19	-0.20
234Q	0.06	0.07	0.05	0.14	0.15	0.12	-0.28	-0.34	-0.23
248A	0.11	0.12	0.10	0.14	0.15	0.14	-0.12	-0.15	-0.08

Table A 13: Beneficial mutations shared by only AmiE WT and AmiE I38V.

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
10H	0.06	0.07	0.05	-0.26	-0.26	-0.26	0.04	0.06	0.03
9L	0.14	0.13	0.15	-0.14	-0.13	-0.16	0.24	0.25	0.23
3Q	0.31	0.31	0.32	-0.07	-0.07	-0.06	0.66	0.65	0.66
339S	0.17	0.16	0.18	-0.03	-0.04	-0.03	0.20	0.20	0.20
237S	0.08	0.04	0.09	-0.19	-0.19	-0.20	0.10	0.13	0.07

Table A 14: Beneficial mutations shared by only AmiE I38V and AmiE I122L.

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
4A	-0.10	-0.13	-0.08	0.08	0.08	0.08	0.55	0.54	0.55
79C	-0.11	-0.11	-0.12	0.15	0.14	0.15	0.73	0.72	0.73
69H	-0.03	-0.04	-0.02	0.04	0.03	0.05	0.03	0.04	0.03
69R	-0.05	-0.07	-0.03	0.05	0.05	0.05	0.11	0.11	0.11

Table A 15: AmiE WT unique beneficial mutations.

Location-mutation	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
5C	0.10	0.09	0.11	-0.20	-0.18	-0.21	-0.32	-0.32	-0.32
7G	0.19	0.15	0.21	-0.07	-0.02	-0.12	-0.12	-0.10	-0.14
47C	0.07	0.08	0.07	-0.44	-0.41	-0.46	-0.53	-0.51	-0.55
67C	0.08	-0.05	0.14	-0.08	-0.04	-0.14	-0.13	-0.12	-0.14
72T	0.08	0.09	0.08	-0.62	-0.60	-0.63	-0.60	-0.57	-0.63
74G	0.12	0.11	0.14	-0.17	-0.15	-0.19	-0.24	-0.25	-0.24
78H	0.11	0.10	0.12	-0.13	-0.13	-0.13	-0.54	-0.59	-0.49
82I	0.05	0.04	0.06	-0.04	-0.04	-0.03	-0.42	-0.45	-0.39
179V	0.11	-0.17	0.17	-0.05	-0.01	-0.10	-0.26	-0.26	-0.27
189Y	0.08	-0.09	0.12	-0.10	-0.05	-0.15	-0.31	-0.31	-0.30
228G	0.05	-0.10	0.09	-0.16	-0.11	-0.23	-0.20	-0.20	-0.21
234T	0.07	0.06	0.08	-0.08	-0.08	-0.09	-0.36	-0.35	-0.37
252D	0.07	-0.25	0.13	-0.21	-0.14	-0.32	-0.54	-0.47	-0.62
277C	0.11	-0.47	0.19	-0.49	-0.35	-0.72	-0.30	-0.29	-0.31
291C	0.07	0.10	0.05	-0.50	-0.36	-0.75	-0.40	-0.42	-0.39
297H	0.08	0.11	0.06	-0.32	-0.42	-0.26	-0.44	-0.46	-0.42
311C	0.16	0.20	0.14	-1.17	-1.11	-1.22	-1.07	-1.28	-0.94
320A	0.05	0.08	0.03	-0.39	-0.52	-0.32	-0.39	-0.44	-0.34
325C	0.14	0.13	0.14	-0.12	-0.13	-0.12	-0.22	-0.18	-0.27
325T	0.05	0.04	0.06	-0.32	-0.30	-0.33	-0.08	-0.06	-0.09
325Y	0.16	0.18	0.14	-0.10	-0.01	-0.23	-0.24	-0.22	-0.28
326H	0.22	0.23	0.22	-0.09	-0.10	-0.08	-0.21	-0.27	-0.16
326Y	0.07	0.09	0.06	-0.27	-0.28	-0.27	-0.51	-0.56	-0.47
336L	0.08	0.07	0.08	-0.04	-0.05	-0.04	-0.16	-0.17	-0.14
341Y	0.18	0.19	0.17	-0.25	-0.23	-0.27	-0.16	-0.14	-0.18

red highlighting indicates samples were dropped from analysis of beneficial mutations due to technical replicate disparities

Table A 16: AmiE I122L unique beneficial mutations.

Location-mutation	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric
47K	0.07	0.07	0.06	-0.07	-0.04	-0.08	-0.07	-0.08	-0.07
47R	0.03	0.03	0.02	-0.07	-0.07	-0.07	-0.13	-0.12	-0.14
58H	0.13	0.01	0.20	-0.24	-0.20	-0.27	-0.17	-0.14	-0.22
69Q	0.04	0.04	0.05	-0.04	-0.06	-0.03	-0.17	-0.16	-0.18
84I	0.17	0.17	0.16	-0.39	-0.44	-0.35	-0.85	-0.92	-0.79
84V	0.15	0.15	0.14	-0.14	-0.16	-0.13	-0.84	-0.83	-0.85
136V	0.05	0.06	0.04	-0.04	-0.04	-0.04	-0.04	-0.06	-0.03
336V	0.02	0.02	0.03	-0.04	-0.05	-0.04	-0.05	-0.03	-0.08

Table A 17: AmiE I38V unique beneficial mutations.

Location-mutation	AmiE I38V cumulative normalized fitness metric	AmiE I38V technical replicate 1 normalized fitness metric	AmiE I38V technical replicate 2 normalized fitness metric	AmiE WT cumulative normalized fitness metric	AmiE WT technical replicate 1 normalized fitness metric	AmiE WT technical replicate 2 normalized fitness metric	AmiE I122L cumulative normalized fitness metric	AmiE I122L technical replicate 1 normalized fitness metric	AmiE I122L technical replicate 2 normalized fitness metric
8V	0.30	0.30	0.31	-0.74	-0.83	-0.68	-0.36	-0.39	-0.33
21A	0.04	0.04	0.05	-0.19	-0.21	-0.18	-0.28	-0.29	-0.27
21P	0.38	0.38	0.38	-0.11	-0.14	-0.09	-0.10	-0.09	-0.11
21S	0.07	0.06	0.08	-0.19	-0.20	-0.19	-0.26	-0.26	-0.26
25C	0.14	0.15	0.14	-0.11	-0.12	-0.10	-0.04	-0.03	-0.04
65V	0.24	0.25	0.23	-0.28	-0.31	-0.27	-0.27	-0.25	-0.29
69Y	0.08	0.08	0.08	-0.12	-0.15	-0.10	-0.27	-0.30	-0.24
72L	0.04	0.03	0.04	-0.35	-0.32	-0.37	-0.43	-0.41	-0.45
75C	0.12	0.13	0.10	-0.39	-0.45	-0.35	-0.19	-0.19	-0.18
75L	0.12	0.11	0.13	-0.27	-0.27	-0.26	-0.23	-0.23	-0.22
102C	0.11	0.09	0.13	-0.24	-0.22	-0.25	-0.41	-0.42	-0.40
102I	0.54	0.53	0.54	-0.20	-0.20	-0.20	-0.07	-0.07	-0.07
143R	0.05	0.01	0.08	-0.19	-0.19	-0.20	-0.13	-0.13	-0.13
143T	0.30	0.29	0.30	-0.09	-0.11	-0.09	-0.04	-0.04	-0.04
149M	0.40	0.39	0.41	-0.25	-0.27	-0.24	-0.25	-0.23	-0.29
149R	0.06	0.04	0.08	-0.16	-0.16	-0.16	-0.36	-0.32	-0.42
190T	0.37	0.38	0.36	-0.49	-0.43	-0.53	-0.37	-0.42	-0.33
193P	0.53	0.54	0.53	-0.47	-0.48	-0.46	-0.08	-0.08	-0.07
197C	0.30	0.34	0.24	-1.20	-1.10	-1.26	-0.81	-0.95	-0.73
201N	0.25	0.27	0.24	-0.11	-0.11	-0.11	-0.07	-0.06	-0.08
204N	0.26	0.26	0.26	-0.28	-0.35	-0.25	-0.16	-0.13	-0.19
212C	0.51	0.50	0.53	-0.37	-0.35	-0.38	-0.33	-0.34	-0.32
212K	0.07	0.09	0.04	-0.33	-0.56	-0.27	-0.60	-0.46	-0.84
212T	0.40	0.40	0.40	-0.25	-0.10	-0.37	-0.24	-0.23	-0.25
212V	0.19	0.20	0.19	-0.48	-0.41	-0.52	-0.37	-0.37	-0.38
228A	0.15	0.13	0.16	-0.32	-0.31	-0.32	-0.25	-0.23	-0.28
230C	0.25	0.25	0.26	-0.18	-0.21	-0.16	-0.31	-0.31	-0.31
251N	0.47	0.47	0.47	-0.10	-0.13	-0.09	-0.15	-0.16	-0.15
253Y	0.33	0.33	0.32	-0.49	-0.63	-0.44	-0.39	-0.40	-0.37
254T	0.15	0.16	0.14	-0.42	-0.43	-0.41	-0.43	-0.43	-0.44
287T	0.12	0.13	0.10	-0.17	-0.17	-0.17	-0.21	-0.21	-0.20
288K	0.07	0.08	0.06	-0.07	-0.07	-0.07	-0.04	-0.04	-0.03
317M	0.17	0.18	0.16	-0.41	-0.34	-0.46	-0.23	-0.17	-0.29
325F	0.15	0.19	0.10	-0.16	-0.16	-0.16	-0.15	-0.09	-0.22
326R	0.06	0.04	0.08	-0.24	-0.22	-0.26	-0.23	-0.23	-0.24
333N	0.25	0.22	0.28	-0.30	-0.31	-0.29	-0.20	-0.21	-0.20
340K	0.15	0.16	0.13	-0.12	-0.11	-0.13	-0.05	-0.04	-0.05

Table A 18: AmiE WT unique beneficial mutations with synonymous codon fitness disparities.

Position	Mutation	Codon	Codon Frequency	Reference population count	Cumulative selected population count	Selected population count - technical replicate 1	Selected population count - technical replicate 2	Cumulative normalized fitness metric	Technical replicate 1 normalized fitness metric	Technical replicate 2 normalized fitness metric	Variance for synonymous codons of normalized fitness metrics - cumulative	Variance for synonymous codons of normalized fitness metrics - technical replicate 1	Variance for synonymous codons of normalized fitness metrics - technical replicate 2	Shared-unique beneficial mutation bin
3	L	CTA	0.04	109	2381	1050	1331	0.29	0.29	0.28	0.08	0.09	0.05	Beneficial for all - 95% CI cutoff
3	L	CTC	0.1	0	2	1	1	None	None	None	0.08	0.09	0.05	
3	L	CTG	0.47	1	2	0	2	-0.27	None	-0.12	0.08	0.09	0.05	
3	L	CTT	0.12	29	109	43	66	-0.10	-0.12	-0.09	0.08	0.09	0.05	
3	L	TTA	0.14	0	6	1	5	None	None	None	0.08	0.09	0.05	
3	L	TTG	0.13	6	0	0	0	None	None	None	0.08	0.09	0.05	
3	Q	CAA	0.34	144	4028	1674	2354	0.33	0.33	0.34	0.52	0.32	1.06	WT and 138V shared beneficial - >10% increase in growth rate cutoff
3	Q	CAG	0.66	16	9	7	2	-0.69	-0.47	-1.12	0.52	0.32	1.06	
3	V	GTA	0.17	22	908	368	540	0.40	0.39	0.41	0.12	0.06	0.34	Beneficial for all ->10% increase in growth rate cutoff
3	V	GTC	0.2	9	144	65	79	0.22	0.23	0.22	0.12	0.06	0.34	
3	V	GTG	0.35	3	6	5	1	-0.27	-0.09	-0.68	0.12	0.06	0.34	
3	V	GTT	0.28	6	0	0	0	None	None	None	0.12	0.06	0.34	
5	C	TGC	0.54	77	799	332	467	0.13	0.13	0.14	0.05	0.06	0.05	WT unique beneficial - 95% CI cutoff
5	C	TGT	0.46	18	48	19	29	-0.19	-0.21	-0.18	0.05	0.06	0.05	
7	G	GGA	0.13	4	3	1	2	-0.58	-0.67	-0.54	0.24	0.37	0.21	Unique WT beneficial - >10% increase in growth rate cutoff
7	G	GGC	0.37	121	1799	639	1160	0.21	0.17	0.23	0.24	0.37	0.21	
7	G	GGG	0.15	9	5	1	4	-0.70	-0.67	-0.58	0.24	0.37	0.21	
7	G	GGT	0.35	0	1	0	1	None	None	None	0.24	None	0.21	
8	A	GCA	0.23	1	11	7	4	0.15	0.23	0.05	0.42	0.39	0.46	Beneficial for all - 95% CI cutoff
8	A	GCC	0.26	10	2	1	1	-0.68	-0.67	-0.68	0.42	0.39	0.46	
8	A	GCG	0.33	41	434	203	231	0.14	0.16	0.12	0.42	0.39	0.46	
8	A	GCT	0.18	7	86	37	49	0.17	0.17	0.17	0.42	0.39	0.46	
8	K	AAA	0.74	2	5	0	5	-0.21	None	-0.07	0.13	None	0.07	Beneficial for all ->10% increase in growth rate cutoff
8	K	AAG	0.26	36	877	376	501	0.31	0.31	0.31	0.13	None	0.07	
9	L	CTA	0.04	6	4	1	3	-0.63	-0.83	-0.54	0.76	0.28	0.54	WT and 138V shared beneficial - >10% increase in growth rate cutoff
9	L	CTC	0.1	4	2	1	1	-0.74	-0.67	-0.80	0.76	0.28	0.54	
9	L	CTG	0.47	47	3	0	3	-2.02	None	-2.51	0.76	0.28	0.54	
9	L	TTA	0.14	66	1317	553	764	0.27	0.26	0.27	0.76	0.28	0.54	
9	L	TTG	0.13	7	69	20	49	0.12	0.04	0.17	0.76	0.28	0.54	
10	A	GCA	0.23	97	3524	1496	2028	0.38	0.38	0.38	0.12	0.12	0.12	Beneficial for all ->10% increase in growth rate cutoff
10	A	GCC	0.26	160	1083	463	620	0.04	0.04	0.04	0.12	0.12	0.12	
10	A	GCG	0.33	9	345	151	194	0.39	0.39	0.39	0.12	0.12	0.12	
10	A	GCT	0.18	46	73	30	43	-0.34	-0.35	-0.34	0.12	0.12	0.12	
10	H	CAC	0.43	80	1246	509	737	0.22	0.21	0.23	0.03	0.03	0.04	WT and 138V shared beneficial - 95% CI cutoff
10	H	CAT	0.57	250	1206	566	640	-0.04	-0.02	-0.06	0.03	0.03	0.04	
47	C	TGC	0.54	234	2225	1029	1296	0.12	0.13	0.12	0.11	0.14	0.09	WT unique beneficial - 95% CI cutoff
47	C	TGT	0.46	80	128	44	84	-0.34	-0.40	-0.30	0.11	0.14	0.09	
52	L	CTA	0.04	66	905	363	542	0.19	0.18	0.20	0.00	0.00	0.00	Beneficial for all - 95% CI cutoff
52	L	CTC	0.1	85	785	333	452	0.11	0.11	0.11	0.00	0.00	0.00	
52	L	CTG	0.47	73	621	258	363	0.09	0.08	0.09	0.00	0.00	0.00	
52	L	CTT	0.12	26	194	72	122	0.06	0.03	0.08	0.00	0.00	0.00	
52	L	TTA	0.14	9	67	30	37	0.06	0.07	0.05	0.00	0.00	0.00	
52	L	TTG	0.13	87	491	199	292	0.00	-0.02	0.00	0.00	0.00	0.00	
62	T	ACA	0.17	23	124	58	66	-0.01	0.01	-0.03	0.01	0.00	0.01	Beneficial for all - 95% CI cutoff
62	T	ACC	0.4	33	401	171	230	0.17	0.17	0.17	0.01	0.00	0.01	
62	T	ACG	0.25	24	187	84	103	0.07	0.08	0.06	0.01	0.00	0.01	
62	T	ACT	0.19	10	67	28	39	0.04	0.03	0.04	0.01	0.00	0.01	
67	C	TGC	0.54	46	427	103	324	0.11	-0.02	0.17	1.05	None	0.74	WT unique beneficial - 95% CI cutoff
67	C	TGT	0.46	7	1	0	1	-2.12	None	-2.12	1.05	None	0.74	
72	T	ACA	0.17	10	1	1	0	-2.57	-2.57	None	0.62	0.31	0.42	WT unique beneficial - 95% CI cutoff
72	T	ACC	0.4	56	692	299	393	0.17	0.17	0.17	0.62	0.31	0.42	
72	T	ACG	0.25	27	122	55	67	-0.06	-0.04	-0.07	0.62	0.31	0.42	
72	T	ACT	0.19	7	3	2	1	-2.01	-2.62	-2.01	0.62	0.31	0.42	
74	G	GGA	0.13	132	655	256	399	-0.03	-0.05	-0.02	0.03	0.03	0.03	WT unique beneficial - 95% CI cutoff
74	G	GGC	0.37	23	86	35	51	-0.10	-0.12	-0.10	0.03	0.03	0.03	
74	G	GGG	0.15	58	1045	448	597	0.25	0.25	0.25	0.03	0.03	0.03	
74	G	GGT	0.35	56	890	324	566	0.22	0.19	0.24	0.03	0.03	0.03	
78	H	CAC	0.43	51	258	101	157	-0.03	-0.05	-0.02	0.03	0.04	0.03	WT unique beneficial - 95% CI cutoff
78	H	CAT	0.57	32	511	214	297	0.22	0.22	0.23	0.03	0.04	0.03	
79	V	GTA	0.17	66	333	152	181	-0.03	-0.01	-0.04	0.00	0.00	0.00	Beneficial for all - 95% CI cutoff
79	V	GTC	0.2	41	273	119	154	0.03	0.04	0.03	0.00	0.00	0.00	
79	V	GTG	0.35	36	348	159	189	0.12	0.13	0.11	0.00	0.00	0.00	
79	V	GTT	0.28	54	386	163	223	0.05	0.05	0.05	0.00	0.00	0.00	
82	I	ATA	0.11	61	363	141	222	0.01	-0.01	0.02	0.00	0.00	0.00	WT unique beneficial - 95% CI cutoff
82	I	ATC	0.39	153	1279	534	745	0.09	0.08	0.09	0.00	0.00	0.00	
82	I	ATT	0.49	62	337	125	212	-0.01	-0.05	0.01	0.00	0.00	0.00	
93	A	GCA	0.23	68	1242	534	708	0.07	0.07	0.08	0.00	0.01	0.00	Beneficial for all - 95% CI cutoff
93	A	GCC	0.26	55	651	266	385	-0.03	-0.05	-0.01	0.00	0.01	0.00	
93	A	GCG	0.33	43	991	420	571	0.12	0.12	0.13	0.00	0.01	0.00	
93	A	GCT	0.18	68	1519	655	864	0.12	0.11	0.12	0.00	0.01	0.00	
179	V	GTA	0.17	14	2	1	1	-2.12	-2.12	-2.12	0.60	0.28	0.65	WT unique beneficial - 95% CI cutoff
179	V	GTC	0.2	226	3082	289	2793	0.16	-0.11	0.21	0.60	0.28	0.65	
179	V	GTG	0.35	27	11	4	7	-0.88	-0.81	-0.92	0.60	0.28	0.65	
179	V	GTT	0.28	16	2	0	2	-2.12	None	-2.26	0.60	0.28	0.65	

Position	Mutation	Codon	Codon Frequency	Reference population count	Cumulative selected population count	Selected population count - technical replicate 1	Selected population count - technical replicate 2	Cumulative normalized fitness metric	Technical replicate 1 normalized fitness metric	Technical replicate 2 normalized fitness metric	Variance for synonymous codons of normalized fitness metrics - cumulative	Variance for synonymous codons of normalized fitness metrics - technical replicate 1	Variance for synonymous codons of normalized fitness metrics - technical replicate 2	Shared-unique beneficial mutation bin
189	Y	TAC	0.41	110	1511	624	1287	0.18	0.00	0.20	11.16	None	6.06	WT unique beneficial - 95% CI cutoff
189	Y	TAT	0.59	52	1	0	1	-2.26	None	-2.26	11.16	None	6.06	
201	R	AGA	0.07	56	498	162	336	0.07	0.08	0.06	0.01	0.01	0.01	WT and 1122L shared beneficial - 95% CI cutoff
201	R	AGG	0.04	0	9	4	5	None	None	None	0.01	0.01	0.01	
201	R	CGA	0.07	103	1037	335	702	0.09	0.11	0.09	0.01	0.01	0.01	
201	R	CGC	0.36	20	89	25	64	-0.10	-0.12	-0.09	0.01	0.01	0.01	
201	R	CGG	0.11	17	180	67	113	0.11	0.15	0.08	0.01	0.01	0.01	
201	R	CGT	0.36	2	20	4	16	0.09	0.00	0.12	0.01	0.01	0.01	
201	T	ACA	0.17	189	3350	1018	2332	0.21	0.21	0.21	0.01	0.01	0.01	WT and 1122L shared beneficial - 95% CI cutoff
201	T	ACC	0.4	0	64	20	44	None	None	None	0.01	0.01	0.01	
201	T	ACG	0.25	10	64	19	45	-0.01	-0.02	-0.01	0.01	0.01	0.01	
201	T	ACT	0.19	34	354	106	248	0.10	0.10	0.10	0.01	0.01	0.01	
221	T	ACA	0.17	110	816	222	594	0.02	0.00	0.04	0.00	0.00	0.01	Beneficial for all - 95% CI cutoff
221	T	ACC	0.4	19	187	54	133	0.09	0.08	0.09	0.00	0.00	0.01	
221	T	ACG	0.25	28	162	64	98	-0.03	0.03	-0.07	0.00	0.00	0.01	
221	T	ACT	0.19	58	557	177	380	0.08	0.09	0.08	0.00	0.00	0.01	
234	Q	CAA	0.34	49	325	97	228	0.00	-0.01	0.00	0.01	0.01	0.01	WT and 1122L shared beneficial - 95% CI cutoff
234	Q	CAG	0.66	33	372	124	248	0.12	0.14	0.11	0.01	0.01	0.01	
234	T	ACA	0.17	40	147	34	113	-0.15	-0.22	-0.12	0.02	0.04	0.02	WT unique beneficial - 95% CI cutoff
234	T	ACC	0.4	295	3407	1004	2403	0.12	0.12	0.13	0.02	0.04	0.02	
234	T	ACG	0.25	34	91	20	71	-0.23	-0.33	-0.20	0.02	0.04	0.02	
234	T	ACT	0.19	47	164	47	117	-0.16	-0.18	-0.16	0.02	0.04	0.02	
237	S	AGC	0.25	20	19	7	12	-0.55	-0.49	-0.59	0.11	0.09	0.12	WT and 138V shared beneficial - 95% CI cutoff
237	S	AGT	0.16	153	401	120	281	-0.24	-0.24	-0.24	0.11	0.09	0.12	
237	S	TCA	0.14	21	31	10	21	-0.41	-0.39	-0.42	0.11	0.09	0.12	
237	S	TCC	0.15	109	282	79	203	-0.24	-0.27	-0.23	0.11	0.09	0.12	
237	S	TCG	0.14	38	87	29	58	-0.28	-0.25	-0.29	0.11	0.09	0.12	
237	S	TCT	0.17	52	2831	680	2151	0.43	0.39	0.44	0.11	0.09	0.12	
251	Y	TAC	0.41	22	493	186	307	0.26	0.30	0.24	0.04	0.06	0.04	Beneficial for all ->10% increase in growth rate cutoff
251	Y	TAT	0.59	25	147	44	103	-0.03	-0.03	-0.03	0.04	0.06	0.04	
252	A	GCA	0.23	31	191	47	144	-0.02	-0.07	0.00	0.01	0.01	0.01	Beneficial for all - 95% CI cutoff
252	A	GCC	0.26	41	380	91	289	0.08	0.02	0.10	0.01	0.01	0.01	
252	A	GCG	0.33	11	64	21	43	-0.03	-0.01	-0.04	0.01	0.01	0.01	
252	A	GCT	0.18	18	219	62	157	0.14	0.12	0.14	0.01	0.01	0.01	
277	C	TGC	0.54	115	5	0	5	3.49	None	3.49	3.49	None	2.60	WT unique beneficial - 95% CI cutoff
277	C	TGT	0.46	53	1441	60	1381	0.34	-0.13	0.41	3.49	None	2.60	
291	C	TGC	0.54	88	410	142	268	-0.04	-0.04	-0.03	0.01	0.02	0.01	WT unique beneficial - 95% CI cutoff
291	C	TGT	0.46	161	1426	600	826	0.11	0.15	0.09	0.01	0.02	0.01	
297	H	CAC	0.43	117	1272	522	750	0.16	0.19	0.14	0.07	0.09	0.06	WT unique beneficial - 95% CI cutoff
297	H	CAT	0.57	69	159	54	105	-0.22	-0.23	-0.21	0.07	0.09	0.06	
311	C	TGC	0.54	70	1328	576	752	0.27	0.31	0.25	1.27	2.68	1.00	Unique WT beneficial - >10% increase in growth rate cutoff
311	C	TGT	0.46	49	7	1	6	-2.57	None	-2.57	1.27	2.68	1.00	
320	A	GCA	0.23	57	99	35	64	-0.30	-0.30	-0.30	0.30	0.13	0.22	WT unique beneficial - 95% CI cutoff
320	A	GCC	0.26	74	52	24	28	-0.59	-0.50	-0.66	0.30	0.13	0.22	
320	A	GCG	0.33	15	3	0	3	-2.57	None	-2.92	0.30	0.13	0.22	
320	A	GCT	0.18	168	1967	794	1173	0.17	0.20	0.16	0.30	0.13	0.22	
325	T	ACA	0.17	18	302	109	193	0.25	0.25	0.25	0.01	0.02	0.01	WT unique beneficial - 95% CI cutoff
325	T	ACC	0.4	121	544	181	363	-0.05	-0.06	-0.04	0.01	0.02	0.01	
325	T	ACG	0.25	10	93	24	69	0.12	0.05	0.15	0.01	0.02	0.01	
325	T	ACT	0.19	41	356	120	236	0.11	0.10	0.11	0.01	0.02	0.01	
326	Y	TAC	0.41	147	1155	450	705	0.09	0.11	0.07	0.10	0.08	0.11	WT unique beneficial - 95% CI cutoff
326	Y	TAT	0.59	11	16	7	9	-0.35	-0.29	-0.40	0.10	0.08	0.11	
330	S	AGC	0.25	82	429	145	284	-0.01	-0.02	0.00	0.00	0.00	0.00	Beneficial for all - 95% CI cutoff
330	S	AGT	0.16	79	519	161	358	0.05	0.01	0.06	0.00	0.00	0.00	
330	S	TCA	0.14	170	947	338	609	0.01	0.01	0.01	0.00	0.00	0.00	
330	S	TCC	0.15	251	1366	589	777	0.00	0.05	-0.03	0.00	0.00	0.00	
330	S	TCG	0.14	105	792	268	524	0.08	0.07	0.08	0.00	0.00	0.00	
330	S	TCT	0.17	52	367	127	240	0.06	0.06	0.06	0.00	0.00	0.00	
336	L	CTA	0.04	62	238	86	152	-0.08	-0.08	-0.09	0.02	0.02	0.02	WT unique beneficial - 95% CI cutoff
336	L	CTC	0.1	100	420	157	263	-0.06	-0.05	-0.07	0.02	0.02	0.02	
336	L	CTG	0.47	57	201	49	152	-0.11	-0.21	-0.07	0.02	0.02	0.02	
336	L	CTT	0.12	148	2191	759	1432	0.22	0.22	0.22	0.02	0.02	0.02	
336	L	TTA	0.14	31	153	47	106	-0.02	-0.06	-0.01	0.02	0.02	0.02	
336	L	TTG	0.13	40	152	62	90	-0.09	-0.05	-0.11	0.02	0.02	0.02	
337	E	GAA	0.68	3	15	4	11	-0.02	-0.09	0.01	0.05	0.07	0.04	Beneficial for all - 95% CI cutoff
337	E	GAG	0.32	17	370	126	244	0.30	0.29	0.30	0.05	0.07	0.04	
341	Y	TAC	0.41	16	272	104	168	0.25	0.26	0.24	0.12	0.13	0.12	Unique WT beneficial - >10% increase in growth rate cutoff
341	Y	TAT	0.59	8	17	6	11	-0.24	-0.24	-0.24	0.12	0.13	0.12	

Table A 19: AmiE I122L unique beneficial mutations with synonymous codon fitness disparities.

Position	Mutation	Codon	Codon Frequency	Reference population count	Cumulative selected population count	Selected population count - technical	Selected population count - technical	Cumulative normalized fitness metric	Technical replicate 1 normalized fitness metric	Technical replicate 2 normalized fitness metric	Variance for synonymous codons of normalized fitness metrics - cumulative	Variance for synonymous codons of normalized fitness metrics - technical replicate 1	Variance for synonymous codons of normalized fitness metrics - technical replicate 2	Shared-unique beneficial mutation bin
3	I	ATA	0.11	14	117	57	60	0.19	0.19	0.18	0.40	0.37	0.43	Beneficial for all - >10% increase in growth rate cutoff
3	I	ATC	0.39	5	160	64	96	0.43	0.40	0.46	0.40	0.37	0.43	
3	I	ATT	0.49	8	2	1	1	-0.76	-0.74	-0.79	0.40	0.37	0.43	
3	L	CTA	0.04	29	438	214	224	0.30	0.30	0.30	0.27	0.14	0.19	Beneficial for all - 95% CI cutoff
3	L	CTC	0.1	7	5	3	2	-0.41	-0.34	-0.49	0.27	0.14	0.19	
3	L	CTG	0.47	5	1	0	1	-0.93	None	-0.93	0.27	0.14	0.19	
3	L	CTT	0.12	24	93	37	56	0.03	-0.01	0.06	0.27	0.14	0.19	
3	L	TTA	0.14	10	168	86	82	0.32	0.33	0.31	0.27	0.14	0.19	
3	L	TTG	0.13	4	1	1	0	-0.76	-0.76	None	0.27	0.14	0.19	
3	V	GTA	0.17	16	105	53	52	0.14	0.15	0.13	0.19	0.12	0.11	Beneficial for all - >10% increase in growth rate cutoff
3	V	GTC	0.2	1	55	25	30	0.52	0.51	0.53	0.19	0.12	0.11	
3	V	GTG	0.35	2	4	2	2	-0.13	-0.12	-0.14	0.19	0.12	0.11	
3	V	GTT	0.28	4	2	2	0	-0.30	-0.30	None	0.19	0.12	0.11	
4	A	GCA	0.23	28	141	57	84	0.08	0.05	0.11	0.07	0.07	0.08	I122L and I38V shared beneficial - >95% CI cutoff
4	A	GCC	0.26	45	395	197	198	0.20	0.20	0.19	0.07	0.07	0.08	
4	A	GCG	0.33	19	58	34	24	-0.03	0.02	-0.08	0.07	0.07	0.08	
4	A	GCT	0.18	32	22	11	11	-0.42	-0.40	-0.44	0.07	0.07	0.08	
4	S	AGC	0.25	61	335	171	164	0.10	0.12	0.09	0.01	0.01	0.03	Beneficial for all - 95% CI cutoff
4	S	AGT	0.16	17	40	26	14	-0.09	-0.02	-0.19	0.01	0.01	0.03	
4	S	TCA	0.14	14	25	17	8	-0.16	-0.07	-0.29	0.01	0.01	0.03	
4	S	TCC	0.15	15	69	33	36	0.06	0.06	0.06	0.01	0.01	0.03	
4	S	TCG	0.14	7	48	29	19	0.15	0.19	0.09	0.01	0.01	0.03	
4	S	TCT	0.17	5	23	11	12	0.06	0.06	0.06	0.01	0.01	0.03	
8	A	GCA	0.23	17	82	33	49	0.07	0.04	0.10	0.23	0.17	0.32	Beneficial for all - 95% CI cutoff
8	A	GCC	0.26	38	8	5	3	-0.83	-0.82	-0.85	0.23	0.17	0.32	
8	A	GCG	0.33	154	691	318	373	0.06	0.05	0.07	0.23	0.17	0.32	
8	A	GCT	0.18	20	214	90	124	0.24	0.21	0.26	0.23	0.17	0.32	
9	Q	CAA	0.34	16	216	96	120	0.28	0.27	0.29	0.11	0.08	0.14	Beneficial for all - 95% CI cutoff
9	Q	CAG	0.66	28	45	26	19	-0.18	-0.13	-0.24	0.11	0.08	0.14	
10	A	GCA	0.23	122	1343	641	702	0.24	0.24	0.24	0.12	0.12	0.11	Beneficial for all - >10% increase in growth rate cutoff
10	A	GCC	0.26	119	584	269	315	0.08	0.07	0.08	0.12	0.12	0.11	
10	A	GCG	0.33	99	1582	757	825	0.31	0.31	0.31	0.12	0.12	0.11	
10	A	GCT	0.18	122	76	34	42	-0.53	-0.52	-0.54	0.12	0.12	0.11	
47	R	AGA	0.07	154	423	215	208	-0.05	-0.04	-0.07	0.00	0.00	0.00	I122L unique beneficial - 95% CI cutoff
47	R	AGG	0.04	108	462	224	238	0.05	0.05	0.05	0.00	0.00	0.00	
47	R	CGA	0.07	314	1132	581	551	0.01	0.03	-0.01	0.00	0.00	0.00	
47	R	CGC	0.36	390	1676	835	841	0.05	0.06	0.04	0.00	0.00	0.00	
47	R	CGG	0.11	149	607	284	323	0.04	0.03	0.04	0.00	0.00	0.00	
47	R	CGT	0.36	230	883	442	441	0.02	0.03	0.01	0.00	0.00	0.00	
69	R	AGA	0.07	102	327	145	182	-0.02	-0.03	0.00	0.00	0.00	0.00	I122L and I38V shared beneficial - 95% CI cutoff
69	R	AGG	0.04	51	300	135	165	0.12	0.10	0.13	0.00	0.00	0.00	
69	R	CGA	0.07	156	589	271	318	0.02	0.01	0.03	0.00	0.00	0.00	
69	R	CGC	0.36	203	899	445	454	0.06	0.06	0.05	0.00	0.00	0.00	
69	R	CGG	0.11	124	537	243	294	0.05	0.04	0.06	0.00	0.00	0.00	
69	R	CGT	0.36	159	780	371	409	0.08	0.08	0.08	0.00	0.00	0.00	
89	S	AGC	0.25	58	424	240	184	0.15	0.15	0.14	0.01	0.01	0.01	Beneficial for all - 95% CI cutoff
89	S	AGT	0.16	52	178	109	69	-0.02	0.01	-0.05	0.01	0.01	0.01	
89	S	TCA	0.14	27	90	54	36	-0.02	0.00	-0.05	0.01	0.01	0.01	
89	S	TCC	0.15	23	95	54	41	0.03	0.04	0.01	0.01	0.01	0.01	
89	S	TCG	0.14	12	97	70	27	0.17	0.22	0.06	0.01	0.01	0.01	
89	S	TCT	0.17	26	88	56	32	-0.02	0.02	-0.07	0.01	0.01	0.01	
93	A	GCA	0.23	128	672	371	301	0.08	0.08	0.07	0.00	0.00	0.00	Beneficial for all - 95% CI cutoff
93	A	GCC	0.26	128	393	215	178	-0.04	-0.04	-0.04	0.00	0.00	0.00	
93	A	GCG	0.33	86	410	225	185	0.06	0.06	0.05	0.00	0.00	0.00	
93	A	GCT	0.18	76	259	157	102	-0.02	0.01	-0.05	0.00	0.00	0.00	
136	V	GTA	0.17	122	750	395	355	0.11	0.10	0.12	0.00	0.00	0.00	I122L unique beneficial - 95% CI cutoff
136	V	GTC	0.2	169	600	339	261	-0.01	0.00	-0.02	0.00	0.00	0.00	
136	V	GTG	0.35	95	378	222	156	0.02	0.04	-0.01	0.00	0.00	0.00	
136	V	GTT	0.28	106	535	321	214	0.07	0.09	0.04	0.00	0.00	0.00	
148	S	AGC	0.25	69	397	214	183	0.10	0.10	0.10	0.00	0.00	0.00	Beneficial for all - 95% CI cutoff
148	S	AGT	0.16	23	133	74	59	0.10	0.10	0.09	0.00	0.00	0.00	
148	S	TCA	0.14	25	169	82	87	0.13	0.11	0.15	0.00	0.00	0.00	
148	S	TCC	0.15	53	373	219	154	0.14	0.15	0.12	0.00	0.00	0.00	
148	S	TCG	0.14	38	212	118	94	0.09	0.10	0.08	0.00	0.00	0.00	
148	S	TCT	0.17	31	101	52	49	-0.03	-0.04	-0.01	0.00	0.00	0.00	
154	D	GAC	0.37	49	288	157	131	0.10	0.10	0.10	0.01	0.00	0.01	I122L and WT shared beneficial - 95% CI cutoff
154	D	GAT	0.63	55	189	118	71	-0.01	0.02	-0.06	0.01	0.00	0.01	
244	A	GCA	0.23	30	215	102	113	0.20	0.19	0.21	0.01	0.01	0.02	Beneficial for all - 95% CI cutoff
244	A	GCC	0.26	70	495	253	242	0.19	0.20	0.19	0.01	0.01	0.02	
244	A	GCG	0.33	10	65	26	39	0.18	0.13	0.21	0.01	0.01	0.02	
244	A	GCT	0.18	94	204	110	94	-0.05	-0.03	-0.08	0.01	0.01	0.02	
336	V	GTA	0.17	233	360	172	188	0.05	0.06	0.04	0.00	0.00	0.00	I122L unique beneficial - 95% CI cutoff
336	V	GTC	0.2	287	341	149	192	-0.01	-0.02	0.00	0.00	0.00	0.00	
336	V	GTG	0.35	233	354	158	196	0.05	0.04	0.05	0.00	0.00	0.00	
336	V	GTT	0.28	190	254	104	150	0.02	-0.01	0.04	0.00	0.00	0.00	

Table A 20: AmiE I38V unique beneficial mutations with synonymous codon fitness disparities.

Position	Mutation	Codon	Codon Frequency	Reference population count	Cumulative selected population count	Selected population count - technical replicate 1	Selected population count - technical replicate 2	Cumulative normalized fitness metric	Technical replicate 1 normalized fitness metric	Technical replicate 2 normalized fitness metric	Variance for synonymous codons of normalized fitness metrics - cumulative	Variance for synonymous codons of normalized fitness metrics - technical replicate 1	Variance for synonymous codons of normalized fitness metrics - technical replicate 2	Shared-unique beneficial mutation bin
3	I	ATA	0.11	23	1937	1034	903	0.64	0.65	0.64	0.13	0.11	0.18	Beneficial for all - >10% increase in growth rate cutoff
3	I	ATC	0.39	3	123	61	62	0.53	0.52	0.54	0.13	0.11	0.18	
3	I	ATT	0.49	5	12	8	4	-0.03	0.03	-0.14	0.13	0.11	0.18	
3	L	CTA	0.04	24	210	99	111	0.25	0.23	0.26	0.34	0.22	0.11	Beneficial for all - 95% CI cutoff
3	L	CTC	0.1	12	5	2	3	-0.54	-0.61	-0.48	0.34	0.22	0.11	
3	L	CTG	0.47	11	1	1	0	-2.10	-0.65	None	0.34	0.22	0.11	
3	L	CTT	0.12	43	75	30	45	-0.11	-0.17	-0.07	0.34	0.22	0.11	
3	L	TTA	0.14	13	65	31	34	0.13	0.12	0.15	0.34	0.22	0.11	
4	A	GCA	0.23	15	1150	594	556	0.63	0.63	0.63	0.19	0.28	0.16	I38V and I122L shared beneficial - 95% CI cutoff
4	A	GCC	0.26	38	2513	1277	1236	0.61	0.60	0.61	0.19	0.28	0.16	
4	A	GCG	0.33	16	44	22	22	0.00	0.00	0.00	0.19	0.28	0.16	
4	A	GCT	0.18	12	13	3	10	-0.24	-0.47	-0.13	0.19	0.28	0.16	
8	A	GCA	0.23	17	927	480	447	0.57	0.57	0.58	0.31	0.29	0.33	Beneficial for all - 95% CI cutoff
8	A	GCC	0.26	20	11	6	5	-0.45	-0.42	-0.48	0.31	0.29	0.33	
8	A	GCG	0.33	69	2591	1350	1241	0.51	0.52	0.51	0.31	0.29	0.33	
8	A	GCT	0.18	7	1830	915	915	0.81	0.80	0.81	0.31	0.29	0.33	
8	V	GTA	0.17	4	456	229	227	0.69	0.68	0.69	0.92	0.74	0.62	I38V unique beneficial - >10% increase in growth rate cutoff
8	V	GTC	0.2	16	1	0	1	-1.41	None	-1.04	0.92	0.74	0.62	
8	V	GTG	0.35	26	10	2	8	-0.27	-1.04	-0.11	0.92	0.74	0.62	
8	V	GTT	0.28	7	143	75	68	0.41	0.41	0.41	0.92	0.74	0.62	
9	L	CTA	0.04	15	3	0	3	-0.83	None	-0.56	0.44	0.47	0.44	I38V and WT shared beneficial - >10% increase in growth rate cutoff
9	L	CTC	0.1	34	4	2	2	-1.04	-1.04	-1.04	0.44	0.47	0.44	
9	L	CTG	0.47	16	15	4	11	-0.28	-0.47	-0.18	0.44	0.47	0.44	
9	L	CTT	0.12	11	4	3	1	-0.59	-0.45	-0.97	0.44	0.47	0.44	
9	L	TTA	0.14	67	982	505	477	0.35	0.35	0.35	0.44	0.47	0.44	
9	L	TTG	0.13	3	231	144	87	0.63	0.66	0.59	0.44	0.47	0.44	
9	Q	CAA	0.34	15	890	489	401	0.59	0.60	0.58	0.28	0.33	0.24	Beneficial for all - 95% CI cutoff
9	Q	CAG	0.66	20	29	12	17	-0.16	-0.21	-0.12	0.28	0.33	0.24	
10	A	GCA	0.23	57	8250	4222	4028	0.72	0.72	0.73	0.29	0.28	0.29	Beneficial for all - >10% increase in growth rate cutoff
10	A	GCC	0.26	85	233	113	120	0.00	-0.01	0.01	0.29	0.28	0.29	
10	A	GCG	0.33	31	2123	1059	1064	0.67	0.67	0.67	0.29	0.28	0.29	
10	A	GCT	0.18	87	53	26	27	-0.41	-0.42	-0.41	0.29	0.28	0.29	
10	H	CAC	0.43	99	491	243	248	0.13	0.13	0.14	0.03	0.02	0.05	I38V and WT shared beneficial - 95% CI cutoff
10	H	CAT	0.57	103	185	117	68	-0.11	-0.05	-0.19	0.03	0.02	0.05	
21	A	GCA	0.23	69	273	136	137	0.08	0.08	0.09	0.01	0.01	0.01	I38V unique beneficial - 95% CI cutoff
21	A	GCC	0.26	86	158	79	79	-0.10	-0.10	-0.10	0.01	0.01	0.01	
21	A	GCG	0.33	72	255	132	123	0.06	0.06	0.05	0.01	0.01	0.01	
21	A	GCT	0.18	63	286	131	155	0.11	0.09	0.13	0.01	0.01	0.01	
21	S	AGC	0.25	61	271	129	142	0.11	0.09	0.12	0.00	0.01	0.00	I38V unique beneficial - 95% CI cutoff
21	S	AGT	0.16	63	223	103	120	0.06	0.04	0.08	0.00	0.01	0.00	
21	S	TCA	0.14	69	248	127	121	0.06	0.06	0.06	0.00	0.01	0.00	
21	S	TCC	0.15	84	224	93	131	-0.01	-0.06	0.03	0.00	0.01	0.00	
21	S	TCG	0.14	59	198	95	103	0.04	0.03	0.06	0.00	0.01	0.00	
21	S	TCT	0.17	29	190	106	84	0.19	0.21	0.17	0.00	0.01	0.00	
69	R	AGA	0.07	58	139	83	56	-0.03	0.01	-0.09	0.02	0.02	0.02	I38V and I122L shared beneficial - 95% CI cutoff
69	R	AGG	0.04	50	745	369	376	0.35	0.35	0.36	0.02	0.02	0.02	
69	R	CGA	0.07	102	270	139	131	-0.01	-0.01	-0.02	0.02	0.02	0.02	
69	R	CGC	0.36	123	538	277	261	0.10	0.11	0.10	0.02	0.02	0.02	
69	R	CGG	0.11	92	349	168	181	0.07	0.06	0.08	0.02	0.02	0.02	
69	R	CGT	0.36	121	460	231	229	0.07	0.07	0.07	0.02	0.02	0.02	
72	L	CTA	0.04	34	161	81	80	0.12	0.12	0.12	0.00	0.00	0.00	I38V unique beneficial - 95% CI cutoff
72	L	CTC	0.1	87	307	155	152	0.06	0.06	0.06	0.00	0.00	0.00	
72	L	CTG	0.47	101	319	136	183	0.03	-0.01	0.06	0.00	0.00	0.00	
72	L	CTT	0.12	63	212	114	98	0.05	0.06	0.03	0.00	0.00	0.00	
72	L	TTA	0.14	40	107	50	57	-0.01	-0.03	0.01	0.00	0.00	0.00	
72	L	TTG	0.13	85	227	115	112	-0.01	-0.01	-0.01	0.00	0.00	0.00	
93	A	GCA	0.23	65	182	83	99	-0.03	-0.07	0.01	0.05	0.05	0.06	Beneficial for all - 95% CI cutoff
93	A	GCC	0.26	37	77	39	38	-0.10	-0.12	-0.09	0.05	0.05	0.06	
93	A	GCG	0.33	41	820	440	380	0.38	0.38	0.39	0.05	0.05	0.06	
93	A	GCT	0.18	22	45	27	18	-0.11	-0.08	-0.15	0.05	0.05	0.06	
102	C	TGC	0.54	31	210	104	106	0.17	0.15	0.19	0.32	0.32	0.32	I38V unique beneficial - 95% CI cutoff
102	C	TGT	0.46	11	4	2	2	-0.63	-0.63	-0.63	0.32	0.32	0.32	
143	R	AGA	0.07	21	32	17	15	-0.18	-0.18	-0.18	0.02	0.01	0.03	I38V unique beneficial - 95% CI cutoff
143	R	AGG	0.04	40	73	50	23	-0.14	-0.07	-0.24	0.02	0.01	0.03	
143	R	CGA	0.07	42	56	31	25	-0.22	-0.21	-0.23	0.02	0.01	0.03	
143	R	CGC	0.36	154	1055	453	602	0.17	0.12	0.22	0.02	0.01	0.03	
143	R	CGG	0.11	59	112	57	55	-0.13	-0.14	-0.11	0.02	0.01	0.03	
143	R	CGT	0.36	53	90	49	41	-0.16	-0.15	-0.16	0.02	0.01	0.03	

Position	Mutation	Codon	Codon Frequency	Reference population count	Cumulative selected population count	Selected population count - technical replicate 1	Selected population count - technical replicate 2	Cumulative normalized fitness metric	Technical replicate 1 normalized fitness metric	Technical replicate 2 normalized fitness metric	Variance for synonymous codons of normalized fitness metrics - cumulative	Variance for synonymous codons of normalized fitness metrics - technical replicate 1	Variance for synonymous codons of normalized fitness metrics - technical replicate 2	Shared-unique beneficial mutation bin
143	T	ACA	0.17	30	67	38	29	-0.09	-0.07	-0.10	0.05	0.05	0.06	I38V unique beneficial - >10% increase in growth rate cutoff
143	T	ACC	0.4	90	2119	1113	1006	0.41	0.40	0.42	0.05	0.05	0.06	
143	T	ACG	0.25	42	141	73	68	0.01	0.01	0.02	0.05	0.05	0.06	
143	T	ACT	0.19	26	70	39	31	-0.04	-0.03	-0.05	0.05	0.05	0.06	
149	R	AGA	0.07	11	3	0	3	-0.95	None	-0.57	0.20	0.28	0.18	I38V unique beneficial - 95% CI cutoff
149	R	AGG	0.04	15	3	1	2	-0.87	-1.07	-0.73	0.20	0.28	0.18	
149	R	CGA	0.07	23	12	7	5	-0.51	-0.47	-0.55	0.20	0.28	0.18	
149	R	CGC	0.36	19	9	3	6	-0.54	-0.70	-0.43	0.20	0.28	0.18	
149	R	CGG	0.11	70	599	298	301	0.22	0.20	0.24	0.20	0.28	0.18	
149	R	CGT	0.36	15	2	1	1	-1.06	-1.07	-1.04	0.20	0.28	0.18	
197	C	TGC	0.54	13	3	1	2	0.70	0.97	0.10	0.65	0.90	0.48	I38V unique beneficial - >10% increase in growth rate cutoff
197	C	TGT	0.46	12	199	124	75	0.42	0.47	0.39	0.65	0.90	0.48	
201	N	AAC	0.51	85	174	91	83	-0.01	0.00	-0.02	0.06	0.06	0.06	I38V unique beneficial - >10% increase in growth rate cutoff
201	N	AAT	0.49	118	1150	617	533	0.33	0.35	0.32	0.06	0.06	0.06	
228	A	GCA	0.23	36	7	4	3	-0.77	-0.71	-0.85	0.29	0.30	0.31	I38V unique beneficial - >10% increase in growth rate cutoff
228	A	GCC	0.26	36	393	186	207	0.36	0.34	0.37	0.29	0.30	0.31	
228	A	GCG	0.33	18	4	2	2	-0.72	-0.71	-0.73	0.29	0.30	0.31	
228	A	GCT	0.18	12	3	1	2	-0.67	-0.93	-0.57	0.29	0.30	0.31	
237	S	AGC	0.25	53	47	24	23	-0.24	-0.23	-0.25	0.06	0.06	0.05	I38V and WT shared beneficial - 95% CI cutoff
237	S	AGT	0.16	84	111	52	59	-0.13	-0.14	-0.11	0.06	0.06	0.05	
237	S	TCA	0.14	14	27	12	15	-0.03	-0.06	0.00	0.06	0.06	0.05	
237	S	TCC	0.15	44	559	338	221	0.39	0.42	0.34	0.06	0.06	0.05	
237	S	TCG	0.14	30	49	24	25	-0.07	-0.07	-0.07	0.06	0.06	0.05	
237	S	TCT	0.17	18	14	7	7	-0.28	-0.28	-0.28	0.06	0.06	0.05	
251	Y	TAC	0.41	37	377	159	218	0.34	0.31	0.37	0.06	0.06	0.06	Beneficial for all - >10% increase in growth rate cutoff
251	Y	TAT	0.59	38	80	36	44	-0.01	-0.03	0.02	0.06	0.06	0.06	
252	A	GCA	0.23	54	104	55	49	-0.03	-0.01	-0.04	0.04	0.04	0.05	Beneficial for all - 95% CI cutoff
252	A	GCC	0.26	133	194	94	100	-0.10	-0.11	-0.09	0.04	0.04	0.05	
252	A	GCG	0.33	52	74	40	34	-0.11	-0.09	-0.13	0.04	0.04	0.05	
252	A	GCT	0.18	43	423	216	207	0.34	0.34	0.33	0.04	0.04	0.05	
287	T	ACA	0.17	24	239	136	103	0.31	0.33	0.29	0.06	0.05	0.07	I38V unique beneficial - 95% CI cutoff
287	T	ACC	0.4	25	72	32	40	0.04	0.01	0.07	0.06	0.05	0.07	
287	T	ACG	0.25	34	37	21	16	-0.21	-0.18	-0.25	0.06	0.05	0.07	
287	T	ACT	0.19	6	7	4	3	-0.19	-0.16	-0.23	0.06	0.05	0.07	
288	K	AAA	0.74	168	855	464	391	0.17	0.18	0.16	0.06	0.05	0.06	I38V unique beneficial - 95% CI cutoff
288	K	AAG	0.26	164	214	115	99	-0.16	-0.15	-0.18	0.06	0.05	0.06	
325	F	TTC	0.42	54	288	179	109	0.18	0.22	0.13	0.18	0.47	0.08	I38V unique beneficial - >10% increase in growth rate cutoff
325	F	TTT	0.58	9	5	1	4	-0.42	-0.25	-0.27	0.18	0.47	0.08	
326	R	AGA	0.07	32	43	19	24	-0.15	-0.19	-0.12	0.09	0.10	0.10	I38V unique beneficial - 95% CI cutoff
326	R	AGG	0.04	15	9	6	3	-0.39	-0.31	-0.53	0.09	0.10	0.10	
326	R	CGA	0.07	34	25	13	12	-0.33	-0.32	-0.34	0.09	0.10	0.10	
326	R	CGC	0.36	53	481	228	253	0.29	0.27	0.31	0.09	0.10	0.10	
326	R	CGG	0.11	24	13	5	8	-0.43	-0.51	-0.36	0.09	0.10	0.10	
326	R	CGT	0.36	27	10	4	6	-0.65	-0.62	-0.69	0.09	0.10	0.10	
329	S	AGC	0.25	72	156	68	88	-0.03	-0.07	0.01	0.02	0.03	0.02	Beneficial for all - 95% CI cutoff
329	S	AGT	0.16	8	35	15	20	0.14	0.09	0.18	0.02	0.03	0.02	
329	S	TCA	0.14	35	97	47	50	0.03	0.02	0.05	0.02	0.03	0.02	
329	S	TCC	0.15	113	258	105	153	-0.02	-0.07	0.03	0.02	0.03	0.02	
329	S	TCG	0.14	48	105	58	47	-0.03	-0.01	-0.05	0.02	0.03	0.02	
329	S	TCT	0.17	43	505	282	223	0.34	0.35	0.33	0.02	0.03	0.02	
329	T	ACA	0.17	43	143	75	68	0.08	0.08	0.07	0.00	0.00	0.01	Beneficial for all - 95% CI cutoff
329	T	ACC	0.4	92	295	151	144	0.07	0.06	0.07	0.00	0.00	0.01	
329	T	ACG	0.25	25	51	29	22	-0.04	-0.02	-0.07	0.00	0.00	0.01	
329	T	ACT	0.19	31	110	52	58	0.09	0.07	0.11	0.00	0.00	0.01	
330	S	AGC	0.25	126	376	180	196	0.05	0.03	0.07	0.03	0.02	0.03	Beneficial for all - 95% CI cutoff
330	S	AGT	0.16	81	172	97	75	-0.03	-0.01	-0.06	0.03	0.02	0.03	
330	S	TCA	0.14	138	1129	586	543	0.27	0.27	0.27	0.03	0.02	0.03	
330	S	TCC	0.15	208	453	213	240	-0.03	-0.05	-0.01	0.03	0.02	0.03	
330	S	TCG	0.14	101	1182	546	636	0.34	0.32	0.37	0.03	0.02	0.03	
330	S	TCT	0.17	130	341	180	161	0.02	0.02	0.01	0.03	0.02	0.03	
333	N	AAC	0.51	26	301	133	168	0.34	0.31	0.37	0.18	0.16	0.19	I38V unique beneficial - >10% increase in growth rate cutoff
333	N	AAT	0.49	17	16	8	8	-0.25	-0.26	-0.25	0.18	0.16	0.19	
336	A	GCA	0.23	130	422	222	200	0.07	0.07	0.07	0.00	0.01	0.00	Beneficial for all - 95% CI cutoff
336	A	GCC	0.26	150	579	317	262	0.11	0.12	0.10	0.00	0.01	0.00	
336	A	GCG	0.33	95	194	102	92	-0.04	-0.04	-0.05	0.00	0.01	0.00	
336	A	GCT	0.18	102	239	113	126	-0.01	-0.03	0.01	0.00	0.01	0.00	
340	K	AAA	0.74	149	831	465	366	0.19	0.21	0.17	0.06	0.06	0.06	I38V unique beneficial - >10% increase in growth rate cutoff
340	K	AAG	0.26	48	63	34	29	-0.16	-0.14	-0.18	0.06	0.06	0.06	

APPENDIX B

Chapter 4 supporting information

Note B 1: DNA sequences of ΦX174 viral capsid protein gene fragments. Sequences provided span from the 5' EcoRV to the 3' BamHI site (bolded sequences). Mutagenic regions are highlighted in pink, BsmBI sites are highlighted in green.

>Fragment_F1
GATATCTGCAGAATTCGCCCTTCGTCTCATTCAAACGGCCTGTCTCATCATGGAAGG
CGCTGAATTTACGGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGA
ATTGTTCGCGTTTACCTTGCGTGTACGCGCAGGAAACACTGACGTTCTTACTGACGC
AGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAGTGATGTAATGTCTAAAGG
TAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACTAAAGG
CAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGG
GCTTCGGCCCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCG
TATGCCGCATGACCTTTCCCATCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATT
ACCATTTCAACTACTCCGGTTATCGCTGGCGACTCCTTCGAGATGGACGCCGTTGGC
GCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTACTGTAGACATTTT
TACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAAG
GATGGTGTTAATGCCACTCCTCTCCCGACTGTGAGACGAAGGGCGAATTCCAGCACA
CTGGGCGGCCGTTACTAGT**GGATCC**

>Fragment_F2
GATATCTGCAGAATTCGCCCTTCGTCTCGACTGTTAACACTACTGGTTATATTGACC
ATGCCGCTTTTCTTGGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTT
TCAGGGTTATTTGAATATCTATAACAACTATTTTAAAGCGCCGTGGATGCCTGACCG
TACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGCTCGTTATGGTTTCCGTTG
CTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTTTCTCGC
CAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCT
AATTTGCATACTGACCAAGAACGTGATTACTTCATGCAGCGTTACCGTGATGTTATT
TCTTCATTTGGAGGTAAAACCTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGC
GCTCTAATCTCTGGGGAGACGAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTA
GT**GGATCC**

>Fragment_F3
GATATCTGCAGAATTCGCCCTTCGTCTCTCTGGGCATCTGGCTATGATGTTGATGGA
ACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACAT
TCTGTGCCGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCG
TTTTCCGCCTACTGCGACTAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGAC
TTATACCGATATTGCTGGCGACCCTGTTTTGTATGGCAACTTGCCGCCGCGTGAAATT
TCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGTTTAAGATTGCTGAG
GGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGAAG
GCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCG
CCACCATGATTATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCA
GGTTAAATTTAATGTGACCGTTTATCGCAATCTGCCGACCACTCGCGATTCAATCAT
GACTTCGTGATAAAAGATTGAGTGTGANNNNNNNNNNNNGGTTATAACGCCGAAGC
GGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGAGACGAANG
GCGAATTCCAGCACACTGGCGGCCGTTACTAGT**GGATCC**

>Fragment_G1
GATATCTGCAGAATTCGCCCTTCGTCTCCGAAGCGCGGTAGGTTTTCTGCTTAGGAG
TTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGC
TGGTTCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGC
TACATCGTCAACGTTATATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTT
CTTCATTGCATTCAGATGGATACATCTGTCAACGCCGCTAATCAGGTTGTTTCTGTTG
GTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGCCTGTTTGGTTCGCTTTGA
GTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGAATGGT
CGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCC
CGTACGGAGACGAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGT**GGATCC**

>Fragment_G2
GATATCTGCAGAATTCGCCCTTCGTCTCCGTACGCCGGGCAATAATGTTTATGTTGG
TTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGTTTCGCTGAAT
CAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTG
GTGCTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTT
TGGAGGCGGTCAAAAAGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCG
ATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTA
ATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTGGTGCTATGGCTAAAGC
TGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGATAA
GTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATAC
TCGGAGACGAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGT**GGATCC**

Note B 2: Amino acid sequences of ΦX174 mutagenized viral capsid protein gene fragments. Only mutagenized regions of gene fragments shown.

>Fragment_F1
MSNIQTGAERMPHDLSHLGFLAGQIGRLITISTTPVIAGDSFEMDAVGALRLSPLRRGLAI
DSTVDIFTFYVPHRHVYGEQWIKFMKDGVNATPL

>Fragment_F2
NTTGYIDHAAFLGTINPDTNKIPKHLFQGYLNIYNNYFKAPWMPDRTEANPNELNQDDA
RYGFRCCHLKNIWTAPLPPETELSRQMTTSTTSIDIMGLQAAYANLHTDQERDYFMQRY
RDVISSFGGKTSYDADNRPLLVMRSN

>Fragment_F3
SGYDVDGTDQTSLGQFSGRVQQTYKHSVPRFFVPEHGTMFTLALVRFPPTATKEIQYLN
AKGALTYTDIAGDPVLYGNLPPREISMKDVFRSGDSSKKFKIAEGQWYRYAPSYVSPAY
HLLEGFPFIQEPPSGDLQERVLIRHHDYDQCFQSVQLLQWNSQVKFNVTVYRNLPTTRD
SIMTS

>Fragment_G1
FQTFISRHNSNFFSDKLVLTSVTPASSAPVLQTPKATSSTLYFDSLTVNAGNGGFLHCIQM
DTSVNAANQVVSVGADIAFDADPKFFACLVRFESSSVPTTLPTAYDVYPLNGRHDGGY
YTVKDCVTIDVLP

>Fragment_G2
PGNNVYVGFMVWSNFTATKCRGLVSLNQVIKEIICLQPLK

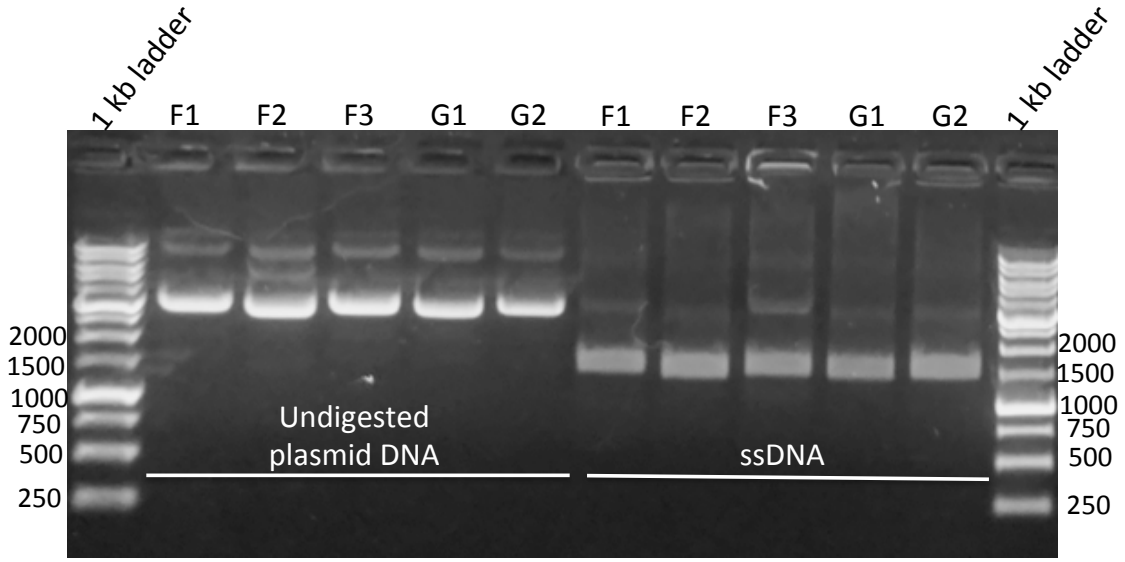

Figure B 1: Introduction of nicking sites into shuttle vectors containing viral genes. Verification of the introduction of the BbvCI nicking site into the shuttle vector containing the viral genes is shown by the generation of ssDNA as in the NSM protocol. Samples were run on a 1% agarose gel with SYBR™ Safe DNA gel stain (Invitrogen) added before casting, the ladder used is the 1 kb DNA ladder from GoldBio.

F1 heatmap of counts

Location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Mutation	M	S	N	I	Q	T	G	A	E	R	M	P	H	D	L	S	H	L	G	F	L	A	G	Q	I	G	R	L	I	T	I	S	T	T	P	V	I	A	G	D	S	F	E	M	D	A	V	G	A	L	R	L	S	P	L	R	R	G	L	A
*	767	467	357	340	380	342	152	190	242	158	302	403	428	881	1189	1879	597	620	718	799	71	336	73	239	144	217	1287	296	513	140	296	292	544	187	203	462	220	231	214	100	221	365	700	294	526	307	706	444	152	85	119	424	444	226	318	566	213	114	337	100
F	610	547	304	421	237	264	82	166	148	132	211	440	451	633	1076	1308	630	646	533	-	70	318	96	35	254	147	544	181	275	101	330	173	706	214	143	654	292	359	190	145	276	-	675	256	500	366	866	504	147	166	215	842	646	243	323	712	359	109	662	62
W	919	544	388	488	325	269	159	201	193	269	345	384	674	1006	1048	1395	697	519	779	766	43	205	70	59	71	128	588	94	176	119	285	224	217	169	169	305	160	209	159	95	158	240	639	409	409	429	705	419	162	117	243	448	462	262	216	636	252	180	418	71
Y	654	658	467	355	203	314	112	325	182	120	261	461	511	728	1096	1705	763	607	640	799	75	228	54	103	180	228	565	328	483	90	277	227	654	238	180	608	261	326	282	112	193	477	557	288	572	286	668	395	114	94	115	439	386	208	312	906	288	90	345	93
P	911	827	341	566	420	586	130	318	176	215	295	-	899	963	1331	1289	1385	531	572	1215	123	347	50	53	96	100	284	104	164	149	177	334	277	184	-	266	131	136	55	101	256	245	604	467	593	417	651	616	146	229	240	471	419	-	186	543	280	97	498	109
M	-	424	511	570	254	400	143	267	253	247	-	550	639	946	1497	1549	1101	556	712	1053	42	257	42	65	117	86	305	106	215	88	457	224	484	160	206	381	126	335	172	81	189	258	587	-	556	403	817	485	232	160	233	410	361	354	318	555	339	217	543	93
I	924	424	447	-	288	470	77	298	246	173	415	502	377	685	1547	1272	647	655	554	888	115	278	79	35	-	173	368	128	-	186	-	191	637	300	164	705	-	395	224	74	249	314	527	308	622	453	1092	526	177	184	229	652	607	360	257	887	378	112	872	65
L	1493	489	437	456	650	496	161	325	248	240	525	748	544	1029	-	1010	1079	-	696	1199	-	272	53	112	183	87	318	-	223	104	281	237	270	116	394	386	169	168	126	80	178	275	462	372	463	280	702	542	136	-	155	-	326	339	-	811	348	123	-	62
V	1180	374	347	471	331	387	275	329	389	139	391	477	476	951	1115	1221	698	687	714	848	61	294	70	73	111	247	383	79	234	122	364	231	438	164	136	-	199	493	252	152	206	272	655	437	545	529	-	518	224	117	252	473	382	370	262	549	253	184	447	187
A	1050	646	356	383	440	610	190	-	596	184	404	755	427	1003	1058	1409	749	558	774	953	48	-	82	53	95	190	295	83	180	289	257	263	343	208	192	353	119	-	219	195	177	248	501	261	480	-	777	622	-	113	221	501	430	399	157	428	294	182	438	-
G	810	583	445	370	302	344	-	467	355	238	309	880	287	1288	1072	1584	992	665	-	925	67	368	-	56	52	-	196	75	110	170	310	177	373	148	164	260	142	154	-	159	150	242	879	294	1037	367	713	-	99	69	171	552	444	172	153	454	263	-	280	88
C	1000	569	270	397	315	394	199	411	198	453	356	870	299	1094	1128	1517	565	679	616	847	76	235	85	31	126	123	439	110	153	185	295	262	425	144	167	380	223	310	182	126	297	372	765	279	677	465	717	422	131	134	261	569	540	394	222	558	392	224	496	73
S	1040	-	415	462	427	498	241	547	319	300	430	1654	521	1298	1425	-	741	584	774	840	53	389	138	44	161	126	328	81	180	221	285	-	520	274	289	307	270	452	306	170	-	535	686	338	738	549	726	521	146	118	178	598	-	465	229	733	319	191	374	90
T	827	720	624	639	361	-	122	418	229	160	313	810	272	913	1316	1125	600	472	727	571	70	263	73	58	222	138	352	140	212	-	311	203	-	-	174	295	208	386	141	188	277	302	684	320	612	611	748	445	199	100	140	903	581	391	223	496	237	150	309	155
N	733	388	-	411	321	512	111	299	225	207	310	730	351	925	950	1059	656	852	677	752	72	271	52	50	143	122	424	124	218	218	317	207	391	176	201	281	217	267	131	251	193	439	632	309	771	483	861	528	156	93	149	894	515	359	238	488	269	107	424	69
Q	986	370	421	533	-	386	169	360	348	264	382	740	783	870	1030	1193	1151	729	675	937	67	290	64	-	149	131	321	136	182	130	278	197	418	130	252	299	100	233	79	80	156	220	658	357	459	269	953	379	188	50	126	380	369	215	233	542	227	114	356	62
D	762	555	531	497	285	685	236	743	303	283	369	516	452	-	816	1774	740	576	956	794	62	298	134	35	121	311	572	162	264	109	285	249	508	191	143	373	148	421	438	-	160	279	694	357	-	383	938	472	179	209	167	572	550	362	252	706	336	177	280	92
E	707	654	349	346	306	367	227	220	-	216	260	803	536	1141	894	1368	575	886	737	658	36	241	66	52	91	179	550	117	213	120	321	234	353	166	170	254	166	338	281	194	230	296	-	285	528	273	891	455	258	145	133	505	520	337	279	409	353	92	301	109
H	746	596	539	323	589	409	147	336	183	409	377	636	-	769	1079	1583	-	964	623	824	111	239	118	83	160	172	396	281	252	132	256	212	523	156	256	396	484	225	170	129	141	321	651	399	540	525	934	383	148	115	268	406	341	292	244	831	422	72	400	63
K	706	298	461	284	418	409	148	357	360	215	373	435	731	819	1244	1435	544	711	726	846	45	217	72	65	100	156	471	52	202	124	265	231	410	153	165	289	187	271	224	148	297	436	742	247	516	344	873	580	186	117	143	561	508	372	267	403	359	201	705	121
R	1149	773	277	363	400	460	146	356	186	-	345	1005	443	1177	1087	1616	704	631	630	992	93	240	74	99	150	129	-	111	177	280	266	174	409	151	250	342	155	242	125	140	207	227	616	366	678	496	786	468	109	142	-	403	317	338	162	-	-	116	421	63

F1 heatmap of counts

Location	61	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	90	91	92	93	94	95
Mutation	I	D	S	T	V	D	I	F	T	F	Y	V	P	H	R	H	V	Y	G	E	Q	W	I	K	F	M	K	D	G	V	N	A	T	P	L
*	174	157	121	190	195	430	219	206	84	65	87	9	10	156	45	59	106	336	119	558	163	445	352	251	241	55	245	139	224	124	266	208	159	165	297
F	266	141	235	235	247	362	332	-	52	-	51	22	56	140	33	49	118	109	164	383	68	113	311	188	-	50	173	115	158	891	219	144	183	183	313
W	107	172	101	131	193	415	151	172	82	57	12	19	26	67	27	34	87	40	84	240	94	-	284	266	189	68	256	113	202	94	175	184	110	51	207
Y	139	125	224	149	194	380	260	246	81	52	-	23	10	261	33	81	96	-	198	357	101	87	334	259	269	47	77	821	145	152	207	185	187	137	269
P	155	176	96	223	204	487	171	135	92	38	23	17	-	178	57	77	86	33	78	245	133	150	273	281	177	42	177	88	108	102	188	178	121	-	259
M	181	129	77	183	199	391	440	174	99	39	23	17	46	130	33	29	137	49	141	277	101	100	629	278	178	-	302	134	242	125	229	172	193	150	255
I	-	154	215	323	297	408	-	250	169	82	40	77	49	113	44	37	168	55	192	837	63	126	-	173	272	205	249	159	176	143	266	141	268	151	314
L	184	131	100	214	208	412	172	233	79	100	14	17	145	213	76	62	120	48	110	205	167	220	280	248	353	93	222	110	203	139	197	180	122	191	-
V	242	221	155	134	-	523	244	205	77	46	19	-	33	103	41	24	-	99	157	333	124	207	362	275	473	85	222	165	398	-	240	633	180	161	289
A	182	207	215	190	205	465	222	244	282	32	18	76	91	118	43	41	107	52	140	270	137	121	258	274	147	56	217	119	193	160	239	-	873	376	230
G	118	266	118	154	215	1133	231	256	102	38	23	87	41	110	25	50	110	11	-	276	83	197	197	217	182	49	216	234	-	154	194	571	148	162	191
C	140	236	212	224	253	1011	200	313	84	41	73	35	53	84	107	52	130	95	368	374	115	192	225	242	593	43	152	141	334	146	209	265	167	165	269
S	188	272	-	241	224	909	305	217	164	82	35	21	105	92	41	29	63	29	151	234	66	207	510	257	271	49	221	203	298	102	603	323	189	201	244
T	214	232	105	-	226	841	212	159	-	52	19	36	200	152	21	26	106	53	104	413	108	97	422	232	165	90	221	106	128	89	30	398	-	419	254
N	203	273	118	290	311	1064	287	201	133	41	25	24	88	174	20	63	74	74	152	430	90	145	364	324	176	27	407	291	172	125	-	356	264	236	249
Q	105	121	83	162	179	412	211	179	137	28	29	21	121	223	56	43	106	58	126	301	-	114	270	221	154	72	298	131	167	91	242	307	179	190	351
D	187	-	132	152	361	-	252	227	130	22	54	97	115	188	48	98	157	73	300	742	78	171	332	259	237	41	199	-	311	168	328	332	234	181	217
E	260	240	128	160	327	715	258	165	66	30	21	23	51	120	35	33	120	39	108	-	69	180	222	317	228	36	337	938	210	117	267	340	192	120	274
H	144	160	138	204	287	1147	355	340	205	122	45	17	207	-	148	-	60	94	160	338	142	130	323	116	150	42	191	151	254	112	171	262	221	483	1383
K	234	167	136	199	252	437	247	165	72	24	16	9	65	122	29	20	84	43	116	514	78	190	252	-	220	501	-	126	173	129	2000	216	197	144	275
R	154	246	139	213	193	765	176	143	87	46	29	25	133	208	-	104	93	62	142	229	254	231	231	244	208	602	231	94	189	153	167	243	161	198	367

Counts

0	50	100	250	500

Figure B 2: F1 heatmap of counts.

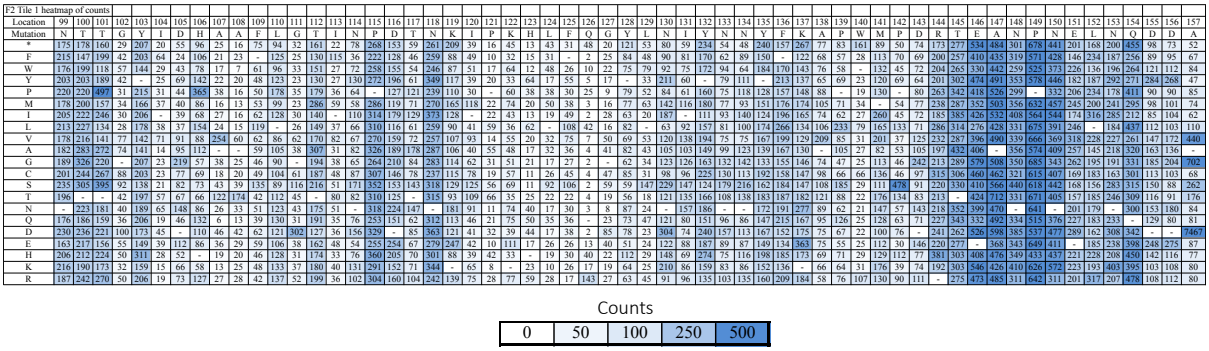

Figure B 3: F2 tile 1 heatmap of counts.

F2 Tile 2 heatmap of counts

Location	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217
Mutation	R	Y	G	F	R	C	H	L	K	N	I	W	T	A	P	L	P	P	E	T	E	L	S	R	Q	M	T	T	S	T	T	S	I	D	I	M	G	L	Q	A	A	Y	A	N	L	H	T	D	Q	E	R	D	Y	F	M	Q	R	Y	R
*	125	631	73	340	239	424	477	513	343	404	604	196	994	226	291	194	801	78	271	451	250	354	657	834	280	640	670	590	336	845	398	281	421	226	316	490	238	65	229	291	51	0	283	64	318	555	608	517	567	912	400	402	121	389	130	126	707	87	384	279
F	113	540	94	-	180	392	378	491	324	307	458	275	680	196	324	147	773	99	246	334	265	318	707	810	292	389	609	474	376	935	313	311	609	299	246	577	205	80	219	75	77	199	265	83	310	577	507	700	402	675	311	344	115	317	-	121	141	114	301	249
W	191	306	92	629	260	389	421	495	315	312	585	188	-	134	314	185	652	134	336	441	166	312	598	712	262	430	670	572	314	950	441	266	569	245	343	502	224	170	216	85	56	244	239	111	295	559	681	599	588	612	354	398	146	260	154	93	233	122	279	358
Y	97	-	134	389	206	603	1073	636	224	291	482	238	799	228	403	207	854	83	226	384	275	263	679	862	238	464	590	509	381	836	373	263	571	233	280	609	234	34	216	101	102	2	-	97	323	517	1430	693	571	701	314	406	213	-	241	80	156	75	-	344
P	199	838	166	428	255	447	407	421	718	297	760	179	753	227	406	-	927	-	-	454	298	291	976	878	377	558	742	663	545	986	555	378	628	245	359	482	238	85	658	67	74	362	177	93	265	638	708	721	522	626	371	560	117	308	161	98	245	160	292	359
M	122	401	109	511	200	389	464	500	290	336	481	196	759	191	363	190	795	151	307	396	241	292	696	776	268	553	-	955	552	855	548	337	565	240	368	613	-	97	370	85	61	301	187	47	287	599	587	572	530	580	379	449	147	14	159	-	175	90	282	306
I	122	451	83	433	192	350	417	419	279	396	568	-	757	300	341	138	811	128	290	326	379	282	844	982	285	503	926	516	400	759	575	439	523	-	326	-	559	89	242	86	84	233	258	56	364	356	562	861	458	585	272	368	148	34	185	255	102	77	262	332
L	214	355	135	728	271	399	438	566	-	321	576	189	1079	191	355	442	-	173	494	382	211	341	-	804	346	614	833	671	445	829	441	370	487	241	304	558	420	76	-	194	92	255	167	120	261	-	712	686	515	687	362	496	129	313	314	160	352	176	259	390
V	138	367	198	415	237	393	359	550	309	327	522	296	693	149	382	165	708	115	264	511	185	409	586	788	346	453	892	489	381	936	478	354	567	302	484	590	379	315	341	56	124	307	182	354	256	535	602	668	732	393	585	362	249	289	191	270	257	140	256	329
A	153	323	185	482	229	856	349	569	341	263	698	252	779	444	-	264	623	154	257	431	420	353	884	830	351	469	735	903	504	778	633	453	607	244	481	452	238	205	274	69	-	-	219	-	308	536	545	877	629	645	524	397	192	360	129	151	205	128	229	360
G	117	353	-	403	286	429	386	439	372	273	610	189	785	167	372	148	578	103	369	571	255	509	538	890	318	459	691	600	380	859	416	313	499	226	796	559	213	-	284	80	101	309	197	170	292	503	526	694	920	616	495	464	246	310	165	131	168	152	285	429
C	264	529	124	490	625	-	-	507	329	277	678	177	853	206	415	147	855	89	377	360	276	297	682	800	681	399	769	507	404	905	452	359	669	261	565	531	232	279	273	98	86	259	287	103	304	551	652	682	744	682	380	760	143	537	204	113	183	293	610	644
S	206	411	233	563	296	455	468	434	385	355	691	236	899	318	388	351	728	210	454	385	412	382	651	-	426	503	714	668	530	-	591	514	-	271	604	700	319	246	261	55	112	388	220	148	444	650	578	724	768	624	393	559	154	178	337	112	161	128	326	415
T	128	358	114	392	273	459	530	528	256	368	876	303	763	-	558	258	713	128	370	365	-	279	665	937	340	484	898	-	-	810	-	-	568	349	447	633	326	75	198	87	138	881	210	230	310	444	535	-	357	596	334	423	153	34	173	266	153	84	335	561
N	138	335	117	318	237	385	463	538	311	421	-	255	889	263	377	208	824	160	366	365	377	369	689	848	338	581	785	513	417	765	569	343	490	269	641	713	242	68	186	105	89	221	214	82	-	439	562	659	759	703	349	474	271	40	232	122	170	69	343	339
Q	188	362	86	422	247	375	352	519	336	395	624	193	842	196	333	296	592	167	320	447	282	400	684	874	385	-	644	608	419	994	532	399	555	254	403	503	264	60	400	-	59	235	208	78	271	569	801	625	628	-	426	532	140	289	159	133	-	132	256	420
D	147	408	227	341	242	375	345	531	265	353	699	198	751	236	415	245	656	130	279	468	345	363	616	996	294	499	727	609	446	810	392	287	462	280	-	624	278	335	282	76	158	288	263	200	425	508	605	753	-	692	542	381	-	331	172	99	214	100	304	365
E	117	314	135	367	216	316	419	484	238	519	564	202	852	151	398	176	523	100	309	-	266	-	729	825	358	612	663	647	404	903	454	325	591	199	500	637	240	121	220	127	86	261	270	90	311	593	509	663	893	702	-	382	229	355	205	124	212	70	337	364
H	316	602	96	302	820	396	323	-	491	301	657	200	327	195	184	261	752	147	382	561	334	345	814	919	1355	689	679	632	467	930	410	373	556	281	563	586	251	87	316	264	98	91	329	93	411	719	-	821	530	749	276	668	167	404	164	105	347	348	417	771
K	106	356	101	331	212	250	485	515	295	-	606	198	841	218	388	207	765	120	339	514	245	501	624	976	275	575	802	423	391	855	418	373	454	321	332	508	245	86	201	177	89	262	208	57	360	396	627	494	536	825	503	452	144	3	161	164	178	149	284	315
R	-	357	115	354	-	606	627	563	341	446	623	242	1138	226	385	202	753	171	477	393	287	413	593	952	-	671	861	598	391	958	503	370	619	243	507	602	192	125	349	272	85	299	243	142	346	714	905	663	773	893	365	-	149	366	203	144	462	-	334	-

F2 Tile 2 heatmap of counts

Location	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242
Mutation	D	V	I	S	S	F	G	G	K	T	S	Y	D	A	D	N	R	P	L	L	V	M	R	S	N
*	292	170	121	115	174	216	177	114	155	243	78	167	82	98	364	1022	922	474	258	170	529	446	223	102	120
F	218	258	153	185	115	-	50	106	63	165	181	99	66	95	290	1060	899	413	427	301	764	390	153	219	91
W	263	203	114	97	103	213	156	250	97	208	79	79	98	131	387	776	831	413	137	140	454	364	192	67	88
Y	337	159	129	146	108	296	99	121	69	258	114	-	112	118	360	1013	985	503	161	176	605	411	200	247	207
P	233	220	110	220	277	186	99	51	86	543	161	76	87	147	324	848	855	-	135	323	522	356	251	150	126
M	226	191	193	78	125	169	85	106	128	395	84	77	80	141	377	991	913	474	170	174	521	-	169	140	106
I	279	315	-	87	76	378	68	113	107	486	96	83	66	108	292	1258	896	454	189	325	799	654	210	83	199
L	235	230	116	107	228	368	92	82	80	274	74	93	116	161	380	789	988	512	-	-	833	585	371	111	103
V	361	-	243	110	127	222	171	331	74	326	81	72	185	252	601	870	1010	570	226	165	-	585	184	136	111
A	338	306	122	182	202	224	151	218	59	508	101	95	80	-	503	841	960	459	168	198	878	377	190	173	79
G	450	265	131	116	108	231	-	-	59	403	109	59	487	192	693	931	892	382	137	241	872	375	284	98	112
C	284	218	175	131	121	257	80	184	57	413	122	167	126	166	438	1068	972	481	171	284	615	379	1003	142	121
S	251	255	152	-	-	285	64	225	92	598	-	67	87	229	461	1212	892	690	242	236	724	455	256	-	206
T	288	234	231	130	135	173	76	107	110	-	155	102	143	367	440	1041	751	631	146	230	710	464	220	108	105
N	377	237	246	118	72	171	91	63	232	454	107	93	164	202	783	-	889	481	156	227	636	370	259	102	-
Q	224	195	121	97	168	216	101	81	113	375	100	88	83	111	478	636	957	452	130	166	626	424	186	126	112
D	-	363	140	125	109	209	137	341	79	253	113	108	-	207	-	1149	1001	441	157	203	876	463	216	122	239
E	427	272	187	102	246	214	233	226	316	350	62	52	154	149	510	1127	963	496	140	210	859	450	223	126	66
H	399	258	172	108	101	190	118	127	87	400	132	238	89	165	428	957	1271	516	149	252	485	511	654	96	111
K	317	188	202	94	123	191	124	87	-	436	111	64	106	157	397	1129	946	498	174	182	660	559	165	112	593
R	260	292	162	70	110	229	175	116	155	445	99	82	120	210	450	996	-	693	192	416	728	885	-	125	87

Counts

0	50	100	250	500

Figure B 4: F2 tile 2 heatmap of counts.

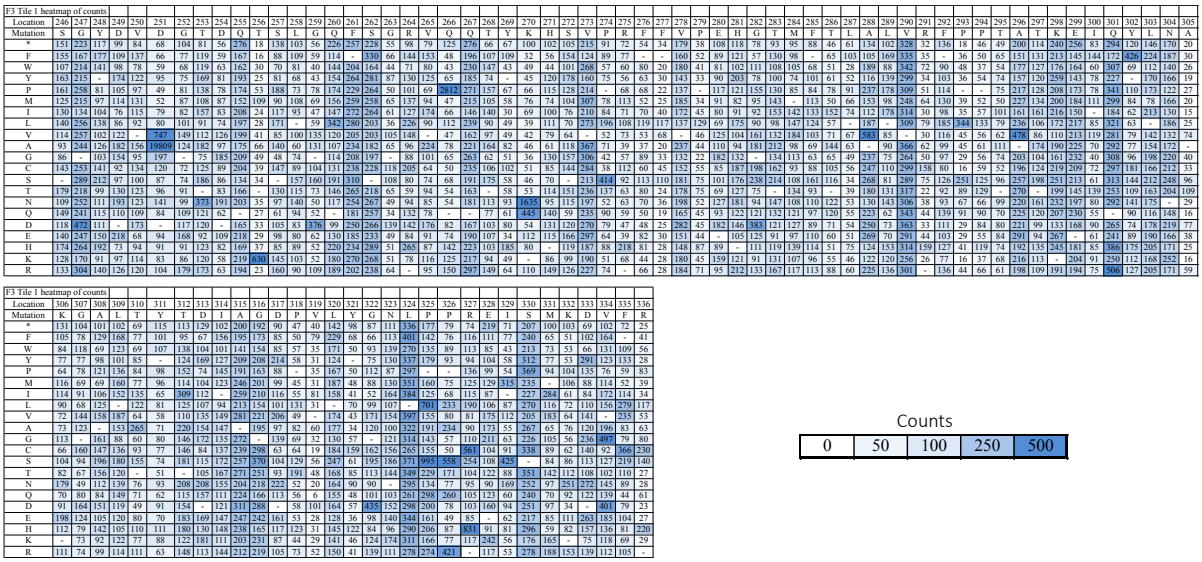

Figure B 5: F3 tile 1 heatmap of counts.

F3 Tile 2 heatmap of counts

Location	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396
Mutation	S	G	D	S	S	K	K	F	K	I	A	E	G	Q	W	Y	R	Y	A	P	S	Y	V	S	P	A	Y	H	L	L	E	G	F	P	F	I	Q	E	P	P	S	G	D	L	Q	E	R	V	L	I	R	H	H	D	Y	D	Q	C	F	Q
*	42	32	62	125	143	109	72	56	65	90	52	205	32	116	193	70	63	96	44	122	156	123	119	89	162	39	207	37	134	25	139	49	152	69	50	71	123	151	148	259	88	73	36	50	624	264	259	194	163	231	161	144	55	171	252	107	113	95	54	152
F	97	34	29	83	155	110	24	-	11	141	50	114	31	64	68	56	48	94	55	118	150	106	143	122	129	52	253	60	150	62	123	27	-	46	-	102	36	88	118	246	67	76	46	138	69	202	173	154	193	217	92	73	69	172	242	99	80	116	-	22
W	61	35	44	173	150	133	27	76	40	109	57	151	25	46	-	29	57	70	45	119	145	84	139	64	112	81	205	62	143	26	104	39	119	89	81	42	45	74	169	240	38	107	41	113	46	172	171	173	173	189	203	131	64	159	208	82	96	109	50	34
Y	116	34	91	61	187	94	32	32	16	173	68	151	30	65	95	-	64	-	54	104	116	-	124	126	156	88	-	121	155	23	97	37	195	55	135	57	47	110	114	283	67	93	79	66	46	205	259	190	118	267	119	193	118	300	-	142	68	125	94	33
P	121	26	36	244	251	136	31	76	40	99	54	150	32	90	88	21	79	65	93	-	219	56	112	133	-	79	198	68	184	98	108	38	142	-	95	65	55	105	-	-	90	67	45	77	44	174	230	201	200	159	183	125	78	206	273	112	114	66	31	61
M	57	18	40	129	99	131	93	54	102	169	57	136	27	78	87	54	114	69	63	95	133	57	113	68	155	74	182	58	140	14	95	64	91	63	66	88	52	111	163	259	54	88	57	122	50	201	227	181	138	204	164	131	85	183	287	100	78	81	70	17
I	36	31	50	42	119	112	24	65	12	-	83	126	42	43	75	35	60	70	47	112	122	103	168	85	158	69	173	63	176	74	97	45	147	68	65	-	44	87	125	203	68	105	37	44	62	172	195	261	172	-	99	110	59	187	228	118	88	84	61	20
L	69	23	51	328	173	182	25	140	24	122	56	117	17	137	726	26	90	60	84	183	211	61	123	63	168	73	195	143	-	-	94	34	255	118	182	86	90	99	285	270	67	83	32	-	105	162	312	262	-	195	219	185	96	181	250	149	113	82	195	66
V	46	84	74	139	102	104	35	124	25	189	214	303	85	71	73	35	58	64	1364	112	141	57	-	67	128	130	191	79	117	26	174	111	105	75	70	119	49	161	187	233	45	69	75	95	40	241	156	-	164	268	157	106	62	193	202	159	116	64	69	4
A	80	69	88	197	122	110	40	56	34	155	-	181	52	53	98	19	71	51	-	102	154	82	190	102	151	-	188	85	105	27	115	44	123	111	54	66	33	114	229	196	62	103	56	100	36	163	126	245	184	146	142	104	55	137	287	129	132	68	44	25
G	45	-	72	101	148	133	44	47	23	163	77	210	-	59	180	15	81	42	109	103	114	74	174	76	162	107	170	88	92	33	176	-	160	60	104	92	80	173	149	177	40	-	75	84	56	210	185	166	205	240	192	122	61	209	202	281	95	91	29	11
C	78	99	49	103	159	122	25	78	15	187	113	127	152	66	167	82	112	94	73	97	145	151	122	95	112	55	243	60	104	20	133	88	156	60	93	81	76	102	164	264	81	192	45	57	48	165	336	171	152	156	434	127	88	218	271	135	133	-	92	19
S	-	109	42	-	-	136	48	102	18	111	170	122	93	58	115	31	100	71	100	196	-	87	129	-	194	128	208	68	87	26	127	106	226	105	137	65	63	87	228	316	-	169	81	179	71	202	253	203	166	187	219	147	62	256	260	159	111	163	139	9
T	64	32	40	94	137	145	89	43	104	174	177	109	19	39	109	30	59	61	149	185	152	118	130	87	250	149	218	65	76	37	113	56	132	119	61	148	45	142	272	329	58	72	63	56	51	164	279	217	160	328	130	182	75	280	258	169	85	71	55	12
N	44	32	78	58	75	169	61	36	102	188	79	163	43	49	80	55	77	82	31	135	126	53	123	50	146	49	216	99	97	30	101	50	120	48	81	66	42	109	184	258	54	77	77	74	60	173	203	225	158	243	137	203	177	276	274	200	121	101	54	18
Q	32	13	50	176	131	241	88	50	49	91	79	152	38	-	95	37	75	85	80	146	127	84	162	83	136	37	225	72	132	35	112	45	120	120	88	61	-	117	284	276	60	85	61	99	-	159	222	215	152	166	160	132	130	228	265	91	-	97	50	-
D	78	169	-	110	139	91	46	93	19	188	185	13264	91	75	52	39	75	64	90	147	102	86	212	89	154	147	182	99	129	26	137	118	112	62	68	112	43	151	174	274	78	216	-	83	73	259	179	221	126	189	165	147	90	-	250	-	112	80	44	10
E	43	49	115	107	117	213	64	47	135	144	108	-	24	84	104	43	110	61	89	116	114	103	173	82	167	69	180	99	104	26	-	63	146	64	75	73	59	-	167	257	67	85	64	48	98	-	228	237	141	267	222	155	82	169	237	219	81	84	81	38
H	56	16	79	89	131	128	20	64	30	135	85	117	28	126	84	182	187	96	92	207	135	261	139	88	218	61	260	-	196	64	93	35	133	78	70	92	139	109	138	334	61	86	56	77	132	203	396	209	147	206	284	-	-	237	333	146	131	90	37	104
K	50	34	31	57	90	-	-	35	-	123	75	214	26	114	94	47	93	67	56	130	136	82	117	83	176	69	181	61	126	30	170	51	149	55	83	117	90	190	133	235	88	92	33	59	114	345	243	194	101	261	177	119	85	123	279	123	124	67	52	35
R	49	49	24	137	161	172	92	66	106	138	75	102	28	145	208	25	-	50	49	123	125	70	172	67	147	80	222	211	163	33	106	61	111	74	77	53	229	84	222	279	59	157	35	78	318	176	-	134	184	178	-	219	102	197	253	170	169	111	66	168

F3 Tile 2 heatmap of counts

Location	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427
Mutation	S	V	Q	L	L	Q	W	N	S	Q	V	K	F	N	V	T	V	Y	R	N	L	P	T	T	R	D	S	I	M	T	S
*	24	43	100	78	27	456	225	141	57	104	61	86	71	98	51	117	287	323	99	133	63	98	249	209	100	124	5992	81	246	81	184
F	58	94	22	142	69	24	40	99	38	59	93	87	-	59	38	63	296	193	91	140	77	105	202	176	142	127	260	141	167	162	101
W	20	54	45	51	24	29	-	134	67	80	63	42	172	56	89	116	186	232	149	164	92	118	166	187	123	93	227	96	251	104	151
Y	83	48	19	66	11	54	60	179	34	36	66	53	82	157	47	119	262	-	129	171	71	113	225	174	125	127	266	101	165	146	82
P	90	39	57	77	27	48	52	132	45	102	74	63	103	80	63	170	231	162	151	159	181	-	197	334	166	91	301	109	217	181	250
M	14	86	50	107	69	34	76	125	61	100	61	85	85	32	83	96	318	266	143	211	141	122	222	167	152	121	295	403	-	186	177
I	25	177	24	68	15	34	58	177	90	59	133	90	72	128	35	251	456	222	74	163	96	106	325	238	7	111	325	-	379	241	73
L	18	40	82	-	-	67	104	120	42	109	117	64	145	34	101	89	258	194	247	212	-	466	179	163	312	105	519	118	275	138	214
V	20	-	42	91	12	30	65	126	73	72	-	108	93	65	-	89	-	262	130	171	115	164	222	144	158	160	231	125	342	93	162
A	31	141	24	89	38	38	46	138	53	74	128	103	105	94	138	197	301	227	154	144	94	213	401	376	124	116	381	71	151	430	253
G	28	68	16	51	23	20	42	144	76	31	65	88	124	178	94	193	182	146	174	180	109	133	191	156	161	162	306	146	198	167	123
C	52	38	28	99	24	26	152	126	88	47	56	74	139	94	89	114	253	244	671	204	64	148	166	163	259	105	198	119	262	211	121
S	-	58	36	125	85	33	54	196	-	59	89	71	157	166	63	218	231	288	204	232	79	867	411	804	238	101	-	156	243	1054	-
T	87	56	39	60	19	59	57	114	71	62	85	62	93	125	35	-	298	261	136	168	91	223	-	-	165	108	345	184	265	-	127
N	28	53	32	93	21	37	75	-	71	54	80	151	119	-	57	363	254	299	110	-	121	157	235	208	149	192	318	170	219	189	102
Q	19	37	-	84	19	-	71	108	60	-	40	96	113	90	35	138	193	68	56	172	127	256	169	162	133	112	268	115	230	135	143
D	30	67	20	61	12	29	63	245	89	33	84	90	57	195	73	102	198	238	135	338	131	86	175	197	102	-	235	101	216	169	126
E	33	44	37	110	40	30	56	138	65	62	94	183	86	149	130	101	240	254	151	188	97	91	216	226	125	161	342	86	243	85	99
H	22	51	75	72	26	93	63	128	66	173	60	88	75	165	45	100	241	79	625	205	116	146	213	189	197	124	318	110	261	155	83
K	22	68	44	63	15	68	55	234	39	158	45	-	52	135	35	147	255	240	65	346	109	96	180	218	117	95	288	106	274	118	110
R	37	58	145	84	22	178	176	176	105	145	78	136	90	127	69	149	256	168	-	178	110	183	208	175	-	89	311	126	307	150	161

Counts

0	50	100	250	500

Figure B 6: F3 tile 2 heatmap of counts.

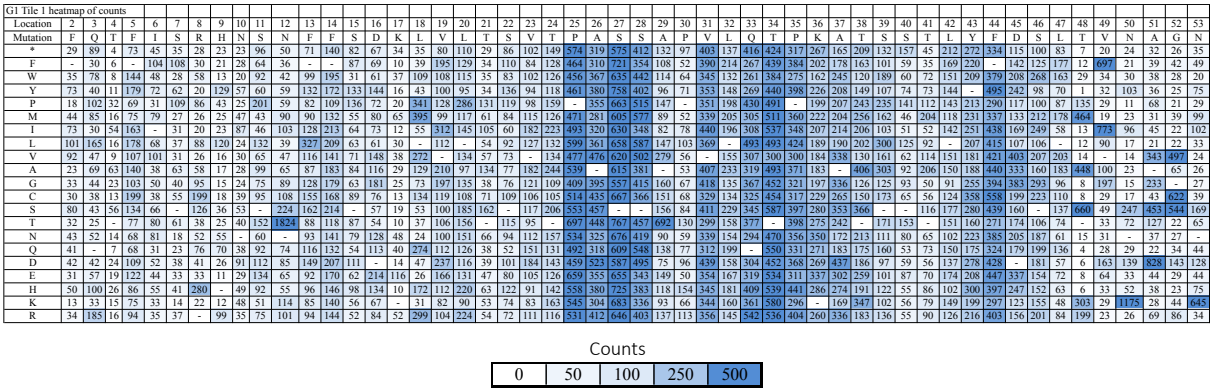

Figure B 7: G1 tile 1 heatmap of counts.

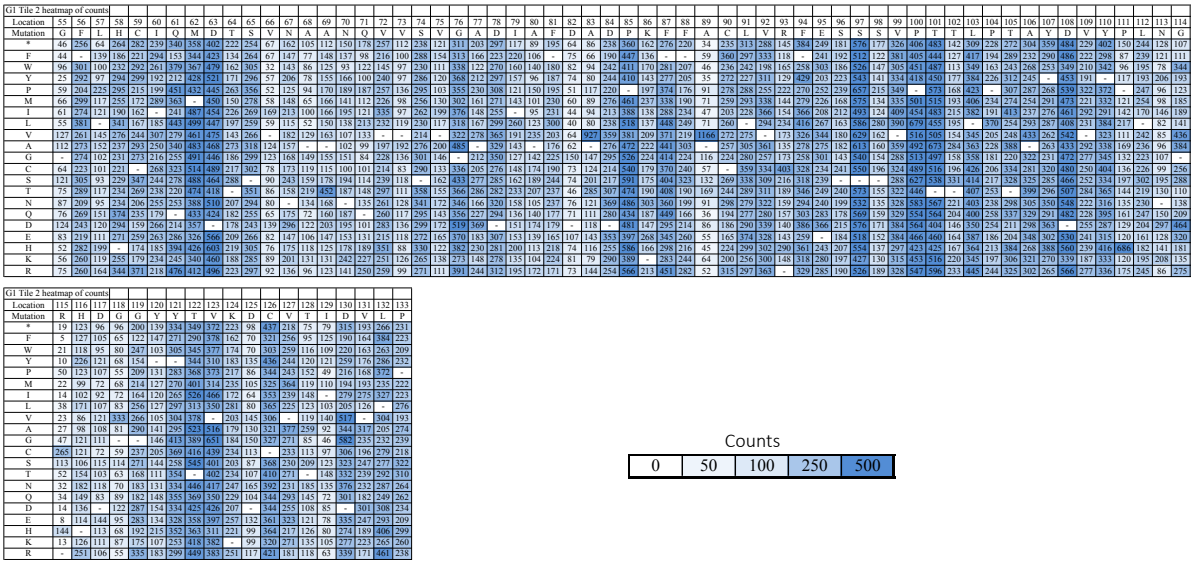

Figure B 8: G1 tile 2 heatmap of counts.

G2 heatmap of counts																																					
Location	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
Mutation	P	G	N	N	V	Y	V	G	F	M	V	W	S	N	F	T	A	T	K	C	R	G	L	V	S	L	N	Q	V	I	K	E	I	I	C	L	Q	P	L	K
*	880	725	785	330	686	1018	769	660	607	348	82	241	504	163	839	422	599	645	651	706	390	844	360	1176	486	457	879	756	980	829	1454	576	458	232	172	174	240	164	252	522
F	623	548	782	243	585	1000	767	771	-	372	147	182	477	156	-	378	444	713	421	607	311	627	311	1175	464	304	871	573	987	842	1018	448	683	525	154	183	112	72	293	386
W	724	708	612	276	788	934	697	1100	592	508	86	-	362	135	890	391	579	700	519	713	281	645	327	841	601	445	850	908	845	691	1128	1060	383	215	524	135	278	135	203	467
Y	733	604	895	380	720	-	793	668	554	425	109	215	493	175	984	439	563	679	551	665	385	665	322	1140	537	309	1121	661	1050	757	1218	408	536	426	202	140	120	152	261	367
P	-	757	689	277	776	882	822	950	610	513	92	293	568	184	964	496	543	859	529	560	329	638	378	1078	715	702	936	1093	1073	751	1263	1015	442	256	118	326	361	-	292	520
M	878	700	894	250	765	946	743	866	544	-	107	259	400	184	1001	390	525	823	627	692	352	703	365	1090	655	585	1110	890	1094	1022	1175	1039	834	566	164	213	291	158	255	527
I	716	648	825	454	722	970	830	753	550	660	158	181	405	142	1072	544	594	594	554	561	344	658	278	1217	446	388	1147	663	1163	-	1406	468	-	-	146	248	122	66	323	485
L	910	771	723	358	682	962	752	839	642	668	101	243	466	155	1009	529	538	751	467	608	352	789	-	1143	704	-	860	1192	960	834	1252	985	488	222	136	-	511	218	-	515
V	776	664	743	243	-	803	-	1196	622	700	-	311	360	149	870	425	737	709	570	660	389	795	411	-	568	586	963	735	-	841	1177	1605	511	345	167	195	273	122	315	526
A	988	763	740	316	956	931	838	1292	658	611	142	281	428	175	1054	547	-	846	537	676	331	666	270	1152	740	494	983	785	1136	911	1265	1648	482	266	201	160	267	177	201	538
G	614	-	655	249	762	915	790	-	673	518	128	265	331	211	880	473	674	690	629	757	386	-	320	1132	550	369	901	881	1182	822	1265	1245	472	184	223	241	194	116	255	464
C	839	1133	792	346	799	1032	797	944	673	550	100	337	451	217	942	517	573	794	546	-	513	649	396	1238	538	421	932	835	1067	903	1258	841	495	281	-	251	221	160	251	474
S	811	1001	884	352	792	914	793	878	758	486	93	243	-	260	926	613	916	776	543	760	492	865	319	1324	-	362	1133	721	1140	931	1540	893	513	348	190	275	291	197	270	542
T	1015	892	882	1487	749	1093	794	854	699	469	74	208	530	248	1051	-	581	-	521	637	433	745	314	1234	532	297	1139	813	960	755	1333	608	474	249	146	197	144	166	256	472
N	935	809	-	-	747	1160	841	816	683	384	96	133	480	-	1006	548	551	783	580	566	383	775	295	1170	522	334	-	931	1077	845	1661	603	501	285	139	252	136	106	249	474
Q	1125	808	808	320	818	872	739	847	534	567	83	265	481	167	892	512	532	785	549	618	347	722	322	1086	693	684	1021	-	1061	831	1292	1229	381	258	138	301	-	233	309	494
D	691	988	907	368	768	1105	958	1141	665	479	169	230	369	196	953	428	775	767	620	617	344	719	272	1236	559	328	1076	796	1233	865	1393	1195	568	411	174	160	186	104	286	460
E	713	701	767	326	940	1009	836	1010	613	427	108	196	436	183	944	421	647	781	598	721	431	911	286	1191	514	330	1025	741	1072	878	1544	-	479	334	181	201	180	174	224	480
H	905	718	776	393	795	995	866	815	586	479	130	161	481	176	1060	400	530	710	535	569	478	786	252	1198	445	520	1154	1023	1065	784	1097	477	621	316	111	292	419	179	368	413
K	783	786	913	338	750	941	749	799	630	405	96	215	400	218	866	440	577	670	-	587	416	872	366	1093	511	283	1140	752	917	832	-	572	487	267	115	213	176	137	255	-
R	848	907	842	380	809	883	807	859	668	459	110	216	502	223	969	540	655	783	501	813	-	838	277	1076	641	546	947	1033	1125	930	1429	839	508	260	213	504	353	200	388	498

Counts

0	50	100	250	500

Figure B 9: G2 heatmap of counts.

Table B 1: Mutant library preparation summary. Summary table of the transformants required for sufficient library coverage, transformants obtained during comprehensive mutant library preparation by nicking scanning mutagenesis, and the fold excess of the number of transformants required for coverage.

Gene mutated	F1	F2	F3	G1	G2
Number of Residues	95	144	182	132	40
Transformants obtained following NSM	620,000	890,000	370,000	620,000	640,000
Required transformants for 99.9% coverage of possible library	13781	20889	26401	19148	5803
Fold excess over amount required for coverage	45	43	14	32	110

Table B 2: Primers for incorporating BbvCI nicking sites into the pCR2.1-topo shuttle vector. Blue text is the overlap region with the shuttle vector, red is the KpnI site, green is the BbvCI site.

Nick_incorporation_fwd
GCTCTACGGGTACCGCTGAGGGAGCTCGGATCCACTAGTAACG
Nick_incorporation_rvs
GCTCTAACGGGTACCAAGCTTGGCGTAATCATGGTC

Table B 3: Inner and outer primers for PCR reactions for Illumina sequencing. Red indicates overhang regions for attaching Illumina adapter primers (inner PCR primers) or overhangs for attaching to inner PCR product (outer PCR primers), black is the overlap region in the gene or the barcode, blue is the Illumina adapter.

Inner PCR Primers	**Sequence (5' to 3')**
Fragment F1 Fwd	gttcagagttctacagtccgacgacccttacttgaggataaatt
Fragment F2 Tile 1 Fwd	gttcagagttctacagtccgacgatcgactcactatagggcgaa
Fragment F2 Tile 2 Fwd	gttcagagttctacagtccgacgatcagcttaatcaagatgatgct
Fragment F3 Tile 1 Fwd	gttcagagttctacagtccgacgatccgtctctctgggca
Fragment F3 Tile 2 Fwd	gttcagagttctacagtccgacgatcgaaggatgttttccgt
Fragment G1 Tile 1 Fwd	gttcagagttctacagtccgacgatcgaattgggccctctag
Fragment G1 Tile 2 Fwd	gttcagagttctacagtccgacgatcggttaatgctggtaatgg
Fragment G2 Fwd	gttcagagttctacagtccgacgatcggccctctagatgca
Fragment F1 Rvs	ccttggcacccgagaattccattcgtctcacagtcgg
Fragment F2 Tile 1 Rvs	ccttggcacccgagaattccacaacggaaaccataacg
Fragment F2 Tile 2 Rvs	ccttggcacccgagaattccattcgtctccccagag
Fragment F3 Tile 1 Rvs	ccttggcacccgagaattccatcttagacgaatcaccaga
Fragment F3 Tile 2 Rvs	ccttggcacccgagaattccatcacactcaatcttttatca
Fragment G1 Tile 1 Rvs	ccttggcacccgagaattccatgcaatgaagaaaacca
Fragment G1 Tile 2 Rvs	ccttggcacccgagaattccacttcgtctccgtacg
Fragment G2 Rvs	ccttggcacccgagaattccattgaccgcctcca
Illumina outer primer adapter	aatgatacggcgaccaccgagatctacacgttcagagttctacagtccga
Illumina outer PCR adapters and barcodes	
RPI31 (Fragment F1)	caagcagaagacggcatacgagatATCGTGgtgactggagttccttggcacccgagaattcca
RPI15 (Fragment F2 Tile 1)	caagcagaagacggcatacgagatTGACATgtgactggagttccttggcacccgagaattcca
RPI16 (Fragment F2 Tile 2)	caagcagaagacggcatacgagatGGACGGgtgactggagttccttggcacccgagaattcca
RPI17 (Fragment F3 Tile 1)	caagcagaagacggcatacgagatCTCTACgtgactggagttccttggcacccgagaattcca
RPI18 (Fragment F3 Tile 2)	caagcagaagacggcatacgagatGCGGACgtgactggagttccttggcacccgagaattcca
RPI19 (Fragment G1 Tile 1)	caagcagaagacggcatacgagatTTTCACgtgactggagttccttggcacccgagaattcca
RPI20 (Fragment G1 Tile 2)	caagcagaagacggcatacgagatGGCCACgtgactggagttccttggcacccgagaattcca
RPI21 (Fragment G2)	caagcagaagacggcatacgagatCGAAACgtgactggagttccttggcacccgagaattcca

APPENDIX C

Purification of the TROP2 extracellular domain from a stable insect cell line using ammonium sulfate precipitation

Abstract

The tumor associated calcium signal transducer 2 (TROP2) is an oncogenic transmembrane protein that is overexpressed in aggressive and late stage cancers. Here we present a facile method for purifying approximately 3 mg/L quantities of the extracellular domain of TROP2 from a stable *Drosophila* Schneider's (S2) cell line. The secreted oncogene is purified by ammonium sulfate fractionation, immobilized metal affinity chromatography, and size exclusion chromatography. The folding fidelity of the highly purified product was confirmed with a binding assay that tests for a unique conformational epitope in TROP2.

Introduction

The tumor associated calcium signal transducer 2 (TROP2) is a 36 kDa type-I transmembrane glycoprotein associated with late stage and aggressively metastatic tumors[1-3]. TROP2 is basally expressed in human trophoblasts when the fetal tissue joins into the maternal circulation[4]. Ectopic expression of TROP2 in cancer cells results in increased proliferation and tumorigenicity[5-8], while gene silencing of TROP2 decreases these same characteristics[5,7,9]. Following an activation event that is currently unknown, TROP2 undergoes regulated intramembrane proteolysis[10] at two distinct sites to release the extracellular domain (TROP2Ex, residues 27-274) and the intracellular domain (TROP2Ic, residues 298-323). TROP2Ic migrates to the nucleus where it increases the expression of genes associated with cell proliferation[10-11]. The liberated extracellular domain interacts with signaling proteins, though with which ones and how is incompletely understood[8,12-13].

Previous biophysical analysis has found that TROP2Ex can assume a dimeric conformation similar to its paralog epithelial cell adhesion molecule (EpCAM)[14-16]. Interestingly, the glycosylation state of the protein seems to dictate whether the monomeric (favored when fully glycosylated) or dimeric (favored when unglycosylated) form predominates at equilibrium[15]. Previous work studying the migration of metastatic cancer cells found that the oligomeric state of TROP2 might govern its function[9]. Whereas binding of TROP2 with the bivalent m7E6 monoclonal antibody (mAb)[17] resulted in significant inhibition of migration of MDA-MB-231 cancer cells while not impacting toxicity or cell proliferation, binding with the monovalent m7E6 Fab did not significantly alter migration. These results suggest that the dimer to monomer transition of TROP2 may play a role in its activation and mechanism of action.

Here we provide a novel and facile approach for the purification of properly folded TROP2Ex from insect cells. We used a stably transformed *Drosophila* Schnieder's (S2) cell line to secrete the protein. The secreted protein is removed from the culture media with a classical but underutilized ammonium sulfate fractionation method. Next, the protein is further purified with immobilized metal affinity chromatography and size exclusion chromatography. We obtain yields above 3 mg/L of induction culture with excellent purity. The correct folding of the protein has been verified with flow cytometry binding assays using a single chain variable fragment (scFv) of the m7E6 mAb displayed on the yeast surface, which recognizes a conformational epitope in TROP2Ex. The purified TROP2Ex can be used in a variety of *in vivo* and *in vitro* experiments to help understand how TROP2 is activated. Additionally, the yeast displayed m7E6 scFv system presented here can be used to map the paratope involved in the TROP2 m7E6 binding interaction.

Results and discussion

Expression of TROP2Ex from a stable cell line

As TROP2Ex contains six potential disulfide bonds and multiple N-linked glycosylation sites, we selected *Drosophila* Schneider 2 (S2) cells for our expression host

because of their potential ability to accurately maintain these post-translational modifications, their low house-keeping requirements, ease of transfection, the advantages of protein secretion, and the ease of forming stably transfected cell lines. The TROP2 gene was truncated to residues

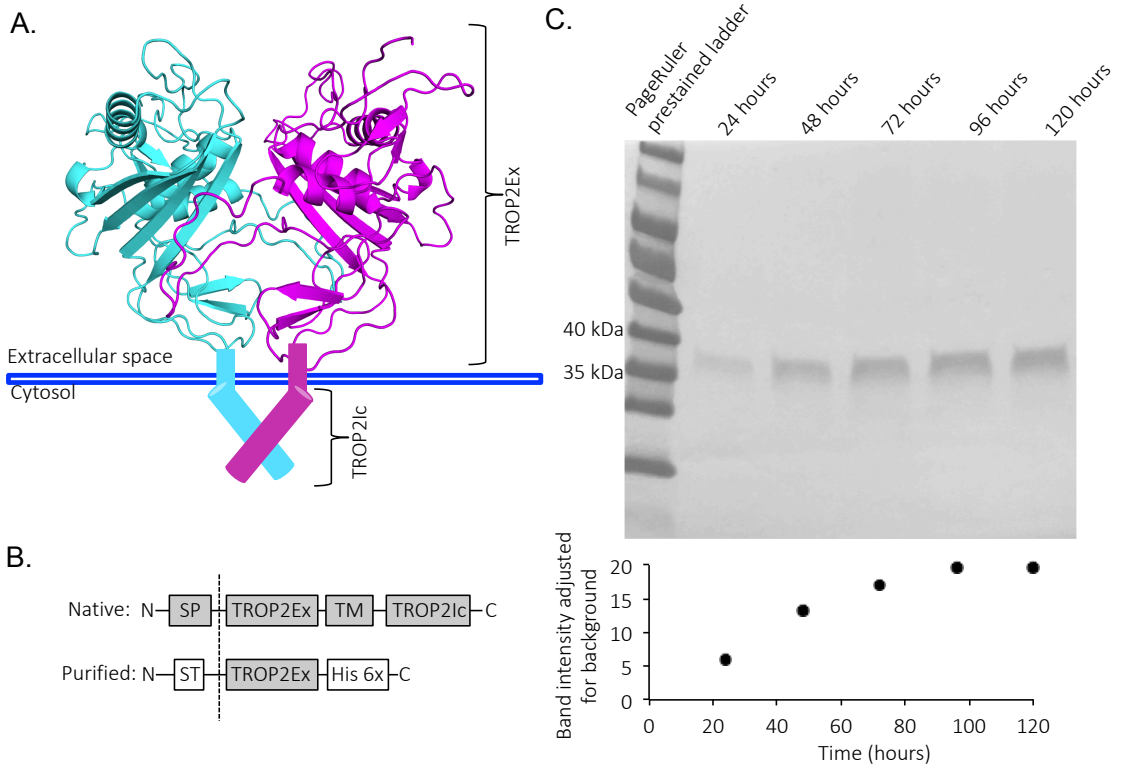

Figure C 1: TROP2Ex and its expression. A. Homology model of the dimeric TROP2Ex domains generated using the paralog EPCAM[14,16] (pdb: 4MZV). This model was previously published[9]. B. Diagram of the native and purified form of TROP2, with the dashed line representing the truncation point in the secretory pathway, SP = signal peptide, ST = secretion tag, TM = transmembrane domain, TROP2Ic = TROP2 intracellular domain. C. Western blot testing for TROP2Ex in equivalent volumes of the stable S2 cell supernatant following initiation of induction and band intensity adjusted for background is plotted against time; bands were quantified with ImageJ by standard protocols.

27-254 (TROP2Ex) for expression and secretion in S2 cells (**Figure C 1A-B**). The native secretion signal peptide sequence (SP) was replaced with the secretion tag needed for expression in S2 cells (ST) (**Figure C 1B**). A His 6x tag was added to the C-terminal end of the truncated TROP2Ex gene (**Figure C 1B**) to facilitate purification.

A stable S2 cell line that secretes TROP2Ex was successfully generated as described in the **Materials and Methods** (**Figure C 1C**). The secreted TROP2Ex runs larger than the predicted molecular weight - ~28.8 kDa – because N-linked glycosylations can occur at up four distinct sites after translation[15]. The robustness of the cell line was verified through multiple successful TROP2Ex purifications from the same lineage over the course of months (data not shown). After generating the stable S2 cell line we next increased the culture volumes possible by transitioning from static to shaking cultures. The addition of a non-ionic surfactant the cells maintained healthy morphologies in the shaking cultures as assessed by confocal microscopy. The optimal induction time for secreted TROP2Ex titers was performed the shaking cultures, with a 96-hour induction period striking the best balance between titers and volumetric productivity (**Figure C 1C**). The shaking cultures were successfully scaled to 300 mL and induced for 96 hours using copper.

Purification of TROP2Ex

TROP2Ex was purified using ammonium sulfate fractionation, IMACs, and chromatography. Fractionation with ammonium sulfate at 40% saturation, followed by fractionation at 65% saturation efficiently removed some of the contaminating proteins and contaminant copper and allowed for resuspension of the secreted TROP2Ex in IMAC buffer (**Figure C 2A**). Batch IMAC resulted in nearly pure TROP2Ex (**Figure C 2A**) and size exclusion chromatography (SEC) using a Superdex 200 column equilibrated with 1x PBS at 4°C removed the remaining impurities (**Figure C 2B-C**). TROP2Ex eluted from SEC as a peak with a shoulder between 78-90 mL elution volume (**Figure C 2B-C**), suggesting that both monomeric and dimeric TROP2Ex exist under elution conditions.

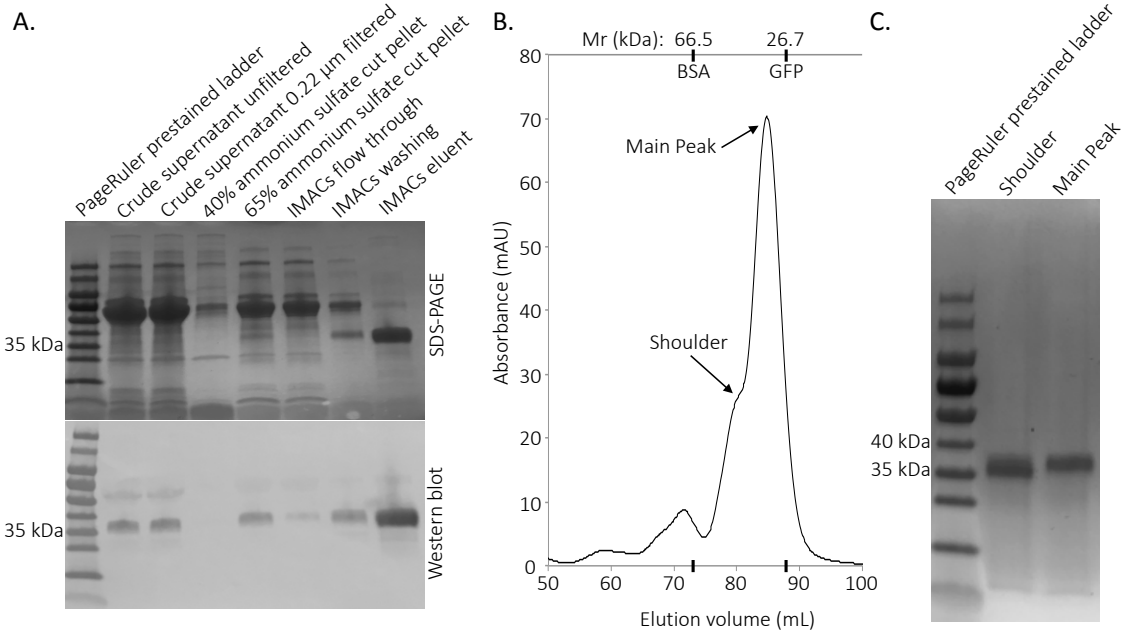

Figure C 2: Purification of TROP2Ex from insect cell cultures. A. Images of an SDS-PAGE gel and a Western blot containing identical samples resulting from the purification of TROP2Ex. B. FPLC-SEC chromatogram of the IMACs eluent, column was equilibrated with 1x PBS at 4°C, main peak and respective shoulder at ~78 – 90 mL are TROP2Ex. C. SDS-PAGE image of the shoulder (80 mL elution volume) and main peak (85 mL elution volume).

Interestingly, when the purified TROP2Ex main peak is run again with the same SEC conditions as before both the main peak and shoulder are resolved (data not shown). All subsequent analysis corresponds only to TROP2Ex from the main peak (**Figure C 2C**). Under conditions used in these experiments TROP2Ex (unglycosylated mass = 28.8 kDa) elutes at ~85 mL (**Figure C 2B**), while BSA (66.5 kDa) elutes at ~73 mL, and GFP (26.7 kDa) elutes at ~88 mL off of the Superdex 200 column in identical conditions. This elution pattern indicates that our purified TROP2Ex is predominantly monomeric. This finding is consistent with previous results finding that only a small fraction of the glycosylated TROP2Ex is in the dimeric form when at equilibrium[15]. In total, we obtain yields of 3.6 mg/L of purified monomeric TROP2Ex from the stable cell line.

Validation of TROP2Ex folding

To verify that TROP2Ex was in a properly folded conformation, we assessed its binding to a yeast displayed scFv derived from the high affinity anti-TROP2 mAb m7E6[17] (**Figure C 3A**). m7E6 recognizes a conformational epitope

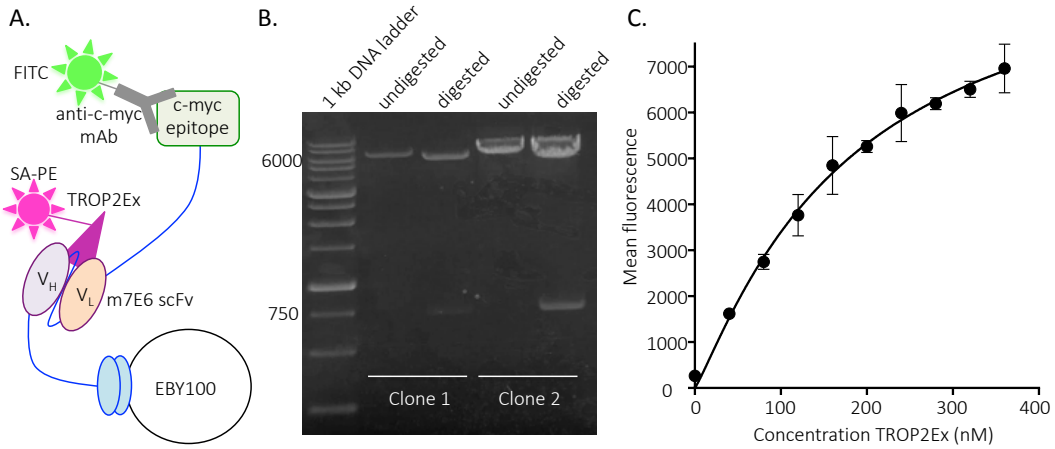

Figure C 3: Verification of the proper assembly of the m7E6 scFv yeast display construct and of binding to TROP2Ex. A. A diagram of the yeast display system, the binding interactions detected, and the respective conjugate fluorophores: SA-PE = streptavidin R-phycoerythrin conjugate. B. A DNA gel containing restriction enzyme analysis products for verifying the assembly of the m7E6 scFv pETconNK construct via homologous recombination. Following digestion the DNA was run on a 1% agarose gel prestained with ethidium bromide at 120 V for 50 minutes, the 1 kb DNA ladder is from GoldBio. Following electrophoresis the gel was imaged using a Safe Imager™ Blue Light Transilluminator. C. Clonal titration of yeast displayed m7E6 scFv with TROP2Ex, n = 3, error bars = 1 S.D.

located in the membrane distal region of TROP2Ex[9]. Overhang PCR and homologous recombination in yeast were used to assemble to the anti-TROP2 m7E6 scFv yeast display construct as described in the **Materials and Methods** section (**Figure C 3B**). Clonal titrations resulted in a binding curve that spanned a >25-fold change in signal intensity and provide a preliminary dissociation constant (K_D) of 180 ± 54 nM (**Figure C 3C**). There are multiple possible reasons the titrations didn't reach saturation: we didn't expand the titrations to high enough concentrations, the yeast displayed form of m7E6 might not be stable, or perhaps the

binding reaction conditions are not fully optimized for this pairing. The obtained K_D is ~4 times larger than the ~44 nM K_D reported for the binding of this pairing where TROP2Ex is displayed on the surface of yeast and recombinant m7E6 fragments of antibodies (Fabs) are applied[9]. The large difference in the K_D could be the result of the titrations failing to reach saturation, different glycosylation patterns on either protein, differences introduced into the CDRs or framework of the antibody fragments resulting from the truncation of the framework region, or it is possible that the orientation of the V_H and V_L domains of the scFv aren't optimal. This pairing binds with a dissociation constant in the hundreds of nanomolar, indicating that the purified TROP2Ex contains the same conformational epitope as the native TROP2Ex and is likely correctly folded.

Conclusion

Here we presented a facile method for purifying the extracellular domain of TROP2 from *Drosophila* Schneider's (S2) insect cells. The cell lines generated are stable and can be seeded for TROP2Ex production once recovered from storage. Removing the need for retransformation allows for an easy scale up of the induction cultures. Additionally, we have presented a method for purifying TROP2Ex using ammonium sulfate fractionation, IMACs, and chromatography. This method allows for the purification of >3 mg/L of properly folded TROP2Ex from an induction culture in a single day.

Materials and methods:

Materials

Antibiotics were purchased from GoldBio and ThermoFisher. All enzymes were from New England Biolabs (NEB), and all other chemicals were purchased from Sigma-Aldrich. The

TROP2Ex gene block, and all primers were purchased from Integrated DNA technologies (IDT); the m7E6 full IgG gene which was purchased from GenScript. Flow cytometry probes were purchased from Miltenyi Biotec (anti-c-myc FITC) and ThermoFisher (streptavidin R-phycoerythrin conjugate). Other materials used have their respective manufacturers listed in the text.

Plasmid Construction

The gene for TROP2Ex (Kozac-Bip-TROP2Ex) was designed as follows following a 5' to 3' convention: a KpnI restriction site, followed by a Kozac sequence and a Bip secretion tag (**Supporting text, Note S1**), then the TROP2Ex gene (residues 27-274), with an AgeI restriction site. Kozac-Bip-TROP2Ex was codon optimized for expression in *D. melanogaster* and was ordered as a single gene block from IDT (**Supporting text, Note S1 and S3**). The Kozac-Bip-TROP2Ex gene was subcloned into the pMT/V5-His B plasmid (Invitrogen) using KpnI and AgeI restriction enzymes and standard cloning methods to produce pMT-Bip-TROP2Ex containing an in-frame C-terminal 6x His tag. The final construct was transformed into *E. coli* XL-1 Blue and a glycerol cell stock was generated. A scratch of the freezer stock was used to inoculate 50 mL LB + 50 μ g/mL carbenicillin which was grown overnight at 37°C. DNA was Midi-prepped (Qiagen) from 25 mL of the overnight culture, and the purified DNA was resuspended in nuclease free H_2O and frozen at -20°C.

The m7E6 scFv[17] (**Supporting text, Note S2 and S4**) comprises (5' to 3') V_H residues 1-118, a flexible $(GGGGS)_3$ linker, and V_L residues 1-111. Using the full m7E6 IgG gene that was previously purchased from Genscript[17] as template, overhang PCR was used to assemble the scFv, introduce the linker, and introduce homology regions from the pETconNK yeast display

plasmid[18]. Primers used for assembly are in **Note S5** of the supporting text. The PCR products were cleaned and concentrated using a Monarch® PCR & DNA Cleanup Kit (NEB) and eluted into nuclease free H_2O. The scFv gene was assembled with pETconNK vector cut with NdeI and XhoI using homologous recombination into *S. cerevisiae* EBY100 cells. For homologous recombination the V_H and V_L inserts where combined with the cut pETconNK plasmid at a ratio of 30 fmol : 1 fmol (respective insert : cut vector) in a final volume of 8 µL. All of the DNA mixture was transformed into EBY100 chemically competent yeast using standard protocols. Transformation mixtures were plated on agar plates containing Synthetic Complete media supplemented with 2% (w/v) dextrose (SDCAA agar) and grown at 30°C for 2-3 days. Colonies were picked and grown in 1 mL of SDCAA and 1x penicillin/streptomycin (SDCAA media) at 30°C with shaking at 250x rpm for ~16 hours. The 1 mL cultures were then used to inoculate 50 mL of SDCAA media that was grown for generation of freezer stocks essentially as described in Klesmith et al.[18]. Both plasmids used in this work are available from Addgene: m7E6 scFv pETConNK (ID: 125731), and pMT-Bip-TROP2Ex (ID: 125730).

Transformation of S2 Drosophila Cells and the Generation of the Stable Cell Line

The S2 *Drosophila* cells were cultured, transformed, and a stable cell line was generated essentially as described by the manufacturer (Invitrogen). Briefly, 3 mL/well of Schneider's S2 *Drosophila* Media (ThermoFisher) supplemented with 10% heat-inactivated fetal bovine serum (full S2 media) was seeded with $1x10^6$ cells/mL in a 6-well plate (Fisher Scientific). Cells were incubated at 28°C until a density of $2\text{-}4x10^6$ cells/mL was reached, after approximately 16 hours. After reaching density the cells were transformed with 19 µg pMT-Bip-TROP2Ex plasmid and 1 µg pCoPuro plasmid[19] via standard calcium phosphate transfection. Calcium phosphate DNA

precipitates were prepared using the Calcium Phosphate Transfection Kit (Invitrogen) as described by the manufacturer. The DNA precipitate solution was added to respective wells such that 5 μg total DNA was applied. Additions were performed in a dropwise fashion with gentle swirling of the culture after each addition. After addition of the DNA precipitates the cultures were incubated at 28°C overnight. The next morning the transformation cultures were spun at 440 xg for 3 minutes and the transfection media was removed via aspiration. Cells were resuspended in 3 mL of full S2 media and spun as before. This media was aspirated and the washing procedure was repeated once more. Cells were resuspended in 3 mL full S2 media, placed back into the original well and incubated at 28°C for 48 hours. Puromycin toxicity was assessed as described in the *Drosophila* S2 Cell Manual (Invitrogen). It was determined that 2 μg/mL puromycin was the optimal concentration for selection of stable S2 cells harboring the pMT-Bip-TROP2Ex and pCoPuro plasmids.

Following the recovery of the transformation cultures, a single culture was brought to a final concentration of inducer - 500 μM $CuSO_4$ - and incubated for 48 hours at 28°C. After 48 hours the culture was spun at 3,200 xg for 10 min to pellet the cells. The supernatant was filtered through a 0.22 μM PES filter and stored at 4°C. A portion of the aliquot was assessed via Western blotting with standard protocols. Briefly, the sample was denatured in Laemmli's buffer supplemented to 1.5% β-mercaptoethanol at 1x for 10 minutes at 98°C. The denatured protein was run on a 4-20% Mini-PROTEAN® TGX™ precast gels (Bio-Rad) for 1 hour at 120 V, and the PageRuler Prestained Protein Ladder (10-180 kDa, Thermo-Fisher) was used as a ladder. Once separated on the gel the proteins were transferred to a nitrocellulose membrane using the iBlot™ Gel Transfer Device with iBlot™ Transfer Stacks as per the manufacturer instructions (ThermoFisher). The primary antibody was Anti-6x His tag® horseradish peroxidase conjugate

antibody (abcam, ab1187), and the blot was resolved using the Pierce™ DAB Substrate Kit as described by the manufacturer (ThermoFisher).

The generation of the stable cell line was performed in parallel to the initial testing of TROP2Ex production. Recovered cells were spun at 440 xg for 3 minutes the supernatant was aspirated, the pellet was resuspended in 3 mL of full S2 media + 2 μg/mL puromycin and placed back into the original well. Every 3-4 days the culture was pelleted, the supernatant removed via aspiration, gently resuspended in 3 mL of fresh full S2 media with 2 μg/mL puromycin, placed back into the original well and grown at 28°C. This practice was continued until the cells had grown a density of 6-20x10^6 cells/mL. Once at density the cells were then passaged 1:2 into full S2 media + 2 μg/mL puromycin to 6 mL final volume. At this point the culture was moved into a Corning® 25 cm^2 culture flask with a vented cap (Sigma-Aldrich) and grown at 28°C. Once the cells reached a density of 6-20x10^6 cells/mL they were passaged 1:2 into 12 mL full S2 media with puromycin in a Corning® 75 cm^2 culture flask with a vented cap (Sigma-Aldrich) and grown at 28°C. When the cells were passaged into the 75 cm^2 flask a 6 well plate was also seeded with 3 mL full S2 media + 2 μg/mL puromycin with 1x10^6 cells/mL and TROP2Ex expression was tested as previously described. The stable cell line was cultured in the 75 cm^2 flask until the cells grew to 1-2x10^7 cells/mL, at which time freezer stocks were prepared.

To prepare freezer stocks the cultures were spun at 440 xg for 3 minutes and the conditioned media was collected in a 50 mL falcon tube via careful pipetting. Cells were resuspended in 10 mL room temperature sterile 1x PBS (10 mM Na_2HPO_4 + 2 mM KH_2PO_4 + 2.7 mM KCl + 137 mM NaCl, pH 7.4) and spun as before, the resulting supernatant was removed via aspiration. Freezing buffer was prepared by combining 9 mL conditioned media with 9 mL full S2 media and 2 mL 100% DMSO. The cells were gently resuspended in the room

temperature freezing buffer at $1.1x10^7$ cells/mL and 1 mL aliquots were dispensed into cryogenic storage vials. A freezer box was lined with paper towel and the samples were stored therein, this box was placed into a Styrofoam container and this was stored at -80°C. After 24 hours at -80°C the vials were moved into liquid nitrogen storage. The remaining culture was continually expanded via 1:2 dilutions into fresh full S2 media with puromycin as before. Once at 30 mL final volume, cells were passaged 1:10 into full S2 media with puromycin every ~6-7 days.

Production and Purification of TROP2Ex:

In a sterile 250 mL Erlenmeyer flask, 50 mL dynamic full S2 media (full S2 media + 2 µg/mL puromycin + 0.1% Plurionic® F68 (Sigma-Aldrich) + 1x penicillin/streptomycin) was seeded with stable S2 cells (pMT-Bip-TROP2Ex-pCoPuro) at $1x10^6$ cells/mL, this was grown at 28°C with shaking at 110 rpm. This culture was grown to a density of ~$2-4x10^6$ cells/mL, once there $CuSO_4$ was added to a final concentration 500 µM. Extractions were performed every 24 hours for 120 hours following initiation of TROP2Ex expression. Collected samples were analyzed via Western blotting as described previously.

For larger scale production 300 mL of dynamic full S2 media was seeded and grown as described previously in a 2 L baffled culture flask. Once at ~$2-4x10^6$ cells/mL (~24-48 hours) the culture was induced $CuSO_4$ as before for 96 hours at 28°C with shaking at 110 rpm. After induction the culture was spun at 3000 xg for 30 minutes at 4°C, the supernatant was saved and the pellet discarded. The volume of the supernatant was measured with a graduated cylinder and then filtered through a 0.22 µm PES filter. After filtration, the supernatant was placed into a sterile 1L beaker with a stir bar, placed into an ice bath on a stir plate, and brought to 0.1 mM PMSF. With continuous gentle stirring solid ammonium sulfate was slowly added to the

supernatant until the solution was 40% of the saturation limit for ammonium sulfate at 0°C. The mass of solid ammonium sulfate to add was calculated using a web-based calculator (http://www.encorbio.com/protocols/AM-SO4.htm). After addition of ammonium sulfate the beaker was covered and stirred for 3 hours in the ice bath. After the equilibration, the 40% saturated ammonium sulfate supernatant was spun at 10,000 xg for 15 minutes at 4°C. Following centrifugation the supernatant and precipitate were separated, the pellet was discarded, and the supernatant was moved back into the ice bath. With gentle stirring the 40% saturated supernatant had solid ammonium sulfate slowly added until the solution was 65% of the saturation limit for ammonium sulfate at 0°C . The 65% ammonium sulfate saturated solution was covered and stirred for 1 hour in the ice bath. After the second equilibration the solution was spun at 15,000 xg for 15 minutes at 4°C. The supernatant was discarded, and the pellets were resuspend in a total of 40 mL 1x PBS + 20 mM imidazole, and pooled together.

Batch immobilized metal affinity chromatography (IMAC) was performed using a 1 mL bed volume of Talon® resin (Takara Bio) pre-equilibrated with 1x PBS + 20 mM imidazole. A 3-hour adherence reaction was performed by rocking the samples with the resin in an ice bath. The slurry was pelleted at 700 xg for 2 minutes and the supernatant was discarded. The slurry was then resuspended in 10 mL of 1x PBS with 20 mM imidazole with gentle pipetting, transferred to a 15 mL conical tube, and incubated for 5 minutes with rocking in the ice bath. The resin was pelleted as described above and the supernatant was discarded, this wash step was repeated four more times. To elute TROP2Ex the pelleted resin was resuspended with 1 mL of 1x PBS + 400 mM imidazole with gentle pipetting and incubated for 5 minutes with rocking in an ice bath. The slurry was spun as before and the supernatant was collected and stored in ice, this was repeated

five more times. IMACs purified TROP2Ex was concentrated from ~6 mL to ~2.5 mL with Amicon® Ultra-4 10,000 NMWL centrifugal filters and filtered through a 0.22 μm PES filter.

The IMAC eluent was further purified by FPLC-SEC using a HiLoad 16/600 Superdex 200 pg size exclusion column equilibrated with 1x PBS with a mobile phase flow rate of 1 mL/min. The main peak eluted at 85 mL with a shoulder that eluted at 80 mL, and the main peak and shoulder were not pooled. The collected fractions were filter through a 0.22 μm PES filter and stored at 4°C. Successful purification of the protein was confirmed by Western blotting and the purity of the final product was confirmed using SDS-PAGE stained with SimplyBlue SafeStain as per the manufacturer instructions (ThermoFisher). The purified protein was quantified using both Bradford assays with standard protocols, and by the Edelhoch method with a predicted molar extinction coefficient at A_{280} of 16680 M^{-1} cm^{-1}.

Yeast display of the m7E6 scFv and clonal titration

Yeast surface display of the m7E6 scFv with EBY100 cells was optimized with the optimal displaying conditions discovered to be 30°C for 16-18 hours with shaking at 250 rpm in a 1 mL culture. TROP2Ex was chemically biotinylated using NHS-Ester chemistry and clonal titrations were performed exactly as described by Medina-Cucurella and Whitehead[20].

Acknowledgements

We would like to thank Dr. C. Chan for allowing us to use her laboratory infrastructure and Dr. A. Oak for her training and help with insect cell culturing.

Supporting text

Note S1. DNA sequence of the TROP2Ex gene block inserted into the pMT vector. *D. melanogaster* optimized TROP2Ex gene (underlined text), Kozac sequence (blue text), Bip secretion tag (orange text), KpnI and AgeI restriction sites are bolded and italicized.

GGTACCACCATGGGAATGAAGTTATGCATATTACTGGCCGTCGTGGCCTTTGTTGGC
CTCTCGCTCGGGCATACGGCAGCCCAGGATAATTGTACTTGTCCAACGAATAAAATG
ACGGTATGCTCCCCGGACGGACCAGGCGGTAGGTGTCAATGTCGGGCACTCGGAAG
TGGAATGGCTGTTGACTGTTCCACACTCACGAGTAAATGTTTGCTCCTGAAGGCAAG
GATGAGCGCTCCTAAAAACGCTCGAACTCTCGTCAGGCCAAGTGAGCACGCGCTGG
TCGACAACGACGGCTTGTACGATCCCGATTGCGACCCAGAGGGCAGGTTTAAAGCT
CGCCAGTGCAACCAAACGTCGGTCTGCTGGTGTGTAAACTCCGTGGGTGTACGACGG
ACCGACAAAGGTGATTTGTCCCTGCGTTGTGACGAACTCGTCCGCACGCATCACATA
CTCATTGATCTGAGGCATCGTCCCACCGCTGGAGCGTTCAATCACTCGGACTTGGAC
GCCGAACTGCGCCGATTGTTCCGAGAACGATATAGGTTGCATCCCAAGTTCGTTGCT
GCGGTACATTACGAGCAACCAACCATACAGATAGAATTGAGGCAAAATACCAGTCA
AAAAGCGGCGGGAGATGTAGATATTGGCGATGCAGCCTATTATTTCGAAAGGGATA
TTAAAGGCGAATCGTTGTTCCAAGGTCGAGGCGGACTCGACCTCCGGGTTAGGGGC
GAGCCATTGCAGGTCGAAAGGACTCTGATCTATTACCTCGATGAAATACCTCCGAAG
TTTAGCATGAAGAGGTTG***ACCGGT***

Note S2. DNA sequence of the m7E6 scFv. The sequence for the V_H domain is blue, the sequence of the linker is underlined, the V_L sequence is lower case.

CAGGTCCAGCTGAAGGAAAGCGGCCCCGGCCTGGTGGCCCCTTCCCAGTCTCTGAG
CATCACCTGCACAGTGAGCGGCTTCTCCCTGACCTCTTACGGCGTGCACTGGGTGCG
GCAGCCTCCTGGCAAGGGCCTGGAGTGGCTGGGCGTGATCTGGACCGGCGGCTCCA
CAGATTATAACTCTGCCCTGATGAGCCGGCTGTCCATCAACAAGGACAATTCCAAGT
CTCAGGTGTTTCTGAAGATGAATAGCCTGCAGACCGACGATACAGCCATGTACTATT
GTGCCCGGGACGGCGATTACGACAGATATACCATGGATTACTGGGGCCAGGGCACC
AGCGTGACAGTGAGCGGTGGAGGGGGTTCAGGCGGGGGTGGAAGCGGTGGAGGGG
GTAGCagtgatatcgtgctgactcagtcccctgcttcactggccgtgagcctgggccagagagctacaatctcctgcagagcctccaag
tctgtgagcacctccggctacagctatatgcactggtaccagcagaagccaggccagccccctaagctgctgatctatctggcctctaacct
ggaaagcggcgtgcctgctcggttctctggcagcggctccggcacagactttaccctgaatattcacccagtggaggaggaggatgccgc
cacatactattgccagcactccagggagctgccctacacattcggcggcggcaccaagctggagatcaag

Note S3. Secreted TROP2Ex amino acid sequence.

HTAAQDNCTCPTNKMTVCSPDGPGGRCQCRALGSGMAVDCSTLTSKCLLLKARMSAP
KNARTLVRPSEHALVDNDGLYDPDCDPEGRFKARQCNQTSVCWCVNSVGVRRTDKGD
LSLRCDELVRTHHILIDLRHRPTAGAFNHSDLDAELRRLFRERYRLHPKFVAAVHYEQPT
IQIELRQNTSQKAAGDVDIGDAAYYFERDIKGESLFQGRGGLDLRVRGEPLQVERTLIYY
LDEIPPKFSMKRLTGHHHHHH

Note S4. **Amino acid sequence of the m7E6 scFv.**

QVQLKESGPGLVAPSQSLSITCTVSGFSLTSYGVHWVRQPPGKGLEWLGVIWTGGSTDY
NSALMSRLSINKDNSKSQVFLKMNSLQTDDTAMYYCARDGDYDRYTMDYWGQGTSV
TVSGGGGSGGGGSGGGGSSDIVLTQSPASLAVSLGQRATISCRASKSVSTSGYSYMHWY
QQKPGQPPKLLIYLASNLESGVPARFSGSGSGTDFTLNIHPVEEEDAATYYCQHSRELPY
TFGGGTKLEIK

Note S5. Primers for overhang PCR for forming the m7E6 scFv and introduction of the cut sites for cloning into pETconNK.

V_H forward:
5'- AGCGGAGGCGGAGGGTCGGCTAGCCATATGCAGGTCCAGCTGAAGGAAAG -3'

V_H reverse:
5'-CTCCACCGCTTCCACCCCCGCCTGAACCCCCTCCACCGCTCACTGTCACGCTGGT -3'

V_L forward:
5'-
AGGCGGGGGTGGAAGCGGTGGAGGGGGTAGCAGTGATATCGTGCTGACTCAGTCCC -3'

V_L reverse:
5'- GAAATAAGCTTTTGTTCGGATCCGCCCCCCTCGAGCTTGATCTCCAGCTTGGTGC -3'

REFERENCES

REFERENCES

1. Fong D, Moser P, Krammel C, Gostner JM, Margreiter R, Mitterer M, Gastl G, Spizzo G: **High expression of TROP2 correlates with poor prognosis in pancreatic cancer**. *Brit. J. Cancer* 2008, 99:1290–1295.

2. Wu M, Liu L, Chan C: **Identification of novel targets for breast cancer by exploring gene switches on a genome scale**. *BMC Genomics* 2011, 12:547.

3. McDougall ARA, Tolcos M, Hooper SB, Cole TJ, Wallace MJ: **Trop2: from development to disease**. *Dev. Dynam.* 2015, 244:99–109.

4. Lipinski M, Parks DR, Rouse RV, Herzenberg LA: **Human trophoblast cell-surface antigens defined by monoclonal antibodies**. *Proc. Natl. Acad. Sci. USA* 1981, 78:5147–5150.

5. Wang J, Day R, Dong Y, Weintraub SJ, Michel L: **Identification of Trop-2 as an oncogene and an attractive therapeutic target in colon cancers**. *Mol. Cancer Ther.* 2008, 7:280–285.

6. Cubas R, Zhang S, Li M, Chen C, Yao Q: **Trop2 expression contributes to tumor pathogenesis by activating the ERK MAPK pathway**. *Mol. Cancer* 2010, 9:253.

7. Trerotola M, Cantanelli P, Guerra E, Tripaldi R, Aloisi AL, Bonasera V, Lattanzio R, Lange R, Weidle UH, Piantelli M, Alberti S: **Upregulation of Trop-2 quantitatively stimulates human cancer growth**. *Oncogene* 2013, 32:222–233.

8. Trerotola M, Jernigan DL, Liu Q, Siddiqui J, Fatatis A, Languino LR: **Trop-2 Promotes Prostate Cancer Metastasis By Modulating β1 Integrin Functions**. *Cancer Res.* 2013, 73:3155–3167.

9. Kowalsky CA, Faber MS, Nath A, Dann HE, Kelly VW, Liu L, Shanker P, Wagner EK, Maynard JA, Chan C, Whitehead TA: **Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing**. *J. Biol. Chem.* 2015, 290:26457-26470.

10. Stoyanova T, Goldstein AS, Cai H, Drake JM, Huang J, Witte ON: **Regulated proteolysis of Trop2 drives epithelial hyperplasia and stem cell self-renewal via β-catenin signaling**. *Genes Dev.* 2012, 26:2271–2285.

11. Pavšič M, Ilc G, Vidmar T, Plavec J, Lenarčič B: **The cytosolic tail of the tumor marker protein Trop2 – a structural switch triggered by phosphorylation**, *Sci. Rep.* 2015, 5:10324.

12. Trerotola M, Li J, Alberti S, Languino LR: **Trop-2 inhibits prostate cancer cell adhesion to fibronectin through the β 1 integrin-RACK1 axis**. *J. Cell Physiol.* 2012, 227:3670–3677.

13. Trerotola M, Ganguly KK, Fazli L, Fedele C, Lu H, Dutta A, Liu Q, De Angelis T, Riddell LW, Riobo NA, Gleave ME, Zoubeidi A, Pestell RG, Altieri DC, Languino LR: **Trop-2 is up-regulated in invasive prostate cancer and displaces FAK from focal contacts**. *Oncotarget* 2015, 6:14318-14328.

14. Litvinov SV, Velders MP, Bakker HA, Fleuren GJ, Warnaar SO: **Ep-CAM: a human epithelial antigen is a homophilic cell-cell adhesion molecule**. *J. Cell Biol.* 1994, 125:437–446.

15. Vidmar T, Pavšič M, Lenarčić B: **Biochemical and preliminary X-ray characterization of the tumor-associated calcium signal transducer 2 (Trop2) ectodomain**. *Protein Expres. Purif.* 2013, 91:69-76.

16. Pavšič M, Gunčar G, Djinović-Carugo K, Lenarčič B: **Crystal structure and its bearing towards an understanding of key biological functions of EpCAM**. *Nat. Commun.* 2014, 5:4764.

17. Liu S, Ho W, Strop P, Dorywalska MG, Rajpal A, Shelton DL, Tran T: **Antibodies specific for TROP-2 and their uses**. *US Patent Application* 2013.

18. Klesmith JR, Bacik J, Wrenbeck EE, Michalczyk R, Whitehead TA: **Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning**. *Proc. Natl. Acad. Sci. U S A* 2017*,* 114:2265-2270.

19. Iwaki T, Figuera M, Ploplis VA, Castellino FJ: **Rapid selection of *Drosophila* S2 cells with the puromycin resistance gene**. *BioTechniques* 2003, 35:482-486.

20. Medina-Cucurella AV, Whitehead TA: **Characterizing protein-protein interactions using deep sequencing coupled to yeast surface display**. *Methods in Mol. Biol.* 2018, 1764:101-121.