DEEP SEQUENCING DRIVEN PROTEIN ENGINEERING: NEW METHODS AND APPLICATIONS IN STUDYING THE CONSTRAINTS OF FUNCTIONAL ENZYME EVOLUTION

By

Emily Elizabeth Wrenbeck

A DISSERTATION

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

Chemical Engineering – Doctor of Philosophy

2017

ABSTRACT

DEEP SEQUENCING DRIVEN PROTEIN ENGINEERING: NEW METHODS AND APPLICATIONS IN STUDYING THE CONSTRAINTS OF FUNCTIONAL ENZYME EVOLUTION

By

Emily Elizabeth Wrenbeck

Chemical engineers have long sought enzymes as alternatives to traditional chemocatalytic routes as they are highly selective and have evolved to function under mild conditions (physiological temperature, neutral pH, and atmospheric pressure). Enzymes, the workhorses of biological chemistry, represent a vast catalogue of chemical transformations. This feature lends their use in a variety of industrial applications including food processing, biofuels, engineered biosynthetic pathways, and as biocatalysts for preparing specialty chemicals (e.g. pharmaceutical building blocks). The totality of an enzymatic bioprocess is a function of its catalytic efficiency (specificity and turnover), product profile (i.e. regio- and enantio-selectivity), and thermodynamic and kinetic stability. For native enzymes, these parameters are seldom optimal. Importantly, they can be modified using protein engineering techniques, which generally involves introducing mutation(s) to a protein sequence and screening for beneficial effects. However, robust enzyme engineering and design based on first principles is extremely challenging, as mutations that improve one parameter often yield undesired tradeoffs with one or more other parameters.

In this thesis, deep mutational scanning - the testing of all possible single-amino acid substitutions of a protein sequence using high-throughput screens/selections and DNA counting via deep sequencing - was used to address two fundamental constraints on functional enzyme evolution. First, how do enzymes encode substrate specificity? To address this question, deep mutational scanning of an amidase on multiple substrates was performed using growth-based selections. Comparison of the resulting datasets revealed that mutations benefiting function on a

given substrate were globally distributed in both protein sequence and structure. Additionally, our massive datasets permitted the most rigorous testing to date of theoretical models of adaptive molecular evolution. These results have implications for both design of biocatalysts and in understanding how natural enzymes function and evolve.

Another fundamental constraint of enzyme engineering is that mutations improving stability (folding probability) of an enzyme are often inactivating for catalytic function, and vice versa. Towards overcoming this activity-stability constraint, I sought to improve the heterologous expression and maintain the catalytic function of a Type III polyketide synthase from *Atropa belladonna*. This was accomplished using deep mutational scanning and high-throughput GFP-fusion stability screening, followed by novel filtering methods to only accept beneficial mutations with high probability for maintaining function.

Lastly, deep mutational scanning relies on the construction of user-defined DNA libraries, however current available techniques are limited by accessibility or poor coverage. To address these limitations, I will present the development of Nicking Mutagenesis, a new method for the construction of comprehensive single-site saturation mutagenesis libraries that requires only double-stranded plasmid DNA as input substrate. This method has been validated on several gene targets and plasmids and is currently being used in academic, government, and industry laboratories worldwide.

I dedicate this thesis to the family and friends who have inspired, encouraged, and delighted in my pursuit of understanding our world through science.

ACKNOWLEDGMENTS

I want to acknowledge the significant impact my graduate advisor, Tim Whitehead, has had on my life and scientific career. Thank you for training me how to think quantitatively, design clever experiments, and for initiating my journey in a lifelong obsession with proteins. Thank you for your assistance in acquiring various fellowships and national conference experiences. I express my deepest appreciation for your patience, support, and for challenging me to do my best. To my labmates – Caitlin Stein, Justin Klesmith, James Stapleton, Matthew Faber, and Carolyn Haarmeyer – thank you for your mentorship, support, and for sharing laughs. I want to thank the Plant Biotechnology for Health and Sustainability graduate training program for providing financial support and an enriching graduate experience.

Lastly, I want to thank my parents for striving to provide me with opportunities to nourish my brain and for their never-ending love and support. My siblings, for shaping who I am. My friends, for thinking of me as cool for being a scientist. And finally, my husband Phil, for always believing me and for your patience, love, and support throughout my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

Ab, *Atropa belladonna*

ACT, acetamide

CSM, comprehensive single-site saturation mutagenesis

DFE, distribution of fitness effects

dsDNA, double-stranded DNA

dU-ssDNA, uracil-containing single-stranded DNA

FACS, fluorescence assisted cell sorting

GFP, green fluorescent protein

GPD, generalized Pareto distribution

HTS, high-throughput screen or selection

IB, isobutyramide

kRBS, knockdown ribosome binding sequence

MPO, methyl-putrescine amine oxidase

NGS, next-generation sequencing

NS, nonsynonymous

PKS, polyketide synthase

PMT, putrescine methyltransferase

PR, propionamide

PSSM, position-specific scoring matrix

RBS, ribosome binding sequence

TA, tropane alkaloids

TRI, tropinone reductase

TS, tropinone synthase

CHAPTER ONE

Introduction to deep sequencing driven protein engineering

Portions of this chapter were adapted with permission from "Deep sequencing methods for protein engineering and design" in *Current Opinion in Structural Biology* 45 (2017) 36-44 by Emily E. Wrenbeck, Matthew S. Faber, and Timothy A. Whitehead.

ABSTRACT

The advent of next-generation sequencing (NGS) has revolutionized protein science, and the development of complementary methods enabling NGS-driven protein engineering have followed. In general, these experiments address the functional consequences of thousands of protein variants in a massively parallel manner using genotype-phenotype linked high-throughput functional screens followed by DNA counting via deep sequencing. We highlight the use of information rich datasets to engineer protein molecular recognition. Examples include the creation of multiple dual-affinity Fabs targeting structurally dissimilar epitopes and engineering of a broad germline-targeted anti-HIV-1 immunogen. Additionally, we highlight the generation of enzyme fitness landscapes for conducting fundamental studies of protein behavior and evolution. We conclude with discussion of technological advances.

INTRODUCTION

Researchers have been engineering proteins for almost 4 decades. Early endeavors involved generation of a handful of point mutations followed by low-throughput assays for function; the 'search space' a protein scientist could feasibly explore was miniscule.

As demonstrated by the seminal works of Fowler et al.[1] and Hietpas et al.[2], the advent of next-generation sequencing (NGS) has presented protein engineers with the ability to economically observe *entire populations* of molecules before, during, and after a high-throughput screen or selection for function (HTS) (**Figure 1.1**). A typical NGS run provides sufficient sequencing data to permit the study of millions of protein variants. Thus, when coupled to HTS, NGS significantly expands the accessible mutational search space. In this way, a researcher can test all possible point mutations or combinations of mutations, for example, and remove the duty of having to design small focused libraries that may miss unpredictable beneficial mutations. As a testimonial to the accessibility of these methodologies, experiments can be performed in a beginning graduate-level course[3].

The intent of this review is to highlight examples where deep sequencing has been applied in different areas of protein engineering and design. As such, we will not provide a comprehensive review of directed evolution or of deep mutational scanning (excellent reviews can be found here[4,5]). We will discuss the use of NGS for engineering protein molecular recognition, membrane proteins, and enzymes, highlight recent technological advances, and offer a perspective on the shape of the field over the next several years.

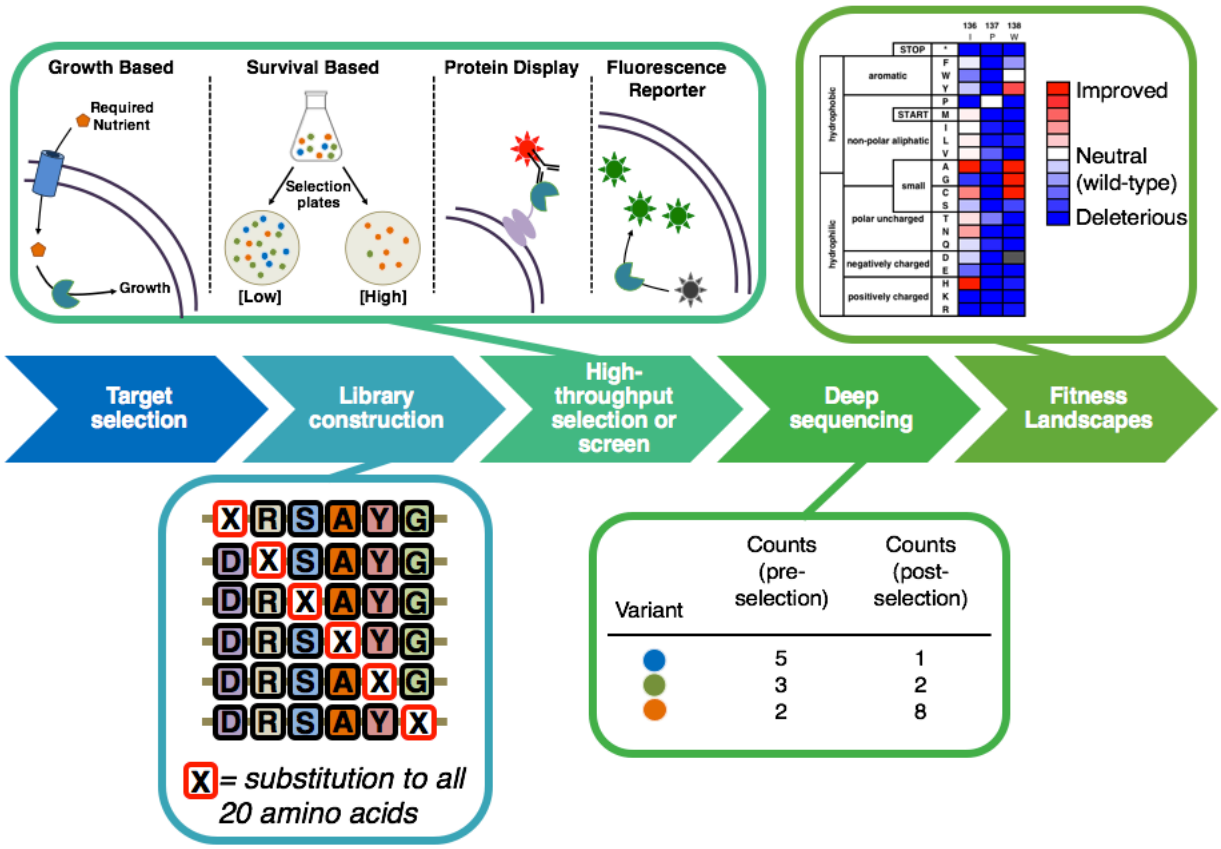


Figure 1.1: Overview of the steps involved in deep mutational scanning. A library of protein variants is generated. Often this is a comprehensive single-site saturation mutagenesis library. The library is subjected to a high-throughput selection or screen for function. Examples of commonly used selections and screens include survival or competitive growth-based selections, protein binding screens like phage or yeast surface display, and fluorescence reporter-based screens. Variants are quantified in the pre- and post-selection populations with counting via deep sequencing. These pre- and post-selection counts are transformed to a normalized functional score and are used to generate fitness landscapes of the target protein.

ENGINEERING PROTEIN MOLECULAR RECOGNITION

Dozens of studies over the past five years have used deep sequencing to identify and engineer protein-ligand interactions. Rapid adoption of deep sequencing by this field is a direct result of mature display-based technologies that can be used to screen very large initial libraries. For example, in the study of protein-protein binding interactions a library of protein variants can

be displayed on the surface of yeast using yeast surface display (**Figure 1.1**). Using a fluorescently conjugated protein binding partner FACS can be used (a thorough review can be found here [6]).

Deep sequencing for screening protein binder libraries

NGS is now frequently used in the evaluation of synthetic or natural libraries to identify antigen-specific binders. Advances in pairing V_H and V_L sequences from individual B cells[7] allows one to identify antigen-specific antibodies directly from sequencing, including panels of antibodies targeting Ebola virus[8] and ricin[9]. Methodological details and limitations associated with identification of rare clones and evaluation of library diversity are presented in a recent review[10].

As an emerging area, engineers now use NGS to refine protein binder libraries[11,12]. In a notable advance, Woldring et al. screened a hydrophilic fibronectin domain library to bind various protein targets[12]. The researchers exploited the site-specific amino acid preferences from an initial library to develop a more focused second library depleted in mutations at the periphery of the binder paratope. Compared to other libraries, this library design afforded far superior performance in isolation of high affinity, stable binders.

Paratope optimization for affinity and specificity

NGS can be used to rapidly improve the affinity and specificity of the binding paratope (**Figure 1.2**)[13,14]. A crucial advantage enabled by NGS is the ability to discriminate very small beneficial changes in binding - on the order of 0.1 kcal/mol or about a 20% improvement in dissociation constant. These small-scale beneficial mutations can be additive, allowing one to "leapfrog" over potential affinity maturation bottlenecks by combining mutations.

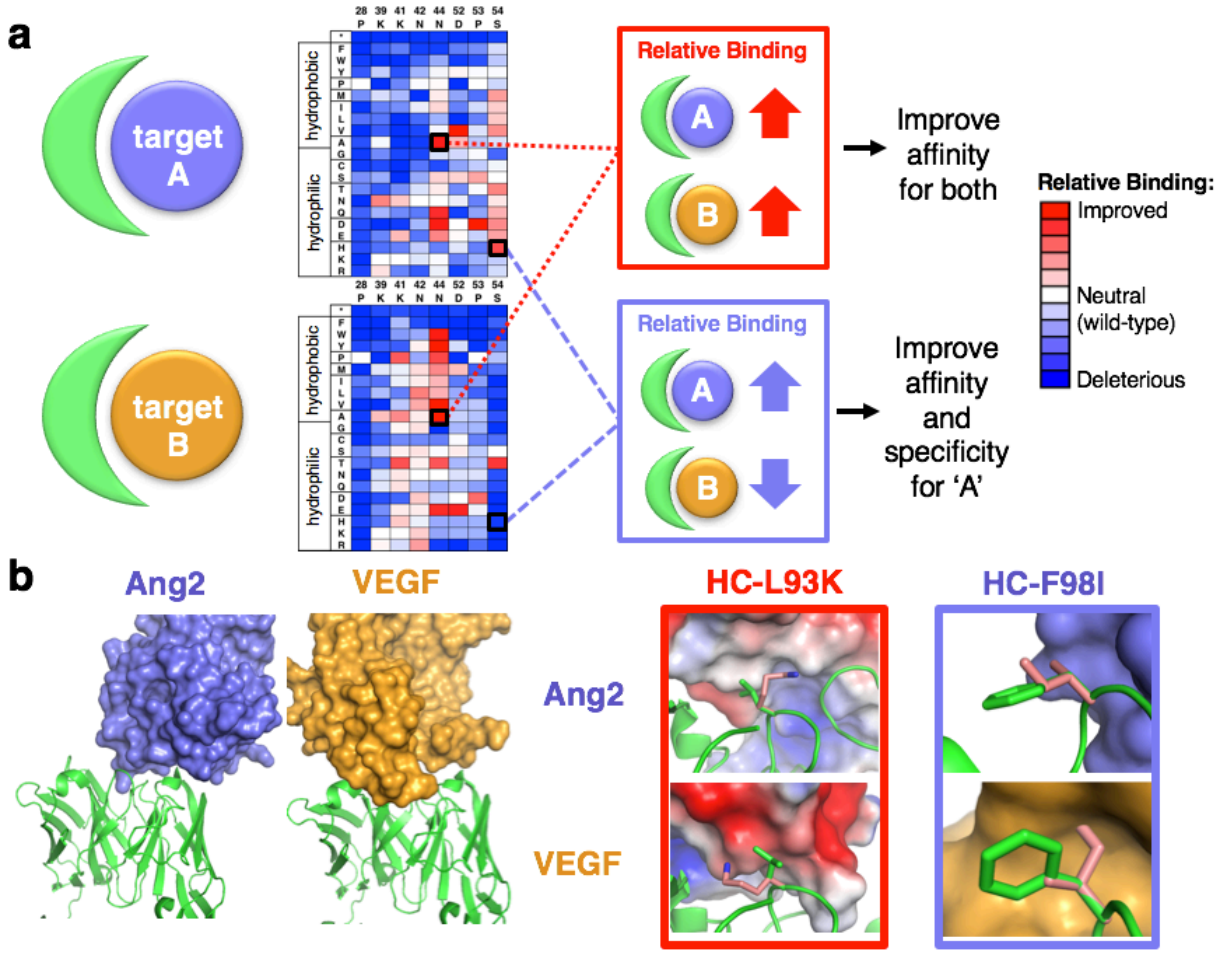


Figure 1.2: Engineering of affinity and specificity in protein-ligand interactions using deep mutational scanning. A.) Consider a protein binder that recognizes two separate targets A and B. Deep mutational scanning is performed against each target in parallel. Site-specific preferences for the protein against each target are visualized by a heatmap. Mutations can be combined to impart binders with greater affinity to both targets (top panel, red box) or restrict specificity to a single target (bottom panel, blue box). In practice, mutations at multiple positions are combined to make a focused library that is subsequently screened. B.) The structural basis for specificity- and affinity- altering mutations identified by deep mutational scanning using a dual action Fab (green cartoon) to Ang2 (purple surface) and VEGF (orange surface) as an example[15]. Heavy Chain (HC) L93K can increase affinity to both targets presumably by increasing electrostatic complementarity. Here Ang2 and VEGF are colored by electrostatic surface potential and HC-L93 (green) and HC-K93 (pink) are shown as sticks. By contrast, HC F98I is strongly depleted for in the VEGF binding population most likely because of steric clashes. Structures were created using PyMol from the PDB IDs 4ZFG, 4ZFF.

Whitehead et al. provide the first example of paratope engineering for affinity and specificity using deep sequencing[16]. The researchers screened a comprehensive single-site saturation mutagenesis library of two *de novo* designed Influenza Hemagglutinin (HA) binders against H1 and H5 HA subtypes. Engineering specificity was demonstrated by comparing site-specific preferences for H1 to the H5 subtype. A single point mutation was identified that gave over a 30-fold specificity switch from the parental designed protein. For affinity maturation, site-specific preferences were encoded into a second library and sorted to improve affinity against both subtypes by approximately 25-fold. The affinity of one designed HA binder, HB36.6, was further improved against seven diverse HA subtypes. HB36.6 showed prophylactic and therapeutic efficacy against lethal challenge of pandemic Influenza in a BALB/c mouse model[17].

Deep mutational scanning approaches have been extended to affinity mature antibodies[18,19]. In an impressive demonstration, Genentech scientists engineered a dual action Fab for high affinity for two unrelated proteins simultaneously[15]. The group used phage display to profile a single and triple site saturation mutagenesis library of a Fab with low nanomolar binding to Ang2 and VEGF. NGS revealed significant site-specific amino acid preferences for each of the two binding paratopes. The researchers combined mutations shown to improve affinity on at least one target and not negatively impact binding on the other target, thus engineering five different sub-nanomolar dual-affinity Fabs.

The apotheosis of deep mutational scanning to identify high affinity binders with defined specificity comes from Jardine et al.[20], who engineered an HIV immunogen that can be recognized by B cell precursors to broadly neutralizing anti-HIV antibodies. Starting with a designed outer domain of the gp120 protein from HIV, they screened a 58-residue site saturation mutagenesis library against 18 germline-reverted and 11 VRC01-class broadly neutralizing antibodies.

Information obtained from the scan was used to encode a second library that was screened against the same antibody panel. One variant showed dramatically improved binding to all antibodies in the panel and could bind naïve B cells in full human repertoires.

Binding surface optimization is not limited to protein-protein binders, provided that there is a suitable HTS. Tinberg et al. used yeast display coupled to NGS to affinity mature a computationally designed anti-steroid binder[21]. Raman and colleagues used an *in vivo* fluorescent reporter coupled to FACS (**Figure 1.1**) to engineer the *E. coli* allosteric transcription factor LacI to recognize four different non-metabolizable inducers, including sucralose[22].

Epitope mapping

An important consideration for the antibody engineer is the identification of the binding epitope. Three recent publications used yeast surface display, site-saturation mutagenesis, FACS, and deep sequencing to identify conformational epitopes for diverse antigenic targets on the order of weeks[23–25]. Doolan and Colby determined epitope regions on prions recognized by conformational-specific antibodies[23]. Van Blarcom et al. performed epitope mapping for a panel of antibodies against the alpha toxin from methicillin-resistant *Staphylococcus aureus*[24]. Kowalsky et al. automated and improved the speed of epitope identification for three different antigens[25].

MEMBRANE PROTEIN ENGINEERING

Plückthun and colleagues screened a near-comprehensive single point mutant library of G protein-coupled receptor (GPCR) rat neurotensin receptor 1 for enhanced heterologous expression, a proxy for protein stability. The library was expressed in the periplasm of *E. coli* and sorted by FACS using a fluorescently conjugated agonist as a probe[26]. NGS was used to quantify variants in

the input library and the enriched FACS selected libraries, and hits identified in the initial library were combined, resulting in variants that express at up to 50-fold higher levels in *E. coli* compared with the wild-type GPCR. Each stability-enhancing mutation contributed a small amount of the overall stability to the protein[27]. Notably, the structure of an engineered GPCR was solved[28], suggesting a general directed evolution strategy of stabilizing membrane proteins for X-ray crystallography structure determination. In a separate effort, Fleishman and colleagues used deep mutational scanning to unravel the energetics associated with membrane protein insertion and homodimerization revealing insights that may facilitate membrane protein design[29].

ENZYME ENGINEERING

In contrast to protein-ligand interactions, the complex and diverse nature of enzyme function has made it challenging to develop robust, sensitive, and generalizable functional screens. As such, far fewer examples of deep sequencing-assisted enzyme engineering exist in the literature (**Table 1.1**).

High-throughput screening and selection for enzyme function

The primary strategy for functional selection of enzymes is to tether enzymatic function to the growth and/or survival (fitness) of a host organism. One type of competitive growth selection is to provide a substrate that the enzyme must catabolize as the sole source of an essential element for growth (carbon, nitrogen) (**Figure 1.1**). Thus, variants enabling higher flux through and enzyme permit faster growth rates and become enriched in the population. Klesmith et al. performed deep mutational scanning of levoglucosan kinase, where levoglucosan was fed as the carbon source[30]. Similarly, Wrenbeck et al. performed deep mutational scanning on amiE, an

aliphatic amidase from *Pseudomonas aeruginosa*, by feeding amides as the nitrogen source[31]. Antibiotic resistance genes also provide straightforward targets for competitive growth selections. Indeed, these represent 4/9 published enzyme scans (**Table 1.1**)[32–34]. In summary, high-throughput screens or selections that are *generalizable* are desired, yet the incredible diversity of enzyme function makes their development a critical challenge for the field.

Table 1.1: NGS-assisted studies of large enzyme libraries.

Gene	Application	Selection employed	Reference
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Deng et al.[32]
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Firnberg et al.[33]
TEM-1 β-lactamase	β-lactam antibiotic resistance	Growth competition	Stiffler et al.[34]
APH(3')II kinase	aminoglycoside antibiotic resistance	Growth competition	Melnikov et al.[35]
Homing endonucleases	Genome engineering	Survival	Thyme et al.[36]
Levoglucosan kinase	Biomass conversion	Metabolic growth	Klesmith et al.[30]
amiE aliphatic amidase	Multiple industrial	Metabolic growth	Wrenbeck et al.[31]
Bgl3 β-glucosidase	Biomass conversion	Micro-fluidic	Romero et al.[37]
Ube4b E3 ubiquitin ligase	E3 ubiquitin ligase	Phage display	Starita et al.[38]

From fitness landscapes to enzyme engineering

Deep mutational scanning experiments afford a richness of knowledge of 'hits'. However, efficiently utilizing ambiguous 'fitness values' to inform enzyme design is still a significant

challenge. To avert this challenge, van der Meer et al. performed over 4000 assays to generate 'mutability landscapes' of a tautomerase enzyme for its expression, Michael-type activities on multiple substrates, and characterization of its enantioselectivity, and used this information to design a novel enantioselective Michaelase[39].

How does one intelligently combine hits to achieve a given design goal? One approach is to biophysically characterize beneficial mutations. For example, Klesmith et al. performed deep mutational scanning of levoglucosan kinase to identify mutations that improved fitness through improved flux of levoglucosan conversion. They characterized a set of beneficial mutations for activity and thermodynamic stability and used this information to generate designs, one of which had greater than 24-fold improvement in activity and 7°C increase in apparent melting temperature[30]. An alternative approach is to generate multiple fitness landscapes under different conditions (concentration and identity of substrate, temperature, etc.) and use differential analysis to generate designs. To that end Melnikov et al. performed deep mutational scanning of APH(3')II, an enzyme responsible for aminoglycoside antibiotic resistance, with several antibiotics at different concentrations and generated designs with orthogonal activities[35].

Datasets from deep mutational scanning can be used to probe the fundamental nature of enzyme behavior and can be used to ask questions related to evolutionary trajectories, rigorously testing theories gleaned from over two decades of directed evolution experiments. Steinberg and Ostermeier analyzed fitness effects for TEM-15 β-lactamase under varying environmental conditions and found that negative selections were able to bridge access to the highest fitness peaks[40]. Wrenbeck et al. performed deep mutational scanning of an aliphatic amidase on three substrates and found that specificity-determining mutations were distributed throughout the protein sequence and structure rather than located near the active site[31].

METHODOLOGICAL ADVANCES AND CURRENT LIMITATIONS

Mutagenic library preparation

Consider a protein of a typical length of 300 residues. A library comprising every possible single or double point mutation would contain $6x10^3$ or $3.6x10^7$ sequences, respectively. Similarly, a library with simultaneous saturation mutagenesis at four defined positions contains $1.6x10^5$ sequences. For a typical experimental workflow there are 10^6-10^7 quality-filtered DNA reads, and accurate estimation of variant frequencies occurs above a statistical background of ~100 sequence reads per variant[41,42]. Dividing the number of sequences from a NGS run by the minimum number needed to estimate frequencies we arrive at an effective maximum population size of 10^4-10^5 per experiment. Thus, even NGS permits only small dances around the local protein sequence-fitness space.

Purchasing thousands to millions of synthetically generated DNA sequences is still not an economically viable option for the average academic lab. Furthermore, established facile protocols for random mutagenesis like error-prone PCR[43] or chemical synthesis by doping[1] provide access only to a minority of possible codon substitutions, and there is often a large variance in the number of mutations introduced. Thus, robust methods for constructing large, user-defined DNA libraries are needed.

Generation of libraries with mutations at 1-4 defined positions have been demonstrated using homologous recombination and cassette mutagenesis. For applications such as lead candidate maturation the generation of comprehensive single-site saturation mutagenesis (CSM) libraries is desired. A CSM library contains all possible single amino acid substitutions at every position in the primary sequence. One could generate such libraries by performing separate saturation mutagenesis reactions for each position using QuikChange or similar methods.

However, there are now three methods that can generate CSM libraries for gene-length targets with a single reaction: PALS[44], PFunkel[45], and Nicking Mutagenesis[46]. In PFunkel mutagenesis, single mutants are generated by thermocycling mutagenic oligos with template DNA at a low primer:template ratio in a single test-tube. While PFunkel has been demonstrated on multiple systems with excellent performance[30,33,42] the method requires a bacteriophage preparation of a Uracil-containing ssDNA template, which can be laborious. To overcome this, Wrenbeck et al. developed a similar method, Nicking Mutagenesis, which uses plasmid dsDNA as the reaction template[46].

DNA read length restrictions

One major limitation of NGS is the inherent short read length (75 to 300 nucleotides for Illumina sequencing platform) (**Figure 1.3a**). As such, a mutation located outside of the read window would be invisible. Longer read lengths are possible using PacBio and Oxford Nanopore instruments but at the cost of reduced throughput and accuracy, respectively. Because of these limitations, many groups perform deep mutational scanning on small genes or on subsets of genes (tiling) (**Figure 1.3b**)[25,27,30,34,42,47].

An emerging strategy is to perform a selection on a full-length gene but 'link' or phase haplotypes from one portion of the gene to the remainder (**Figure 1.3c**)[44,48–54]. For example, Sarkisyan et al. introduced a random 20-nucleotide barcode at the C-terminal end of a library of green fluorescent protein variants whilst performing error-prone PCR[54]. Genotypes were barcode linked by sequencing both the N- and C- termini, with the N-terminus brought into proximity of the barcode with successive digestion and ligation reactions.

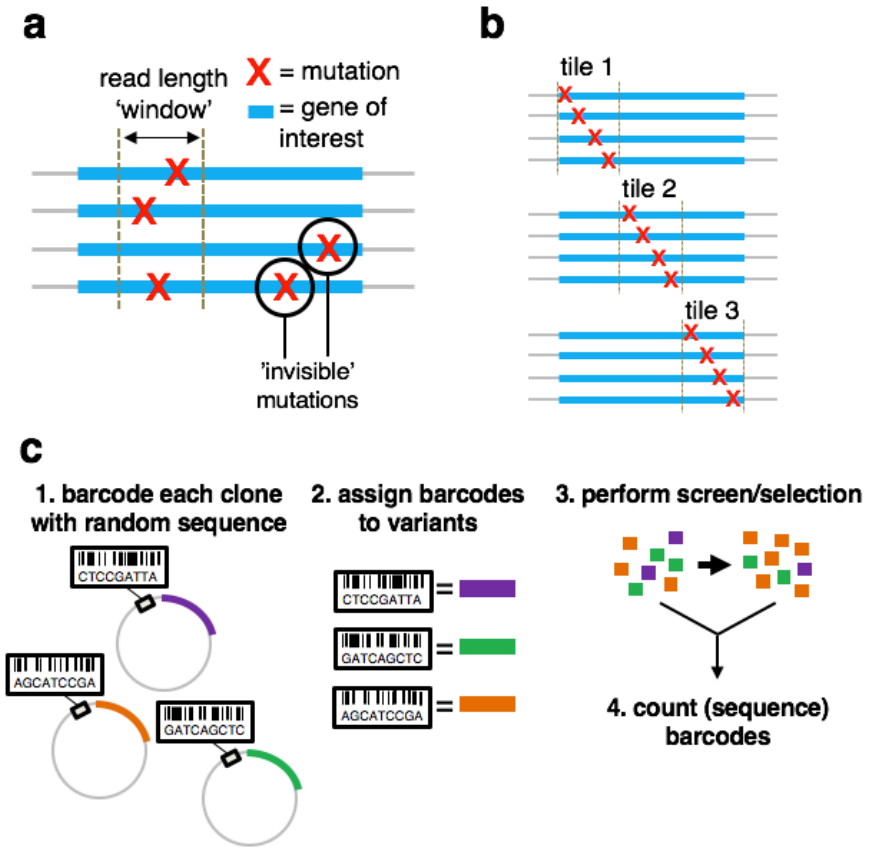


Figure 1.3: Strategies to overcome read length limitations of NGS. A.) Mutations falling outside of a length 'readable' by current sequencing technologies would be invisible. B.) In a gene tiling approach, mutational libraries are prepared such that mutations are restricted to a stretch of DNA readable by NGS platforms. Parallel screens or selections for function are performed. C.) Molecular barcoding of library members provides a means to overcome NGS sequencing read length restrictions. Randomized DNA barcodes are assigned to library member (1). Variants and their corresponding barcodes are linked and cataloged (haplotyped) (2). After functional selection (3), variants in the pre- and post-selection populations are counted by sequencing barcodes (4).

Sequencing analysis

A crucial step in any NGS-utilizing experiment is to extract useful phenotypic data - binding, kinetics, thermodynamic stability, host organismal fitness, etc. - from raw sequencing reads. Many groups report site-specific preferences as an enrichment ratio. To that end, Fowler et al. developed Enrich, a python-based software that transforms raw sequencing counts from pre-

and post-selection populations into per-allele enrichment ratios[55]. Similarly, Bloom developed a software that calculates enrichments using a likelihood-based treatment of mutation counts instead of simple ratios[56]. Woldring et al. developed ScaffoldSeq, a Python-based software for the analysis of partially diverse protein sequences for single site and pairwise amino acid frequencies across the population[57].

Normalization of these enrichment ratios to an unambiguous fitness metric like binding or catalytic efficiency is perhaps the least standardized portion of the deep mutational scanning pipeline and there is a need for a community-wide consensus on how to normalize. Kowalsky et al. describe a mathematical framework for normalizing enrichment ratios of variants assayed in deep mutational scanning experiments for FACS and growth-based selections[42]. Similar approaches are used for plate-based selections[33]. Finally, Abriata et al. developed a webserver, PsychoProt, for the analysis of functional data from saturation mutational libraries and protein sequence alignments for biophysical constraints using structural information[58].

CONCLUSION

NGS has been a transformative technology for many fields in the biological sciences, with protein science and engineering being no exception. Generation and analysis of fitness landscapes can inform on mechanisms of natural evolution and fundamentals of enzyme behavior. Notable advances in our ability to engineer affinity and specificity in protein-ligand interactions has been enabled by NGS, while enzyme engineering has lagged behind largely because of the lack of generalized HTS strategies. For this same reason, the application of NGS to membrane protein engineering has even further lagged behind. The utility of NGS enabled enzyme and membrane protein engineering awaits screening technology breakthroughs. Accurate and facile sequencing

of non-contiguous mutations (haplotyping), either through the use of barcoding or the advent of longer-read technologies, will improve and expand the utility of NGS protein engineering.

REFERENCES

REFERENCES

1. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

2. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **108,** 7896–7901 (2011).

3. Mavor, D. *et al.* Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife* **5,** e15802 (2016).

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

5. Boucher, J. I., Bolon, D. N. A. & Tawfik, D. S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Sci.* (2016). doi:10.1002/pro.2928

6. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17,** 467–73 (2007).

7. Dekosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31,** 166–169 (2013).

8. Wang, B. *et al.* Facile discovery of a diverse panel of anti-Ebola virus antibodies by immune repertoire mining. *Sci. Rep.* **5,** (2015).

9. Wang, B. *et al.* Discovery of high affinity anti-ricin antibodies by B cell receptor sequencing and by yeast display of combinatorial VH: VL libraries from immunized animals. *MAbs* (2016). doi:10.1080/19420862.2016.1190059

10. Glanville, J. *et al.* Deep sequencing in library selection projects: What insight does it bring? *Curr. Opin. Struct. Biol.* **33,** 146–160 (2015).

11. Mahon, C. M. *et al.* Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J. Mol. Biol.* **425,** 1712–1730 (2013).

12. Woldring, D. R., Holec, P. V, Zhou, H. & Hackel, B. J. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* **10,** e0138956 (2015).

13. Strauch, E.-M., Fleishman, S. J. & Baker, D. Computational design of a pH-sensitive IgG

binding protein. *Proc. Natl. Acad. Sci.* **111,** 675–680 (2014).

14. Procko, E. *et al.* A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* **157,** 1644–1656 (2014).

15. Koenig, P. *et al.* Deep Sequencing-guided Design of a High Affinity Dual Specificity Antibody to Target Two Angiogenic Factors in Neovascular Age-related Macular Degeneration. *J. Biol. Chem.* **290,** 21773–21786 (2015).

16. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–8 (2012).

17. Koday, M. T. *et al.* A Computationally Designed Hemagglutinin Stem-Binding Protein Provides In Vivo Protection from Influenza Independent of a Host Immune Response. *Plos Pathog.* **12,** e1005409 (2016).

18. Forsyth, C. M. *et al.* Deep mutational scanning of an antibody mammalian cell display and massively parallel Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs* **5,** (2013).

19. Fujino, Y. *et al.* Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochem. Biophys. Res. Commun.* **428,** 395–400 (2012).

20. Jardine, J. G. *et al.* HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science (80-.).* **351,** 1458–1463 (2016).

21. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501,** 212–216 (2013).

22. Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat. Methods* **13,** 177–183 (2016).

23. Doolan, K. M. & Colby, D. W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *J. Mol. Biol.* **427,** 328–340 (2015).

24. Van Blarcom, T. *et al.* Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. *J. Mol. Biol.* **427,** 1513–1534 (2015).

25. Kowalsky, C. A. *et al.* Rapid Fine Conformational Epitope Mapping Using Comprehensive Mutagenesis and Deep Sequencing. *J. Biol. Chem.* **290,** 26457–26470 (2015).

26. Schlinkmann, K. M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proc. Natl. Acad. Sci.*

109, 9810–9815 (2012).

27. Schlinkmann, K. M. *et al.* Maximizing detergent stability and functional expression of a GPCR by exhaustive recombination and evolution. *J. Mol. Biol.* **422,** 414–428 (2012).

28. Egloff, P. *et al.* Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proc. Natl. Acad. Sci.* **111,** E655–E662 (2014).

29. Elazar, A. *et al.* Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *Elife* **5,** e12125 (2016).

30. Klesmith, J. R., Bacik, J., Michalczyk, R. & Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth. Biol.* **4,** 1235–1243 (2015).

31. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8,** 15695 (2017).

32. Deng, Z. *et al.* Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* **424,** 150–67 (2012).

33. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

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

35. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** gku511 (2014).

36. Thyme, S. B. *et al.* Massively parallel determination and modeling of endonuclease substrate specificity. **42,** 13839–13852 (2014).

37. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl. Acad. Sci.* **112,** 7159–7164 (2015).

38. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci.* **110,** E1263–E1272 (2013).

39. van der Meer, J. Y. *et al.* Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective Michaelases. *Nat. Commun.* **7,** 1–16 (2016).

40. Steinberg, B. & Ostermeier, M. Environmental changes bridge evolutionary valleys. *Sci.*

Adv. **2,** e1500921 (2016).

41. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9,** 2267–2284 (2014).

42. Kowalsky, C. A. *et al.* High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PLoS One* **10,** e0118193 (2015).

43. Cirino, P. C., Mayer, K. M. & Umeno, D. in *Directed Evolution Library Creation: Methods and Protocols* 3–9 (Springer, 2003).

44. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12,** 203–206 (2015).

45. Firnberg, E. & Ostermeier, M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One* **7,** e52031 (2012).

46. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* (2016). doi:10.1038/nmeth.4029

47. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19,** 1537–1551 (2013).

48. Borgstrom, E. *et al.* Phasing of single DNA molecules by massively parallel barcoding. *Nat. Commun.* **6,** (2015).

49. Cho, N. *et al.* De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries. *Nat. Commun.* **6,** (2015).

50. Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7,** 119–122 (2010).

51. Hong, L. Z. *et al.* BAsE-Seq : a method for obtaining long viral haplotypes from short sequence reads. *Genome Biol.* **15,** (2014).

52. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).

53. Stapleton, J. A. *et al.* Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. *PLoS One* **11,** e0147229 (2016).

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

55. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27,** 3430–3431 (2011).

56. Bloom, J. D. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* **16,** 1–13 (2015).

57. Woldring, D. R., Holec, P. V & Hackel, B. J. ScaffoldSeq: Software for characterization of directed evolution populations. *Proteins Struct. Funct. Bioinforma.* **84,** 869–874 (2016).

58. Abriata, L. A., Bovigny, C. & Peraro, M. D. Detection and sequence/structure mapping of biophysical constraints to protein variation in saturated mutational libraries and protein sequence alignments with a dedicated server. *BMC Bioinformatics* **17,** 1–13 (2016).

CHAPTER TWO

Nicking Mutagenesis: a plasmid-based, one-pot saturation mutagenesis method

This chapter is adapted with permission from "Plasmid-based one-pot saturation mutagenesis" in *Nature Methods* 13:11 (2016) 928-930 by Emily E. Wrenbeck, Justin R. Klesmith, James A. Stapleton, Adebola Adeniran, Keith EJ Tyo, and Timothy A. Whitehead.

ABSTRACT

Deep mutational scanning is a foundational tool for addressing the functional consequences of large numbers of mutants, yet a more efficient and accessible method for construction of user-defined mutagenesis libraries is needed. Here we present Nicking Mutagenesis, a robust single-day, single-pot saturation mutagenesis method that is performed on routinely prepped plasmid dsDNA. The method can be used to produce comprehensive, single-, or multi-site saturation mutagenesis libraries.

INTRODUCTION

Mutational studies have been used for over six decades to probe protein sequence-function relationships. Deep mutational scanning has emerged as a method to assess the effect of thousands of mutations on function using massively parallel functional screens and DNA counting via deep sequencing[1]. Information rich sequence-function maps obtained from such methods allow a researcher to address a variety of aims, including the generation of biomolecular fitness landscapes[2–6], therapeutic protein optimization[7], and high-resolution conformational epitope mapping[8]. Although other technical challenges have been resolved[9,10], a robust and accessible method for the construction of high quality, user-defined mutational libraries is lacking.

Random mutagenesis methods such as error-prone PCR suffer from limited codon sampling and imprecise control over the number of mutations introduced[11]. Of the published comprehensive saturation mutagenesis methods[2,4,11–14], PFunkel[12] offers the best combination of library coverage, mutational efficiency, control over number of mutations introduced, and scalability (**Table A 2.1**). In particular, PFunkel can be used to prepare libraries covering all possible single point mutations, with most members of the library having exactly one mutation. However, PFunkel is limited by the required preparation of a uracil-containing ssDNA template by phage infection. dU-ssDNA yields are highly variable and the preparation adds at least two days to the mutagenesis procedure. By analogy to site-directed mutagenesis, PCR-based methods like QuikChange have mostly supplanted the highly efficient Kunkel mutagenesis that also requires dU-ssDNA[15].

RESULTS

Here we present Nicking Mutagenesis, a method that does not rely on dU-ssDNA (**Figure 2.1**). Nicking mutagenesis is flexible, as any plasmid dsDNA can be used provided that it contains a single 7-bp BbvCI restriction site. The key mechanism in nicking mutagenesis is the successive creation and degradation of a wild-type ssDNA template. This is accomplished via a pair of nicking endonucleases (Nt.BbvCI and Nb.BbvCI)[16,17] that recognize the same site but nick one strand or the other, followed by exonuclease digestion. First, ssDNA template is created from dsDNA plasmid via a strand-specific nick introduced by Nt.BbvCI followed by selective digestion of the nicked strand with Exonuclease III (step 1; **Figure 2.1**). Mutant strands are then synthesized by thermal cycling template DNA with mutagenic oligos at a low primer-to-template ratio to promote annealing of effectively one primer to each template[12] (step 2). The highly processive and high fidelity Phusion Polymerase extends the primer around the circular template. *Taq* DNA Ligase closes the new strand to form a dsDNA plasmid with a mismatch at the mutational site. The heteroduplex DNA is then column purified to avoid buffer incompatibility issues and prevent potential competition between Phusion and Exonuclease III.

To resolve the heteroduplex, the opposite strand nicking endonuclease, Nb.BbvCI, creates a nick in the template strand, which is subsequently degraded by Exonuclease III (step 3). A secondary primer is then added and synthesis of the complementary mutant strand follows as above (step 4). To reduce wild-type background, the final reaction is treated with DpnI to digest methylated and hemi-methylated parental DNA. The complete protocol can be performed in a single day with minimal hands-on time (**Table A 2.2**).

We first optimized nicking mutagenesis using a green/white fluorescent screen based on reversion of a non-fluorescent green fluorescent protein (GFP) mutant (**Note A 2.1,**

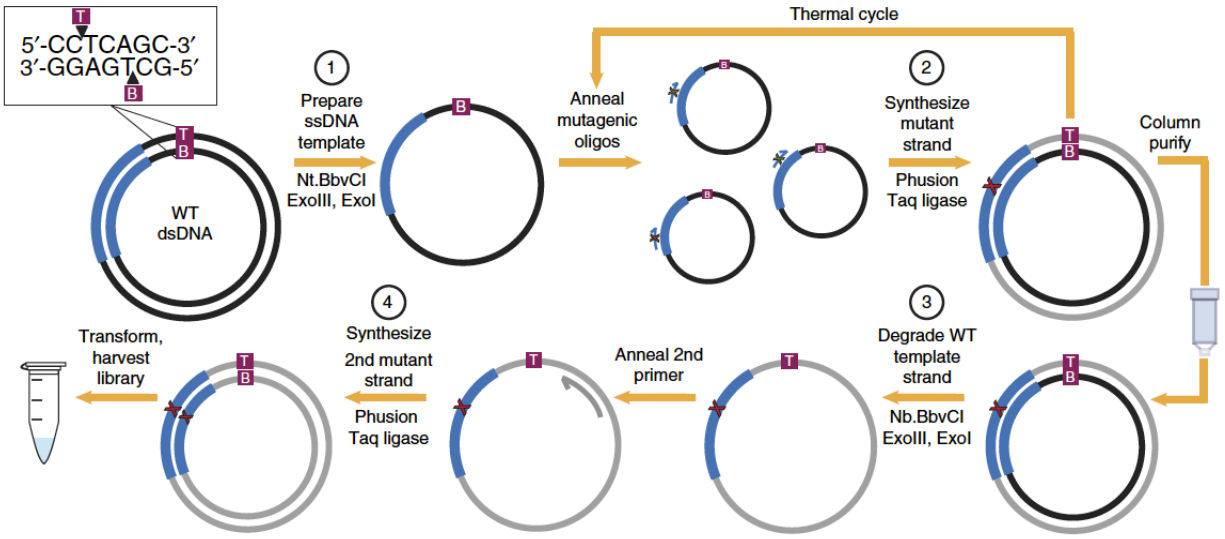


Figure 2.1: Comprehensive single-site Nicking Mutagenesis. Plasmid dsDNA containing a 7-bp BbvCI recognition site is nicked by Nt.BbvCI. Exonuclease III degrades the nicked strand to generate an ssDNA template (step 1). Mutagenic oligos are then added at a 1:20 ratio with template, Phusion Polymerase synthesizes mutant strands, and *Taq* DNA Ligase seals nicks (step 2). The reaction is column purified, and then the wild-type template strand is nicked by Nb.BbvCI and digested by Exonuclease III digestion (step 3). A second primer is added and the complementary mutant strand is synthesized to yield mutagenized dsDNA (step 4).

Figure A 2.1, and **Table A 2.3**). Next, we used nicking mutagenesis to prepare comprehensive single-site saturation mutagenesis libraries for two different 71 codon stretches of an aliphatic amidase encoded by the gene *amiE* from *Pseudomonas aeruginosa* (reaction 1 and 2 correspond to residues 100-170 and 171-241, respectively)[18]. A mixture of 71 degenerate NNN oligo sets, each with three consecutive randomized bases (NNN) corresponding to one of the 71 codons, was used at a 1:20 primer:template ratio. We deep sequenced the resulting libraries to an average depth of coverage of 2,200 reads per variant and processed the data using Enrich[19]. We observed 100% of possible single non-synonymous (NS) mutants (2840 total) and 100% of all possible programmed codon mutations (8946 total) with at least 10 reads (library coverage statistics are shown in **Table 2.1**). 64.4% and 63.5% of library members had exactly one

Table 2.1: Nicking mutagenesis library coverage statistics.

	Theoretical	*amiE* reaction 1	*amiE* reaction 2	*bla*
Sequencing reads post quality filter[19] (fold coverage)		4273346 (941x)	5378051 (1184x)	414417 (74x)
Number of transformants		$1.3x10^7$	$1.4x10^7$	$1.5x10^5$
Number of mutated codons		71	71	88
Total plasmid length (nucleotides)		4,612	4,612	6,907
Percent of reads with:				
No nonsynonymous mutations	1.6	27.2	26.3	30.1
One nonsynonmymous mutation	98.4	64.4	63.5	59.9
Multiple nonsynonymous mutations	0	8.4	10.2	9.7
Frameshift mutation	0	0.05	0.05	0.34
Percent of mutant codons with:				
1-bp substitution	14.3	32.2	31.4	25.4
2-bp substitution	42.9	32.8	31.5	41.7
3-bp substitution	42.9	35.0	37.1	33.0
Percent of possible codon substitutions observed				
1-bp substitution		100.0	100.0	99.7
2-bp substitution		100.0	100.0	83.5
3-bp substitution		100.0	100.0	77.8
All substitutions		100.0	100.0	83.4
Coverage of possible single amino acid substitutions with ≥5 reads		**100.0**	**100.0**	**91.5**
Coverage of possible programmed mutant codons with ≥5 reads		**100.0**	**100.0**	**75.4**

NS mutation for *amiE* reaction 1 and 2, respectively. The incidence of non-programmed indel mutations was 0.05% for both reactions 1 and 2. The frequency of individual mutations in each library followed a log-normal distribution, which is consistent with libraries prepared by PFunkel mutagenesis[6,9] (**Figure A 2.2**). In deep mutational scanning experiments, the initial library is typically sequenced at approximately 200-fold depth of coverage of the expected diversity.

Normalizing the above sequencing results to a 200-fold depth of coverage reveal that 93.2% and 97.8% of possible NS mutations would be represented above the typical threshold of 10 sequencing reads for *amiE* reaction 1 and 2, respectively (**Figure A 2.3a)**. This compares favorably with PFunkel mutagenesis (91.7% using the same threshold), although we note that the library distributions between the two methods are essentially identical (**Figure A 2.3b)**. We next assessed the libraries for off-target mutations by shotgun sequencing the input plasmid, pEDA3_amiE (no intended mutations), and library dsDNA from *amiE* reactions 1 and 2 and found the corresponding mutant tiles had significantly higher percent mutant allele rates (p-value $< 2.2 \times 10^{-16}$ for both reaction 1 and 2, **Figure A 2.4**).

To demonstrate performance on larger plasmids, we used nicking mutagenesis to prepare a comprehensive single-site saturation mutagenesis library for an 88-codon stretch of the gene *bla* encoding *E. coli* TEM-1 β-lactamase from a 6.9 kb plasmid and sequenced to 74-fold coverage of codon space. We observed nearly identical library composition with 91.5% coverage of possible amino acid substitutions (**Table 2.1**), which is consistent with expected coverage at this depth of sequencing (**Figure A 2.5**). Of note, we observed an order of magnitude fewer transformants when preparing this library compared to *amiE,* consistent with larger plasmids having lower transformation efficiency. One potential strategy to improve transformation efficiency is to use ultra-competent cells. Alternatively, the library can be constructed on a smaller plasmid and then transferred to a desired plasmid via subcloning.

To further expand the utility of nicking mutagenesis, we developed a single- and multi-site protocol (**Figure A 2.6**). The protocol was modified by adding primer at a 5:1 molar ratio to template and altering the thermal cycling steps for mutant strand synthesis. We tested the method by performing three single- and one triple-mutation nicking mutagenesis reaction on *bla* (plasmid

pSALECT-wtTEM1/csTEM1). Sanger sequencing of two clones from each of the three single-site reactions revealed that 5/6 clones contained a single mutation. For the multi-site reaction, 5 out of 10 sequenced clones contained the desired three programmed mutations.

Robust and effective molecular biology methods are characterized by the ease of their adoption by laboratories outside of where they were developed. To evaluate the accessibility of the method, the Tyo lab (Northwestern University) tested the method by performing single-site nicking mutagenesis on the pEDA5_GFPmut3_Y66H plasmid with the restore-to-function oligo GFP_H66Y. The resulting mutational efficiency, calculated by counting fluorescent (mutant) and non-fluorescent (wild-type) colonies, was 86.8 ± 6.1% (n=3 independent experiments).

DISCUSSION

We have demonstrated a single-pot single-day method for the preparation of comprehensive single- and multi-site saturation mutagenesis libraries from plasmid dsDNA (method cost detailed in **Table A 2.4**). The utility of nicking mutagenesis is not limited to saturation mutagenesis. Codon substitutions are user defined, making it possible to restrict diversity to specific residues such as hydrophobic or charged substitutions. An inherent limitation is that if multiple BbvCI nicking sites on a plasmid exist they must be in the same orientation. In the human genome, BbvCI has a mean distance between sites of 2058 base pairs, thus a considerable fraction of human genes will have nicking sites. Solutions include either cloning the gene of interest into a plasmid with a compatible nicking site orientation or using custom gene synthesis to remove extra BbvCI sites.

To validate the performance of nicking mutagenesis we used “testers” from an external lab; we propose using such testers to enhance reproducibility and accessibility of new molecular

biology methods. To aid in method adoption, the GFP plasmid used for green/white screening (pEDA5_GFPmut3_Y66H) has been deposited to the AddGene repository (www.addgene.org, plasmid #80085) as a tool for practicing and troubleshooting the method.

MATERIALS AND METHODS

Reagents

All chemicals were purchased from Sigma-Aldrich unless otherwise noted. All enzymes were purchased from New England Biolabs. All mutagenic oligos were designed using the QuikChange Primer Design Program (Agilent, Santa Clara, CA). Mutagenic oligos and sequencing primers were ordered from Integrated DNA Technologies (Coralville, IA).

Plasmid construction

All primer sequences used in this work are listed in **Table A 2.3**. Plasmid pEDA5_GFPmut3_Y66H was prepared by modification of pJK_proB_GFPmut3 as described in Bienick et al.[18] by a single Kunkel[15] reaction with two mutagenic primers: one encoding a BbvCI site (primer pED_BbvCI) and the second to introduce a Tyr66His point mutation (primer GFP_Y66H). pEDA3_amiE was constructed by altering pJK_proK17_amiE as described in Bienick et al.[18] with a single Kunkel[15] reaction with two primers: one encoding a BbvCI site (pED_BbvCI) and the second encoding a mutated ribosome binding sequence (pED_kRBS3). pEDA5_GFPmut3_Y66H has been deposited in the AddGene repository (www.addgene.org, plasmid #80085).

Plasmid pSALECT-wtTEM1/csTEM1 was created as follows. Overhang PCR was used to add in an XhoI and BbvCI site after the existing NdeI site and before the original stop codon of

plasmid pSALECT-EcoBam (Plasmid #59705, acquired from AddGene). A Δ2-23 truncation of wild-type TEM-1 was cloned in-frame between the NdeI and XhoI sites. A codon swapped Δ2-23 truncation of wild-type TEM-1 with a C-terminal His_{6x} tag and double stop codon was ordered as a gBlock (IDT) and was cloned in-frame between the XhoI and BbvCI site. This second TEM-1 is a C-terminal fusion to the wild-type TEM-1.

Plasmid pETconNK-TEM1(S70A,D179G) was created as follows. Gibson assembly was used to remove the ampicillin gene from pETcon(-) (Addgene plasmid #41522) and insert a kanamycin gene with a 3' BbvCI site on the coding strand. A Δ2-23 truncation of TEM-1 with point mutations S70A and D179G was cloned in-frame between the NdeI and XhoI sites.

Comprehensive nicking mutagenesis optimization

The final optimized comprehensive nicking mutagenesis protocol is supplied in Supplementary Protocol 1 and on the Protocols Exchange (DOI 10.1038/protex.2016.061). 1X CutSmart Buffer (NEB) was used as an enzyme diluent when necessary. Two reactions were set up as follows: 0.76 pmol pEDA5_GFPmut3_Y66H was incubated with 10 U each of Nt.BbvCI and Exonuclease III in 1X CutSmart Buffer (20 µL final volume) for 60 minutes at 37°C followed by 80°C for 20 minutes (heat kill). 40 U of DpnI was added and the reaction was incubated at 37°C for 60 minutes followed by 80°C for 20 minutes (heat kill). One reaction was then column purified by Zymo Clean & Concentrator (5:1 v/v ratio of binding buffer to sample), eluted in 6 µL Nuclease-Free H_2O (NFH_2O, Integrated DNA Technologies), transformed into XL1-Blue electrocompetent cells, and dilution plated. The following was added to the second reaction: 200 U of *Taq* DNA Ligase, 2 U Phusion High-Fidelity DNA Polymerase, 20 µL 5X Phusion HF Buffer, 20 µL 50 mM DTT, 1 µL 50 mM NAD^+, 2 µL 10 mM dNTPs, 29 µL NFH_2O (final reaction volume

of 100μL). The tube was placed into a preheated (98°C) thermal cycler set with the following program: 98°C for 2 minutes, 15 cycles of 98°C for 30 seconds (denature), 55°C for 45 seconds (anneal oligos), 72°C for 7 minutes (extension), followed by a final incubation at 45°C for 20 minutes to complete ligation. The reaction was column purified, transformed, and dilution plated as described above.

The optimization experiment including addition of Exonuclease I was performed as described below with the following modifications. A single mutagenic primer, His66Tyr (restores wild-type chromophore sequence), was used at a 1:20 primer:template ratio. The reaction was column purified and transformed into XL1 Blue electrocompetent cells as above. Green fluorescent (mutated) and white (parental) colonies were counted to calculate transformational and mutational efficiencies.

Comprehensive nicking mutagenesis of amiE and bla

Three separate reactions targeting residues 100-170 and 171-241 of amiE and 201-289 of TEM-1 were performed. Mutagenic oligos programming degenerate codons (NNN) for each reaction were mixed in equimolar amounts to a final concentration of 10 μM. 20 μL of each primer mix was added to a phosphorylation reaction containing 2.4 μL of T4 Polynucleotide Kinase Buffer, 1 μL 10 mM ATP, 10 U T4 Polynucleotide Kinase, and incubated for 1 hour at 37°C. Secondary primer pED_2ND was phosphorylated in a reaction containing 18 μL NFH_2O, 2 μL T4 Polynucleotide Kinase Buffer, 7 μL 100 μM secondary primer, 1 μL 10 mM ATP, and 10 U T4 Polynucleotide Kinase. The reaction was incubated for 1 hour at 37°C. Phosphorylated NNN and secondary primers were diluted 1:1000 and 1:20 in NFH_2O, respectively.

ssDNA template was prepared in a reaction containing 0.76 pmol plasmid dsDNA, 2 µL NEB CutSmart Buffer, 10 U Nt.BbvCI, 10 U Exonuclease III, 20 U Exonuclease I, and NFH_2O to 20 µL final reaction volume in a PCR tube. The following thermal cycle program was used: 37°C for 60 minutes, 80°C for 20 minutes (heat kill), hold at 4-10°C. Next, for mutant strand synthesis the following was added to each PCR tube on ice: 20 µL 5X Phusion HF Buffer, 20 µL 50 mM DTT, 1 µL 50 mM NAD^+, 2 µL 10 mM dNTPs, 4.3 µL 1:1000 diluted phosphorylated NNN mutagenic oligos, and 26.7 µL NFH_2O (final reaction volume of 100µL). The tube contents were mixed, spun down, and placed on ice. 200 U of *Taq* DNA Ligase and 2 U Phusion High-Fidelity DNA Polymerase were added to each reaction, mixed, spun down, and placed into a preheated (98°C) thermal cycler set with the following program: 98°C for 2 minutes, 15 cycles of 98°C for 30 seconds (denature), 55°C for 45 seconds (anneal oligos), 72°C for 7 minutes (extension), followed by a final incubation at 45°C for 20 minutes to complete ligation. Additional 4.3 µL of oligos were added at the beginning of cycles 6 and 11. Each reaction was then column purified using a Zymo Clean & Concentrator kit (5:1 DNA Binding Buffer to sample). Each reaction was eluted in 15 µL NFH_2O, and 14 µL was transferred to a fresh PCR tube.

Next, for the template degradation reaction the following was added to each tube: 2 µL 10X NEB CutSmart Buffer, 1 U Nb.BbvCI, 2 U Exonuclease III, and 20 U Exonuclease I (20µL final volume). The following thermocycler program was used: 37°C for 60 minutes, 80°C for 20 minutes (heat kill), hold at 4-10°C. To synthesize the second (complementary) mutant strand, the following was added to each reaction: 20 µL 5X Phusion HF Buffer, 20 µL 50 mM DTT, 1 µL 50 mM NAD^+, 2 µL 10 mM dNTPs, 3.3 µL 1:20 diluted phosphorylated secondary primer (0.38 pmol), and 27.7 µL NFH_2O (final reaction volume of 100 µL). The tube contents were mixed, spun down, and placed on ice. 200 U of *Taq* DNA Ligase and 2 U Phusion High-Fidelity DNA

Polymerase were added to each reaction, mixed, spun down, and placed into a preheated (98°C) thermal cycler set with the following program: 98°C for 30 seconds, 55°C for 45 seconds, 72°C for 10 minutes (can be extended for longer constructs), and 45°C for 20 minutes.

To degrade methylated and hemi-methylated wild-type DNA, 40 U of DpnI was added to each reaction and incubated at 37°C for 1 hour. The final reaction was column purified using the Zymo Clean & Concentrator-5 kit as described above but eluted in 6 μL NFH_2O. The entire 6 μL was transformed into 40 μL of XL1-Blue electroporation competent cells (Agilent) and plated on Corning square bioassay dishes (Sigma-Aldrich, 245mm x 245mm x 25mm). The following day, colonies were scraped with 15 mL of TB, vortexed, and 1 mL was removed and mini-prepped using a Qiagen Mini-prep Kit.

Single and multi-site nicking mutagenesis

Mutagenic primers were phosphorylated separately following the protocol described above for the secondary primer, then diluted 1:20 with NFH_2O. For multi-site nicking mutagenesis, 2 μL of each primer was mixed in a single tube and diluted to a final volume of 40 μL. ssDNA template preparation was performed as described above. For mutant strand synthesis, oligos were annealed in the absence of polymerase as suggested by Firnberg et al.[11]. 3.3 μL of 1:20 phosphorylated oligos (single or mixed), 10 μL 5X Phusion HF Buffer, and 16.7 μL NFH_2O were added to the appropriate tube. Oligos were annealed with the following thermocycler program: 98°C for 2 minutes, decrease to 55°C over 15 minutes, 55°C for 5 minutes, and hold at 55°C. While the reactions were held on the block, the following was added to each tube from a master mix: 20 μL 5X Phusion HF Buffer, 20 μL 50 mM DTT, 1 μL 50 mM NAD^+, 2 μL 10 mM dNTPs, and 11 μL NFH_2O (final reaction volume of 100μL). The tube contents were mixed by pipetting, then 200 U

of *Taq* DNA Ligase and 2 U Phusion High-Fidelity DNA Polymerase were added to each reaction, mixed, spun down, and returned to the thermocycler for the following program: 72°C for 10 minutes, 45°C for 20 minutes. The remainder of the protocol proceeded as described in the comprehensive protocol.

DNA deep sequencing and analysis

Plasmids obtained after transformation of the reaction mix and miniprep were used for deep sequencing analysis of library coverage. Samples were prepared for deep sequencing as described in Kowalsky et al.[9] following Method B. Sequences of PCR primers are listed in **Table A 2.3**. Samples for shotgun sequencing were prepared at the Michigan State University sequencing core (approximate median insert size of 360bp). *amiE* libraries were sequenced on an Illumina MiSeq with 250bp PE reads at the University of Illinois Chicago sequencing core. All other samples were sequenced on an Illumina MiSeq with 300bp PE reads at Michigan State University. Read statistics are given in **Table 2.1**. Raw FASTQ files were analyzed with Enrich software[19] with modifications as described in Kowalsky et al.[9]. Analysis of libraries for frameshift and off-target mutations was done using the Burrows Wheeler Aligner[20] followed by processing with SAMtools[21]. Library statistics (**Table 2.1**) and read coverage plots (**Figure A 2.2 and A 2.5a**) were obtained using custom scripts freely available at Github (user JKlesmith). Sequencing data has been deposited to the NCBI Sequence Read Archive (accession numbers SRR4105481-SRR4105486).

Statistics

For analysis of the shotgun sequencing data, the mean of the background subtracted per-position percent mutant allele values for *amiE* reactions 1 and 2 at positions inside and outside the targeted

region for mutagenesis were computed. Welch two sample t-tests were performed using the R statistical software[22] to calculate significance between averages from the inside regions and the outside regions for reaction 1 (p-value < $2.2*10^{-16}$, t = -14.846, df = 697.06) and reaction 2 (p-value < $2.2*10^{-16}$, t = -19.259, df = 214).

APPENDIX

APPENDIX

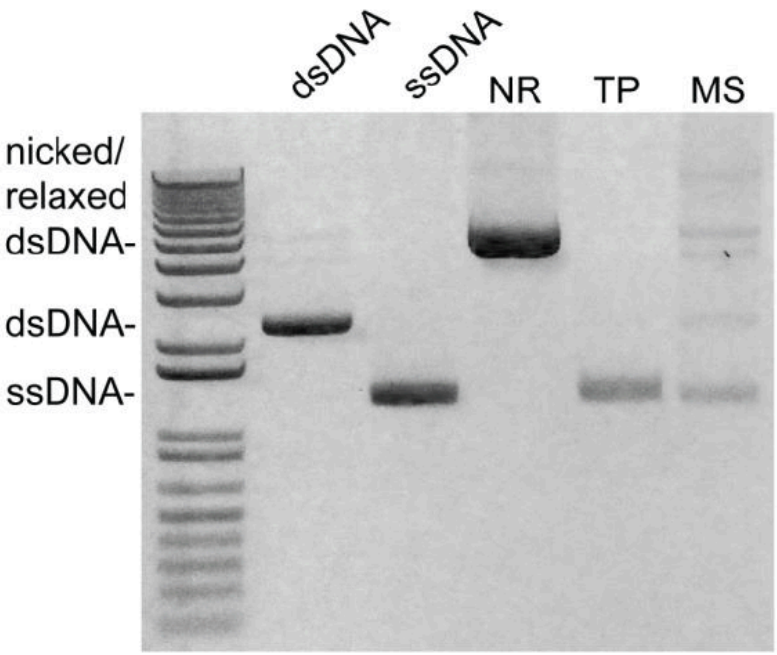


Figure A 2.1: Gel snapshots along the optimized nicking mutagenesis method. Plasmid dsDNA and ssDNA (prepared from bacteriophage) of pEDA5_GFPmut3 are included for size reference. NR = nicking reaction; 2 µg of pEDA5_GFPmut3_Y66H was placed in a 20 µL reaction with 10 U Nt.BbvCI in 1X CutSmart buffer. TP = template preparation; a reaction was ceased after the template preparation phase. MS = mutant strand; a reaction was ceased after the synthesis of the mutant strands, where regeneration of relaxed dsDNA can be seen. 1 kb Plus Ladder (Thermo Fischer Scientific, lane 1) included for size reference. Gel image has been cropped to size.

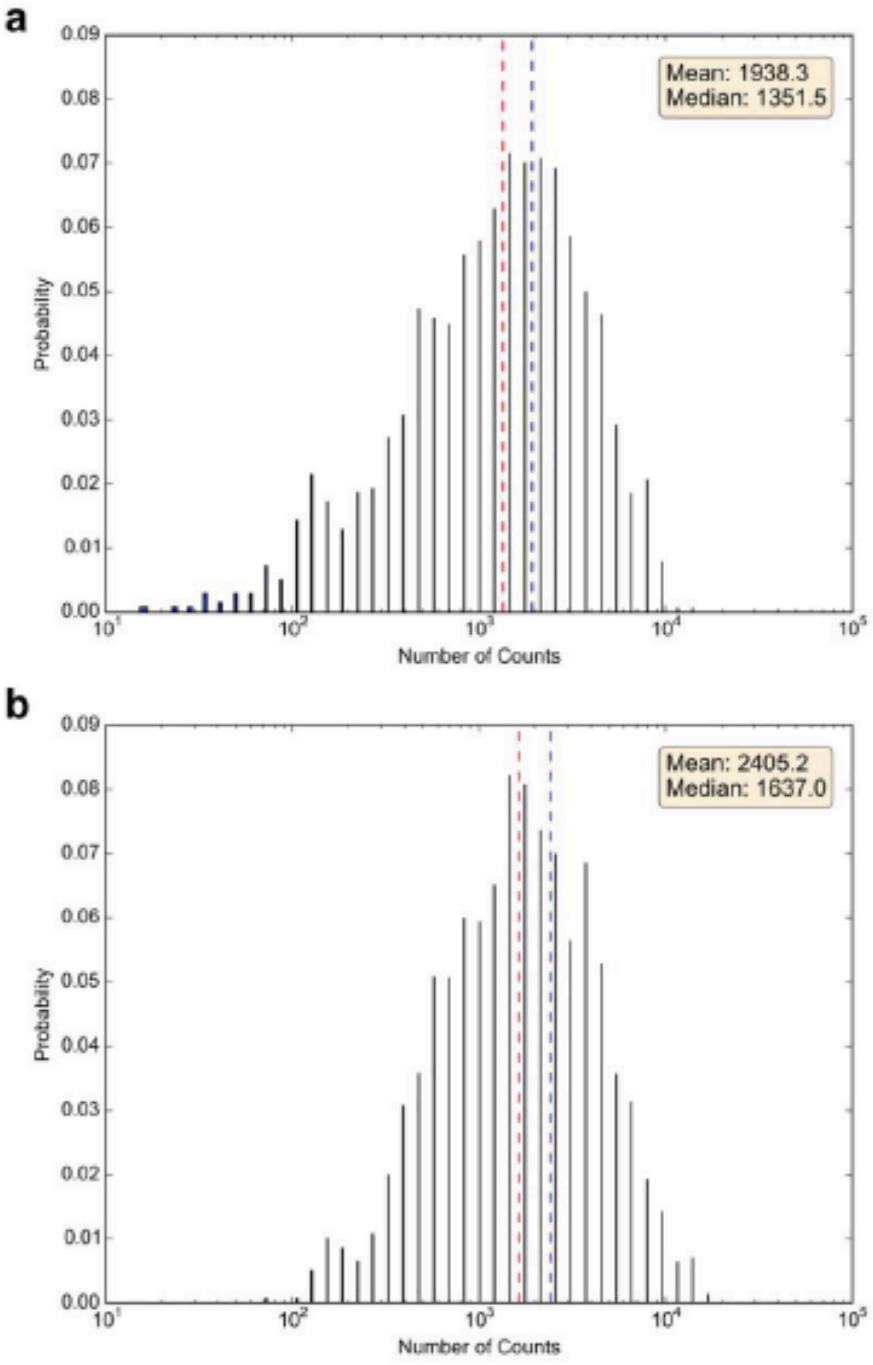


Figure A 2.2: Probability distribution of mutation counts in *amiE* comprehensive nicking mutagenesis libraries. Dashed vertical lines represent median (red) and mean (blue) library member read coverage. Panel a shows distribution for reaction 1 and panel b shows the distribution for reaction 2.

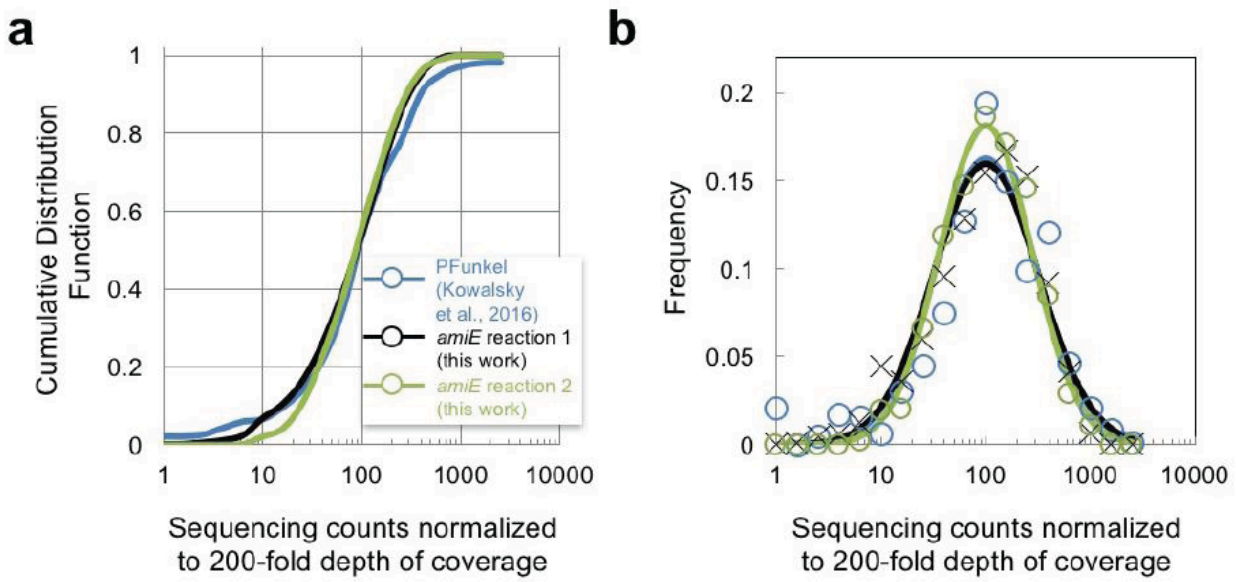


Figure A 2.3: Comparison of the probability distributions of site-saturation mutagenesis libraries resulting from nicking mutagenesis or PFunkel mutagenesis. Because the depth of sequencing coverage varied between the three methods, all samples were normalized to a 200-fold depth of coverage of possible single non-synonymous mutations. The expected library diversity is 820 for Kowalsky et al.1,2 and 1420 for *amiE* reaction 1 & reaction 2 (this work). **a.** Cumulative distribution function for the three libraries as a function of normalized sequencing counts. 91.7%, 93.2%, and 97.8% of the library is represented above a threshold of 10 sequencing counts for PFunkel library, *amiE* reaction 1, and the *amiE* reaction 2 libraries, respectively. **b.** Frequency is plotted as a function of sequencing counts for the same three libraries. The experimental data are plotted as symbols, with lines representing a best fit of the data using a log-normal distribution (PFunkel: μ=2, σ=0.49, *amiE* reaction 1: μ=2, σ=0.50. *amiE* reaction 2: μ=2, σ=0.44).

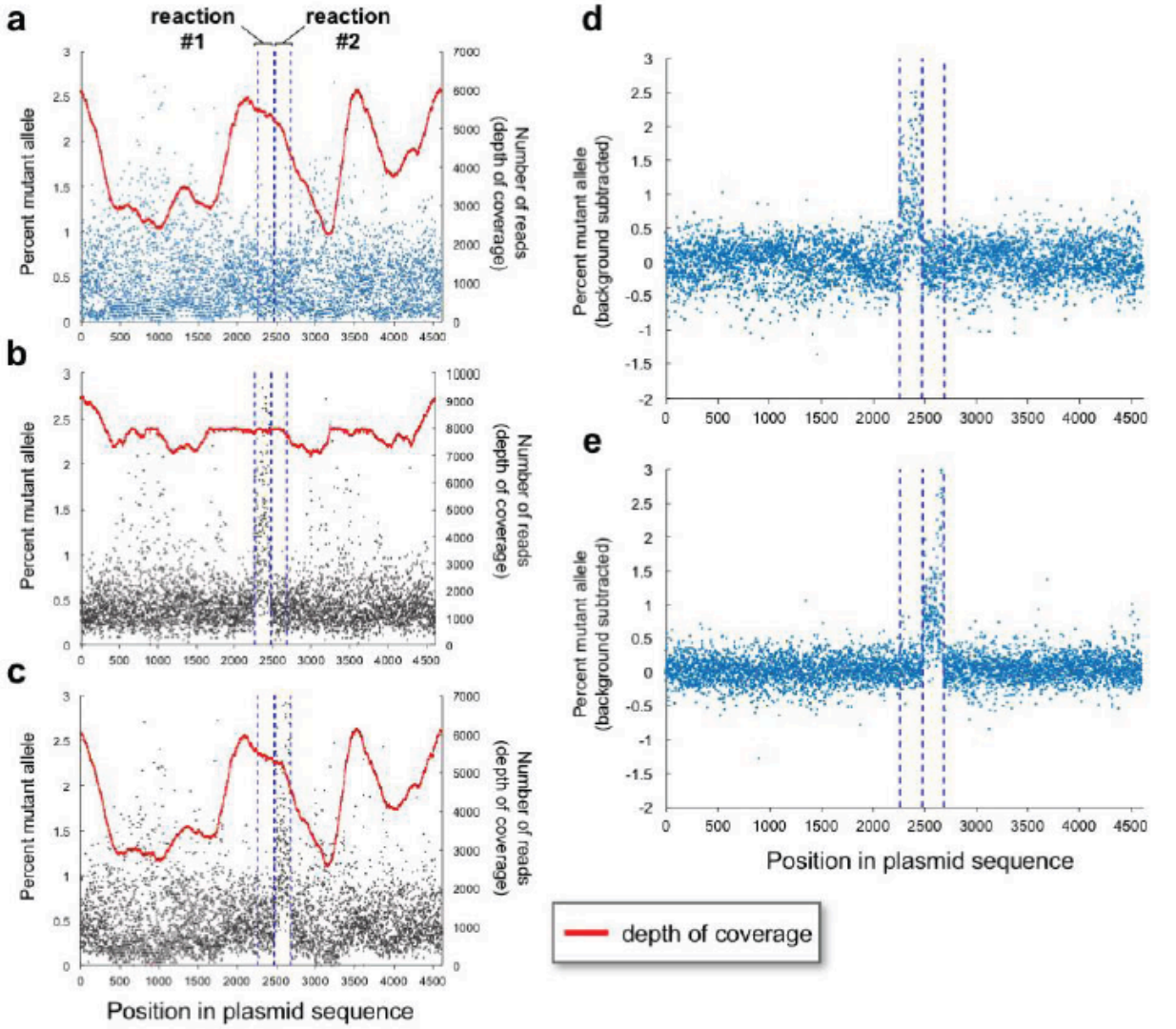


Figure A 2.4: Off-target mutational analysis of *amiE* input plasmid and mutational libraries by shotgun sequencing. a-c. Percent mutant allele at each position in the plasmid sequence for the input plasmid (**a**) *amiE* reaction 1 library (**b**) and *amiE* reaction 2 library (**c**). Shotgun sequencing reads were aligned to the pEDA3_amiE plasmid using BWA aligner[3,4] and the frequency of each base at each position was counted using bam-readcount (www.github.com). Percent mutant allele was calculated for each position by summing all non-wildtype allele counts and diving by total reads at that position. Overlain red curves indicate depth of sequencing coverage at each position. **d-e**. Background subtracted percent mutant allele for each position in plasmid sequence of *amiE* reaction 1 library (**d**) and *amiE* reaction 2 library (**e**).

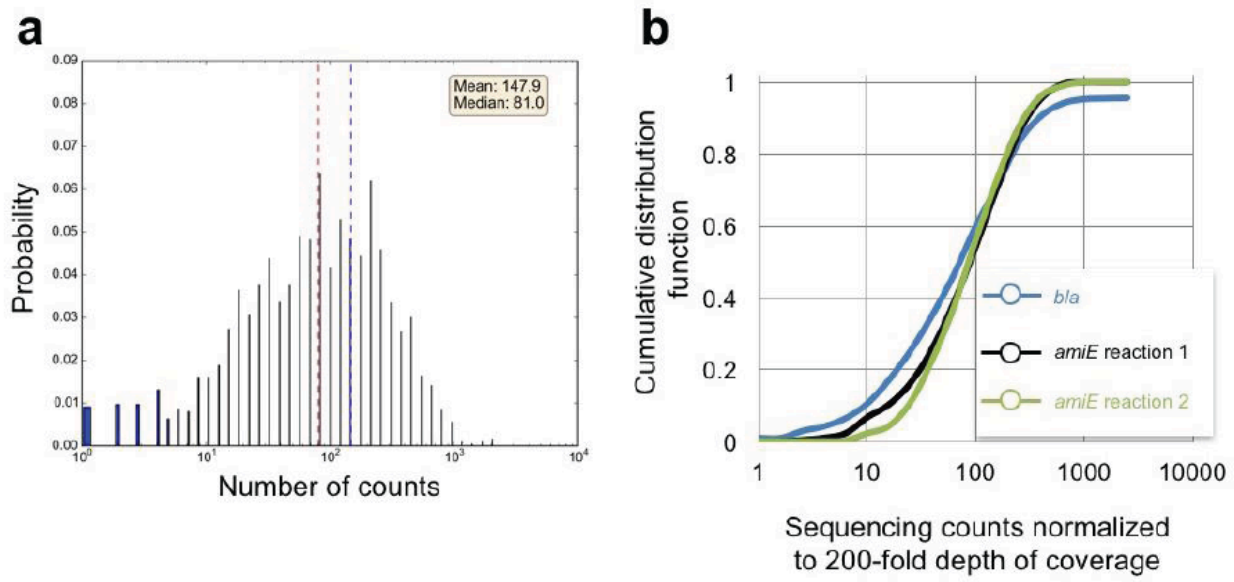


Figure A 2.5: *bla* library coverage distributions. Probability distribution of mutation counts in *bla* comprehensive nicking mutagenesis libraries. Dashed vertical lines represent median (red) and mean (blue) library member read coverage. **b.** Cumulative distribution function for the three libraries as a function of normalized sequencing counts.

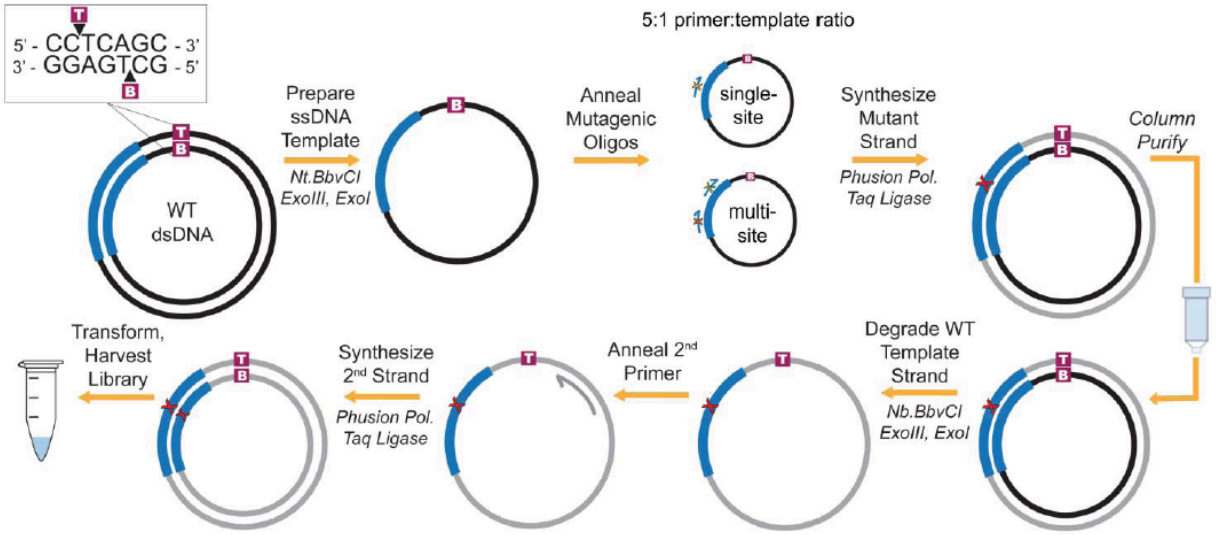


Figure A 2.6: Schematic overview of single- or multi-site nicking mutagenesis. After the preparation of an ssDNA template, an annealing reaction is set up with a single or mixed set of mutagenic oligos at a 5:1 primer:template ratio (for each oligo). Next, reagents and enzymes necessary to synthesize the mutant strands are added. The remainder of the protocol is identical to comprehensive nicking mutagenesis.

Table A 2.1: Performance metrics of published comprehensive mutagenesis methods[2,11,12,14,23]**.** Bolded text indicates metrics that are comparatively inefficient to nicking mutagenesis and PFunkel mutagenesis. NS = nonsynonymous.

Mutagenesis method Publication data gathered from Gene (# codons mutated)	**Library type**	**Library coverage**	**Percent of mutants with NS mutations**			**Scalability** *mutatable codons/ reaction*
			Single	**Zero**	**Multiple**	
Cassette Mutagenesis Hietpas et al.[2] Hsp90 (9)	user-defined	100%	nd	nd	nd	**20**
Error-Prone PCR Doolan et al.[23] mouse PrP (211)	random	nd	**28.2%**	**60.6%**	**11.08%**	all
Chemical Synthesis Fowler et al.[14] hYAP65 WW domain (25)	random	**83.2%**	nd	20*	nd	**30**
PALS Mutagenesis Kitzman et al.[11] Gal4 DBD and p53 (457 total)	user-defined	**94.3%**	**35%**	**29.2%**	**33%**	all
PFunkel Mutagenesis Kowalsky et al.[24] Ct Cohesin (162)	user-defined	97.1%	73.6%	20.5%	5.9%	all
Nicking Mutagenesis This work amiE (142)	user-defined	100.0%	64%	26.8%	9.3%	all

*estimated from Supplementary Figure 3 of original publication

Table A 2.2: Estimated time required for comprehensive library construction using nicking mutagenesis.

Step number		Hands-on time (min)	On-thermal cycler time (min)
1a*	Phosphorylate oligos	30	60*
1b*	ssDNA template strand preparation	5	80*
2a	Comprehensive codon mutagenesis strand 1	10	146
2b	Column purification I	5	
3	Degrade template strand	5	80
4a	Synthesize complimentary mutagenic strand	10	32
4b	DpnI DNA cleanup	2	60
4c	Column purification II	5	
	Subtotal (hr):	1.2	6.6
	Total (hr):	**7.8**	

*steps can be performed simultaneously

Table A 2.3: Primer sequences.

Plasmid construction primers	
pED_BbvCI	gcggccccacgggtcctcagcgcgcatgat
pED_kRBS3	gacgagctaatatcgccatgtctcatatgtataaaaacttcttaaagttaaacaaaattatttctagaaagttaaa
GFP_Y66H	gcaaagcattgaacaccatgaccgaaagtagtgacaagt
Green/white screening mutagenic oligos	
GFP_H66Y	gcaaagcattgaacaccataaccgaaagtagtgacaagt
GFP_H66Y_RC	acttgtcactactttcggttatggtgttcaatgctttgc
Green/white screening secondary primer	
pED_2ND	ggtgattcattctgctaa
amiE and TEM-1 secondary primers	
pED_2ND (amiE)	ggtgattcattctgctaa
pSALECT/pETconNK_2ND (TEM-1)	ggtttcccgactggaaag
Gene amplification: inner primers	
amiE_NMT1_FWD	gttcagagttctacagtccgacgatcgcaaatgtttggggtgtg
amiE_T2_FWD	gttcagagttctacagtccgacgatcctgcgatgacggtaat
amiE_T1_REV	ccttggcacccgagaattccactctccaaatttccggata
amiE_NMT2_REV	ccttggcacccgagaattccattcgccgcattcacccagagt
TEM1_T3_FWD	gttcagagttctacagtccgacgatcattaactggcgaactacttact
pETconNK_REV	ccttggcacccgagaattccaaagcttttgttcggatc
blue = Illumina sequencing primer; black = gene overlap	
Gene amplification: outer primers	
Illumina_FWD	aatgatacggcgaccaccgagatctacacgttcagagttctacagtccga
RPI30	caagcagaagacggcatacgagat**CCGGTG**gtgactggagttccttggcacccgagaattcca
RPI31	caagcagaagacggcatacgagat**ATCGTG**gtgactggagttccttggcacccgagaattcca
RPI21	caagcagaagacggcatacgagat**CGAAAC**gtgactggagttccttggcacccgagaattcca
red = Illumina adapter sequence; **BOLD** = barcode; blue = Illumina sequencing primer	

Table A 2.4: Cost analysis of nicking mutagenesis compared with PFunkel[12]. Library preparation cost was calculated by totaling cost of enzymes (price information gathered from New England Biolabs) and reagents (price information gathered from Sigma-Aldrich, Qiagen, and Zymo Research) on a per reaction basis. Price of chemically synthesized degenerate NNN oligos based on IDT pricing for a 40bp primer[6] at the 500 pmole scale: $0.10/base*40bp = $4/codon. Prices obtained February 2016.

	PFunkel	**Nicking Mutagenesis**
Library preparation cost per reaction	$53	$55
NNN oligo cost per codon (source)	$4 (IDT)	$4 (IDT)
Total cost per 100 scanned codons	**$453**	**$455**

Note A 2.1: Optimization of nicking mutagenesis using green/white screening

A previously constructed GFPmut3 expression plasmid[18] was modified by incorporating a BbvCI site and by changing the amino acid sequence of the GFPmut3 chromophore, Gly65-Tyr66-Gly67, to Gly65-His66-Gly67, resulting in a non-fluorescent protein. We performed nicking mutagenesis on this construct (pEDA5_GFPmut3_Y66H) with a restore-to-function mutagenic oligo (primer GFP_H66Y, see **Supplementary Table 3** for sequences). **Figure A 2.1** shows gel snapshots at different stages along the optimized process.

Initial experiments with the full nicking mutagenesis protocol showed a mutational efficiency of 23% with $3x10^5$ transformants. To determine the sources of high wild-type background, we performed a series of control experiments containing no mutagenic primer. Thus, any resulting transformants could be unambiguously attributed to wild-type. The number of background transformants was 10^3 after the template preparation step and incubation with DpnI, but increased to 10^6 if the reaction was allowed to proceed through the thermal cycling steps. We hypothesized that short stretches of incompletely degraded DNA were priming and regenerating wild-type constructs. To remedy this, Exonuclease I, which specifically degrades ssDNA, was added to both the template preparation and degradation reactions. The addition of Exonuclease I improved mutational efficiency to 56% with $>5x10^5$ transformants. Incubation of the final reaction mixture with DpnI to remove methylated and hemi-methylated wild-type DNA increased the mutational efficiency to 68% with $>3x10^5$ transformants.

In oligonucleotide-programmed mutagenesis, mutagenic oligos are designed to be complementary to the wild-type template sequence on either side of the programmed mutation such that they can anneal to the template. For Kunkel mutagenesis[15], the ssDNA template strand

is made by replication and packaging within a phage host. The directionality of the ssDNA template strand (sense or anti-sense) is dependent upon the directionality of the F1-origin of replication. If the F1-origin is such that the template strand made is sense, then mutagenic oligos are designed anti-sense.

For nicking mutagenesis, the directionality of the template strand is dependent upon the orientation of the BbvCI site. The set of enzymes, Nt.BbvCI (Nick-top BbvCI) and Nb.BbvCI (Nick-bottom BbvCI) will create nicks on the strands containing their respective recognition sequence. If the Nt.BbvCI nicking enzyme is used for template preparation and its recognition sequence is encoded on the anti-sense strand, the ssDNA template formed will be sense. Thus, mutagenic oligos should be designed anti-sense. The opposite is true if Nb.BbvCI was used to create the template strand.

To confirm that the order of nicking enzymes could be switched, we performed nicking mutagenesis using green/white screening in two reactions: one with Nt.BbvCI then Nb.BbvCI using the GFP_H66Y mutagenic primer (priming one strand), and the second using Nb.BbvCI first with the GFP_H66Y_RC primer (priming the opposite strand at the same location as GFP_H66Y). We observed mutational efficiencies of 46% and 44% with $>8x10^4$ and $>9x10^4$ total transformants, respectively, confirming that the order of nicking enzymes can be switched.

Another consideration is that a target gene of interest may contain a BbvCI nicking site. In such a case, confirm that the orientation of the BbvCI nicking site is the same on the gene as on the backbone.

REFERENCES

REFERENCES

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

2. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **108,** 7896–7901 (2011).

3. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

4. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** gku511 (2014).

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

6. Klesmith, J. R., Bacik, J., Michalczyk, R. & Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth. Biol.* **4,** 1235–1243 (2015).

7. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–8 (2012).

8. Kowalsky, C. A. *et al.* Rapid Fine Conformational Epitope Mapping Using Comprehensive Mutagenesis and Deep Sequencing. *J. Biol. Chem.* **290,** 26457–26470 (2015).

9. Kowalsky, C. A. *et al.* High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PLoS One* **10,** e0118193 (2015).

10. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9,** 2267–2284 (2014).

11. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12,** 203–206 (2015).

12. Firnberg, E. & Ostermeier, M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One* **7,** e52031 (2012).

13. Jain, P. C. & Varadarajan, R. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.* **449,** 90–98 (2014).

14. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

15. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci.* **82,** 488–492 (1985).

16. Chan, S.-H., Stoddard, B. L. & Xu, S. Natural and engineered nicking endonucleases - from cleavage mechanism to engineering of strand-specificity. *Nucleic Acids Res.* **39,** 1–18 (2011).

17. Heiter, D. F., Lunnen, K. D. & Wilson, G. G. Site-Specific DNA-nicking Mutants of the Heterodimeric Restriction Endonuclease R.BbvCI. *J. Mol. Biol.* **348,** 631–640 (2005).

18. Bienick, M. S. *et al.* The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate. *PLoS One* **9,** e109105 (2014).

19. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27,** 3430–3431 (2011).

20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

21. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

22. R Core Team, R. R: A language and environment for statistical computing. (2015). at <https://www.r-project.org/>

23. Doolan, K. M. & Colby, D. W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *J. Mol. Biol.* **427,** 328–340 (2015).

24. Kowalsky, C. A. & Whitehead, T. A. Determination of binding affinity upon mutation for type I dockerin-cohesin complexes from Clostridium thermocellum and Clostridium cellulolyticum using deep sequencing. *Proteins* (2016).

CHAPTER THREE

Exploring the sequence-determinants to specificity of an enzyme using deep mutational scanning

This chapter is adapted with permission from "Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded" in *Nature Communications* 8 (2017) 15695 by Emily E. Wrenbeck, Laura R. Azouz, and Timothy A. Whitehead.

ABSTRACT

Our lack of total understanding of the intricacies of how enzymes behave has constrained our ability to robustly engineer substrate specificity. Furthermore, the mechanisms of natural evolution leading to improved or novel substrate specificities are not wholly defined. Here we generate near-comprehensive single-mutation fitness landscapes comprising >96.3% of all possible single nonsynonymous mutations for hydrolysis activity of an amidase expressed in E. coli with three different substrates. For all three selections, we find that the distribution of beneficial mutations can be described as exponential, supporting a current hypothesis for adaptive molecular evolution. Beneficial mutations in one selection have essentially no correlation with fitness for other selections and are dispersed throughout the protein sequence and structure. Our results further demonstrate the dependence of local fitness landscapes on substrate identity and provide an example of globally distributed sequence-specificity determinants for an enzyme.

INTRODUCTION

Understanding the sequence determinants to substrate specificity for enzymes is a significant challenge in protein science that impacts fields as diverse as evolutionary biology and biocatalysis[1,2]. The dynamic relationship between protein structure and function makes it difficult to predict perturbations to the primary sequence that will improve or alter activity for a given substrate[2]. More fundamental concerns relate the nature of protein fitness landscapes to a biophysical basis underlying molecular evolution and adaptation[3,4]. What is the distribution of fitness effects (DFE) for mutations, and do they correspond with existing theory of adaptation[5–7]? Are the DFE of mutations correlated between substrates[8]? Are specificity-modulating mutations correlated to bulk properties of enzymes (e.g. distance to active site)?

Over the past 20 years directed evolution experiments have provided a number of insights to the above questions[9–11]. For engineering enzyme specificity, it has been shown that a rational mutagenesis approach – primarily focused on residues lining a substrate binding pocket - provides greater payoffs than random mutagenesis (i.e. error-prone PCR)[1,12,13]. However, it is no secret that distant (>10 Å) mutations can have significant effects on catalytic function[13–20]. For example, in a classic paper Oue et al. evolved the specificity of an aspartate aminotransferase to valine and found only one mutation in direct contact with the substrate out of seventeen accumulated in the final construct[20]. However, the spatial distribution of specificity-modulating substitutions is still unclear, as typical experiments assay the effects of less than 100 mutations. Large scale mutational studies, such as deep mutational scanning to generate local fitness landscapes[21,22], provide a more comprehensive purview and can potentially be used to resolve the above open questions[23].

From the protein engineer's perspective, the ability to predict fitness effects would greatly improve the discovery rate of beneficial mutations. In recent years, theoretical work on adaptive

molecular evolution has experienced a revolution with the availability of new experimental tools. Recognizing the rare nature of beneficial mutations, Gillespie[24] borrowed extreme value theory mathematics to predict that the distribution of fitness effects (DFE) for beneficial mutations, drawn from the extreme tail of DFEs, would be of the Gumbel or 'typical' type (exponential, gamma, Weibull, etc.). Orr later proposed that beneficial mutations from a high fitness parent should be roughly exponentially distributed[6]. While generally providing support for these theories, discerning the parameterization of a mathematical model from experimental data has yielded mixed conclusions as summarized by Orr[25].

To explore the question of how enzymes encode specificity and scrutinize adaptive molecular evolution theory, we evaluate the sequence determinants to substrate specificity for an enzyme by generating comprehensive single-mutation fitness landscapes – the effects of all possible single point mutations - on multiple substrates. As a model system we use the aliphatic amide hydrolase encoded by *amiE* from *Pseudomonas aeruginosa*[26] because the structure is solved[27], amidases are an industrially-relevant class of enzymes[28,29], and amiE has activity against multiple substrates. In particular, amiE maintains comparatively higher activity on acetamide and propionamide compared with the bulkier isobutyramide. Thus, our experimental system allows comparison of adaptation between similar and structurally dissimilar substrates.

RESULTS

Local Fitness Landscapes of amiE on Multiple Substrates

We first developed growth selections for three short-chain aliphatic amides: acetamide (ACT), propionamide (PR), and isobutyramide (IB) (**Figure 3.1**), such that only *E. coli* cells harboring a functional *amiE* gene product can grow when an amide is provided as the sole nitrogen source in selective minimal growth media[30,31]. Following passive diffusion into cells, amiE catalyzes hydrolysis of the amide to its corresponding carboxylic acid, liberating ammonium (a bioavailable nitrogen source). To allow variants supporting higher ammonium flux to become enriched in the population relative to wild-type, we tuned amiE expression levels by screening

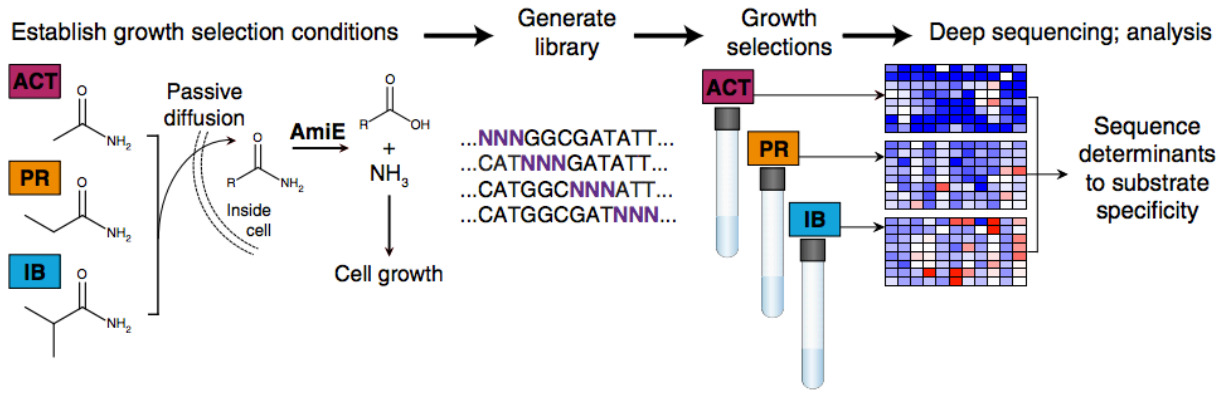


Figure 3.1: Experimental overview. Growth selections for acetamide (ACT), propionamide (PR), and isobutyramide (IB) were established. Amides passively diffuse into host cells harboring amiE variants that produce ammonia necessary for cell growth. Comprehensive site-saturation mutagenesis libraries of amiE were made and selected in media containing an amide as the sole nitrogen source. The pre- and post-selection populations from each selection were deep sequenced and each variant was assigned a fitness metric (ζ) value.

synthetic, insulated constitutive promoters[32,33] such that the specific growth rate in selection media relative to that in defined minimal media ($\mu_{S,wt}/\mu_{M9,wt}$) is 0.4-0.6[34]. Promoter proK14 with the high translational efficiency RBS from gene 10 of T7 bacteriophage (t7RBS) had a suitable $\mu_{S,wt}/\mu_{M9,wt}$ at 0.54 ± 0.11 for IB selection media (plasmid pEDA6_amiE, **Figure 3.2** and **Table B 3.1**). However, the weakest promoter of the set, proK17, had a $\mu_{S,wt}/\mu_{M9,wt}$ of 0.92 ± 0.05 for ACT.

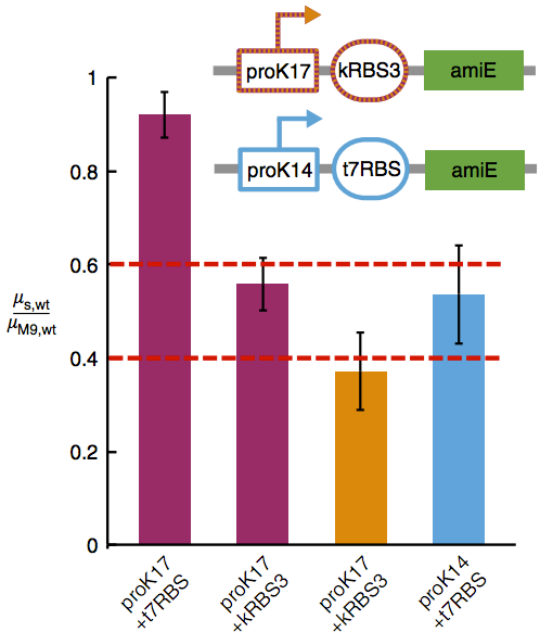


Figure 3.2: **Establishing growth-based selection conditions**. amiE expression was tuned using promoter and RBS engineering. $\mu_{S,wt}/\mu_{M9.wt}$ is the ratio of growth rate of wild-type amiE harboring cells in selection media ($\mu_{S,wt}$) to M9 minimal media ($\mu_{M9,wt}$). Error bars represent 1 s.d. of $n = 4$, 12, 10, and 12 independent measurements. Inset represents the final promoter/RBS combination used for each selection.

To further decrease protein expression, plasmids containing an altered RBS were tested. One construct containing promoter proK17 and a knockdown RBS 3 (kRBS3) sequence had a $\mu_{S,wt}/\mu_{M9,wt}$ of 0.56 ± 0.06 for ACT and 0.37 ± 0.08 for PR (plasmid pEDA2_amiE, **Figure 3.2** and **Table B 3.1**).

A significant concern with this growth selection is the potential for cells containing a non-functional enzyme variant to propagate in a population by acquiring ammonium that has leaked into the extracellular medium. We assessed the risk of such 'cheating' by competing wild-type amiE on the ampicillin-resistant expression construct described above (pEDA2_amiE) against a catalytic knockout, amiE_C166S, with kanamycin-resistance on an otherwise identical expression construct (pEDK2_amiE_C166S), in acetamide selection medium containing no antibiotics. *E. coli* cells harboring either pEDA2_amiE or pEDK2_amiE_C166S were mixed in equal proportion

and competed for 4.12 and 4.17 generations for replicates 1 and 2, respectively. Cells from the pre- and post-selection populations were dilution plated on ampicillin and kanamycin containing plates, and the resulting colonies counted to calculate the frequency of each member in the pre- and post-selection populations. Using the fitness equations laid out in Kowalsky et al.[34], the average fitness metric for the C166S mutant was -2.40 ± 0.9 (n = 2), close to the fitness metric expected if no cheating occurred (-2.46). Thus, we conclude that non-functional variants minimally propagate under the conditions of the selection.

Next, we used PFunkel mutagenesis[35] to construct comprehensive single-site saturation mutagenesis amiE libraries and transformed them into *E. coli* MG1655 rph$^+$. We carried out growth selections for each of the three substrates for approximately 8 generations, starting with an initial population size of >6x10^6 cells. Deep sequencing of the pre- and post-selection populations was used to determine a relative fitness metric (ζ_i) for each amiE variant i, defined as[34]:

$$\zeta_i = \log_2\left(\frac{\mu_{s,i}}{\mu_{s,wt}}\right) \qquad (1)$$

The pre-selection populations were comprised of >51.8% single nonsynonymous mutations and represented >96.3% of the 6820 possible single nonsynonymous mutations for all libraries (**Table B 3.2, Figure B 3.1)**. Given that the read counts per variant in the pre-selection population were log-normally distributed (**Figure B 3.1**) and underrepresented variants could show biased fitness metrics, we calculated Pearson's product moment correlation coefficients for pre-selection read counts and fitness and found them to be to 0.047, 0.033, and -0.064 for the ACT, PR and IB selections, respectively (**Figure B 3.2**). This confirmed that resulting fitness metrics were not biased by a wide distribution of pre-selection read counts. Furthermore, we determined a lower bound fitness metric for each selection that can be discriminated based on depth of sequencing coverage, such that while below this value fitness effects can be described

categorically as 'deleterious', the quantitative effect cannot be reliably predicted. The lower bounds were found to be -1.3, -0.8, and -0.6 for the ACT, PR, and IB selections, respectively. Heat map representations of the local fitness landscapes for each selection can be found in **Figures B 3.3-3.5.**

We tested the validity of using deep sequencing to reconstruct fitness in multiple ways. First, we performed replicate growth selections using the same pre-selection library. The resulting two post-selection libraries were prepared for sequencing in parallel attaching unique Illumina barcodes to each, and normalized fitness metrics were calculated for each replicate. To assess whether the selection results were reproducible we calculated the Pearson product moment correlation coefficients of fitness metrics between replicates and found them to be 0.661, 0.842, and 0.889 for the ACT, PR, and IB selections, respectively ($P < 2.2 \times 10^{-308}$, n = 6627, 6630, and 6569). When we excluded variants with fitness metrics below the lower bounds the correlation coefficients improved to 0.932, 0.949, and 0.943 for the ACT, PR, and IB selections, respectively ($P < 2.2 \times 10^{-308}$, n = 3834, 2954, 4977, **Figure 3.3a** and **Figure B 3.6)**.

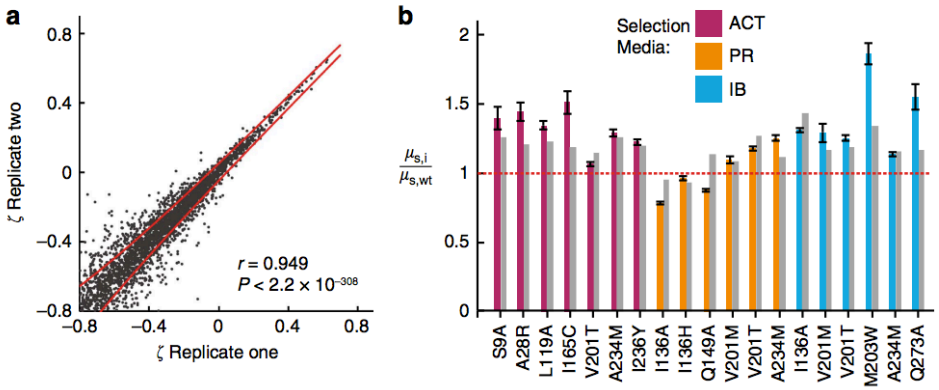


Figure 3.3: **Validation of deep sequencing results.** A.) Fitness metrics from replicate growth selections in the propionamide selection (n = 2954). Red lines indicate two standard deviations from theoretical error estimation[34]. The reported p-value for the Pearson's product moment correlation coefficient was calculated using a two-tailed t-test. B.) Comparison of relative growth rates calculated from the selection experiments (grey bars) and isogenic growth rate assays (colored bars). Error bars represent 1 s.d. of at least four independent measurements.

Second, we compared relative isogenic growth rates ($\mu_{S,i}/\mu_{S,wt}$) to deep sequencing-calculated growth rates for a set of mutations (**Figure 3.3b**, **Table B 3.3**). Deep-sequencing derived fitness corresponded to increased growth rates for 16/17 beneficial mutations, near wild-type growth rates for 2/2 neutral mutations, and no growth for 1/1 deleterious mutation tested. To confirm that improved growth rates were a result of increased flux through amiE, we performed lysate activity assays for a subset of these variants and found that all samples save one improved flux relative to wild-type (**Table B 3.3**).

The Distribution of Fitness Effects (DFE) of amiE

The DFE, both at the organismal and protein level, demarcates evolution[5,7,36]. Specifically, the DFE for a protein is related to its evolvability: the number and type of available beneficial mutations for a new function compared with effects on existing functions is illustrative of how natural proteins evolve. While theoretical and experimental work has advanced our understanding of the available pathways for adaptive molecular evolution[4,6,37–44], the exact form of the distribution, which determines these pathways, is still a subject of debate. **Figure 3.4a** shows the DFE for the three selections. For each, nonsense mutations had a median fitness metric below the detection limit of the deep sequencing method. Nonsense mutations with increased fitness metrics (ζ>0.15) cluster in the last 19 residues of the C-terminus, a relatively unstructured region likely to have no influence on catalytic activity, suggesting that translation of these residues plus the C-terminal His_6-tag used for purification is deleterious to fitness. Missense mutations were on average deleterious for the ACT and PR selections, with 75.8 and 74.2% of variants yielding at least 20% reduction in growth rate relative to wild-type, respectively. By contrast, only 45.4% fell below this threshold for the IB selection.

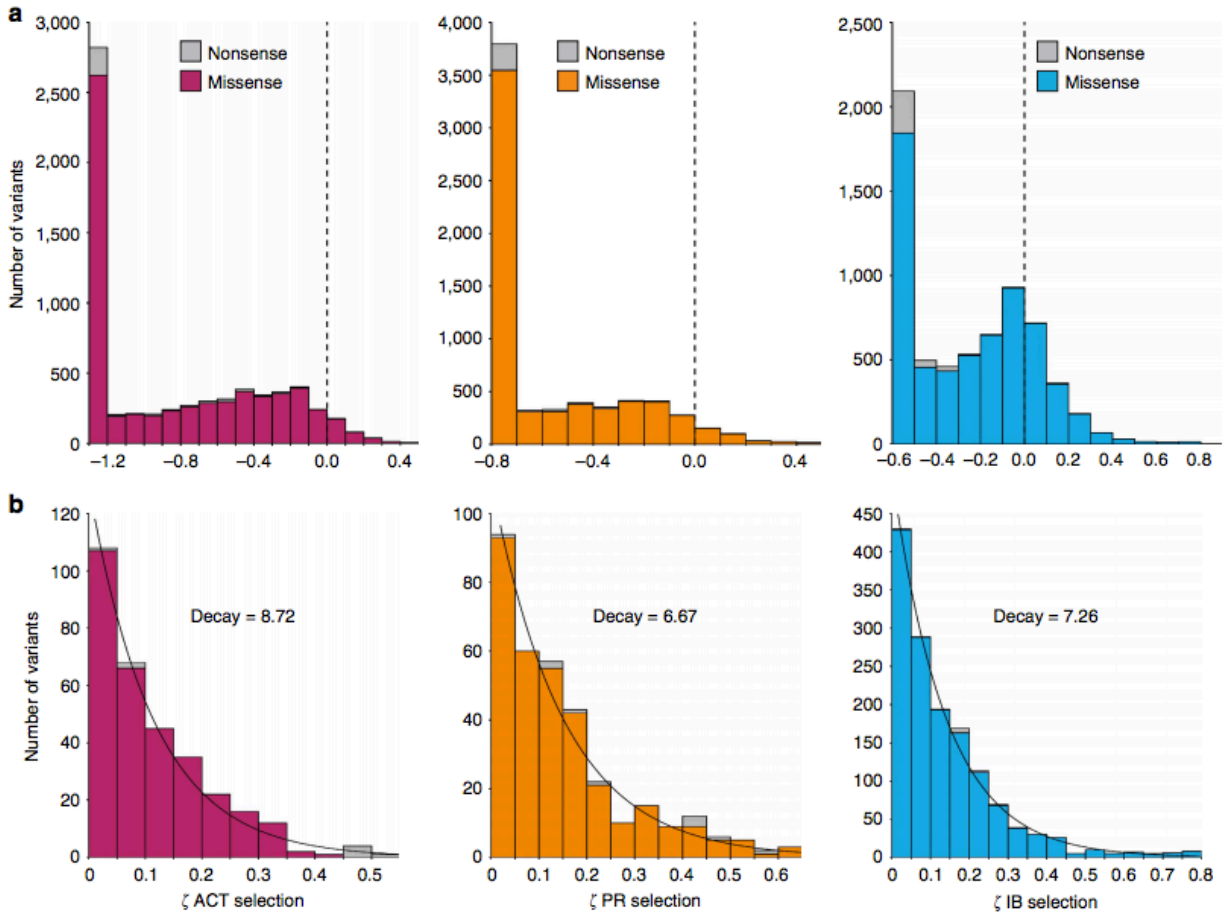


Figure 3.4: Distribution of fitness effects (DFE) are exponentially distributed for beneficial mutations. A.) The DFE of missense and nonsense mutations for ACT (cranberry), PR (orange), and IB (cyan) growth selections. The dashed vertical line demarcates the wild-type fitness metric (ζ=0.0). B.) DFE for beneficial mutations identified in the ACT, PR, and IB selections, respectively. Overlain curves are best-fit exponential distributions estimated from the data.

Remarkably, 21.5% (n = 1394) of missense mutations had above wild-type fitness metrics for the IB selection, with 483 (7.5%) variants having at least 10% increased growth rate (ζ>0.15). There were appreciably less enhanced variants found in the ACT and PR selections, with 4.7 and 5.1% (n = 306 and 328) having fitness metrics above wild-type, respectively.

Modern theories of adaptive molecular evolution predict the DFE for beneficial mutations is scale-free and exponentially distributed[6,24]. However, the available experimental data is

conflicted[25], and most studies have low statistical power due to the rare nature of beneficial mutations. While synthetically constructed, our competitive growth selection results yield fitness metrics for a large effective population size, and the hundreds of beneficial mutations observed provides high statistical power for model fitting. Predictions of beneficial DFE are derived from extreme value theory that describes many distributions falling under the umbrella of the generalized Pareto distribution (GPD)[24]. GPD includes three domains of attraction defined by their shape parameter (κ): Gumbel (κ=0), Fréchet (κ>0), and Weibull (κ<0). We first performed bootstrap goodness of fit tests to a GPD and concluded a failure to reject the null hypothesis that the datasets belonged to a GPD ($P < 0.066$) and estimated κ to be -0.292, -0.309, and -0.195 for the ACT, PR, and IB datasets, respectively. This finding indicates that the tail behavior for the observed beneficial DFE for amiE is slightly truncated, yet our results are consistent with the predictions of Orr that if departures from the Gumbel domain are observed they will be minimal ($-1/2<\kappa<1/2$)[6].

We next conducted log-likelihood ratio tests for fitted exponential distributions (null hypothesis) against fitted gamma and Weibull distributions (alternative hypotheses) for the DFE of beneficial mutations (**Figure 3.4b**). These alternative models were chosen as previous empirical studies have observed tail behavior indicative of these such distributions[40,45,46]. We concluded a failure to reject the null hypothesis for the IB dataset, yet found that the ACT dataset best fit a Weibull distribution ($P = 0.05$) and that gamma and Weibull were both better fits for the PR dataset ($P = 0.023$ and 0.039, respectively, **Table 3.1**). Interestingly, one-sample Anderson-Darling tests for goodness-of-fit to each distribution indicated a failure to reject the null hypothesis that the data fit any of the distributions (**Table 3.1**). To assess the null hypothesis that the three data sets came from a single, statistically indistinguishable distribution, we performed a k-sample Anderson-

Darling test and concluded they were not from a single distribution ($P = 0.0124$). Thus, all datasets can be described as exponentially distributed, though the ACT and PR datasets best fit the higher parameter models.

Table 3.1: Model fitting results for distribution of beneficial mutations

		ACT	PR	IB
Exponential	rate	8.72	6.67	7.26
	A-D test *P*-value	0.527	0.338	0.635
	LL	365.9	303.2	1382.6
Gamma	shape	1.14	1.17	1.03
	rate	9.90	7.81	7.46
	A-D test *P*-value	0.834	0.459	0.723
	LL	367.4	305.8	1382.9
Weibull	shape	1.094	1.094	1.012
	scale	0.119	0.155	0.138
	A-D test *P*-value	0.851	0.451	0.712
	LL	367.8	305.3	1382.8
Log-likelihood ratio tests				
H_0	**H_A**			
Exponential	Gamma	3.11	5.14	0.62
	P-value	0.078	0.023	0.43
Exponential	Weibull	3.8	4.3	0.31
	P-value	0.050	0.039	0.58

Beneficial Mutations result from Protein, not mRNA, effects

We addressed whether effects at the mRNA level could explain beneficial mutations, as variants can achieve higher fitness by increasing total active amiE concentration through improvements to the rate of transcription, the degradation rate of mRNA, and the efficiency of translation. The fitness metrics of synonymous codons for beneficial mutations (ζ>0.15) showed

low variance in most cases except near the N-terminus (**Figure B 3.7**). A recent mRNA model[47] could explain up to 5% of the variance in the first 15 residues but only 0.2% of the variance over the entire sequence length (**Table B 3.4**). We conclude that the observed fitness effects are the result of changes at the protein level, not at the mRNA level.

Comparison of DFE Between Selections

Promiscuous activity of enzymes is believed to be the driving force of evolution towards new activities[3]. Our fitness maps allow us to address the question of how mutations impact fitness in multiple substrate backgrounds. At the outset of this work, we anticipated that the majority of 'hits' or beneficial mutations would be shared across selections. This null hypothesis is grounded in the biophysical argument that most beneficial mutations would improve protein expression, not activity, and these would be beneficial regardless of the substrate selected on. Additionally, we anticipated that the pool of mutations available for improving activity for a single substrate would predominately localize to the vicinity of the active site, thus rendering few specificity-altering mutations. Consequences of this prediction are that there should be significant correlation of fitness between amides, with specificity-determining mutations localized near the active site.

We first assessed whether there was a significant correlation of fitness between different amides (**Figure 3.5a**). Correlation for the ACT and PR selections ($r = 0.827$, $P < 2.2\text{x}10^{-308}$) was notably higher than that for the IB and ACT ($r = 0.317$, $P = 8.6 \text{ x } 10^{-85}$) or IB and PR selections ($r = 0.367$, $P = 6.7 \text{ x } 10^{-95}$). PCA revealed that a single principal component could explain 96.8% and 87.8% of the variance of the ACT and PR datasets, respectively, while two principal components are sufficient to explain over 99% of the variance for the IB dataset (**Figure 3.5b, Figure B 3.8**). These results are inconsistent with our null hypothesis, pointing towards global alterations in the

protein structure to adapt to different substrates. Of note, the fitness data is non-normal and could be analyzed with other multivariate statistical analysis methods.

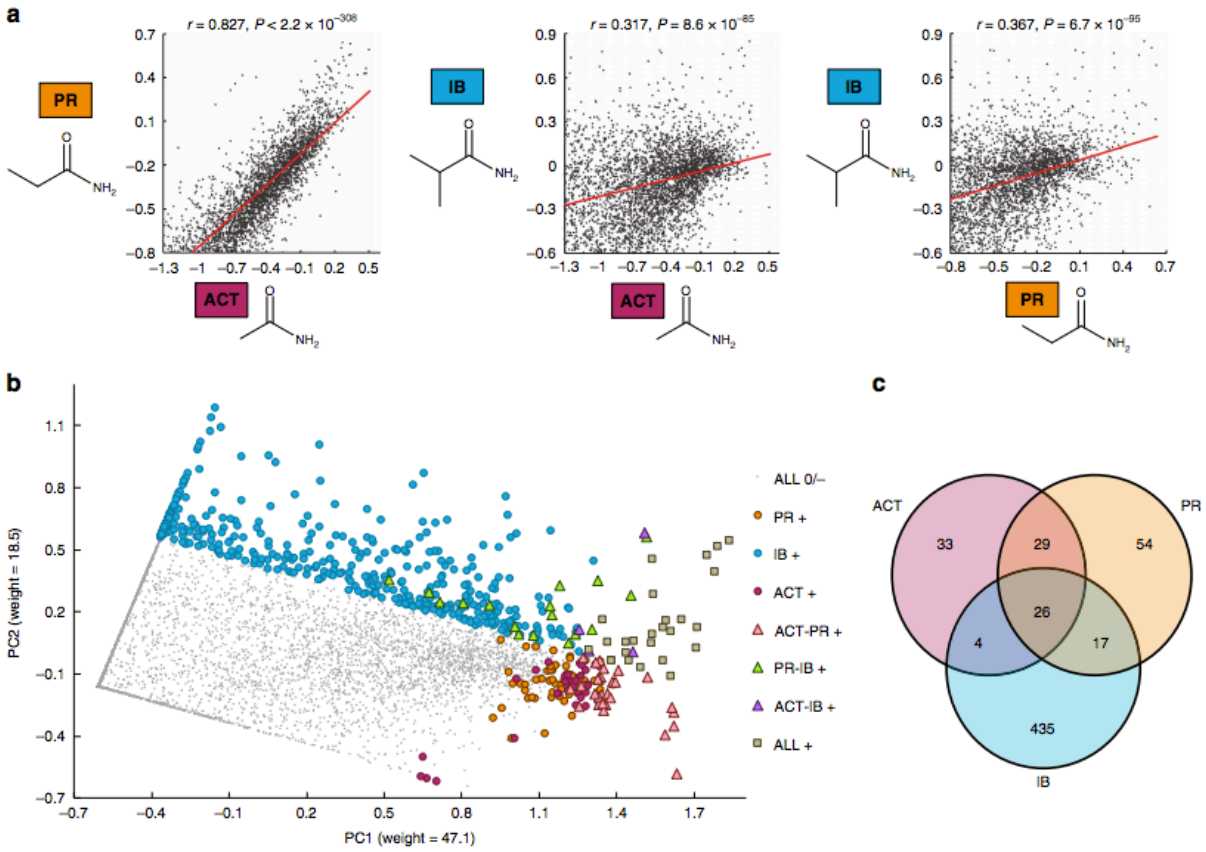


Figure 3.5: Correlative analysis of fitness effects. A.) Correlation of variant fitness metrics between selections. Variants with fitness metrics above the lower bounds are compared between each selection. Plots represents n = 3054, 3600, and 2959 points for panels ACT vs PR, ACT vs. IB, and PR vs. IB, respectively. The reported p-value for the Pearson's product moment correlation coefficient was calculated using a two-tailed t-test. B.) Principal component analysis of substrate-specific fitness effects. Black dots show common neutral and deleterious mutations, while substrate-specific beneficial mutations (ζ>0.15) are colored according to 7 bins. C.) 3-way Venn diagram representing 7 specificity bins.

Restricting our correlative analysis to only beneficial mutations (ζ>0.15) revealed that fitness-enhancing mutations for ACT were, on average, likely to be beneficial for PR (mean ζ = 0.236). By contrast, beneficial mutations for IB were likely to be deleterious in both the ACT and

PR selections (mean ζ = -0.480 and -0.319). This result is consistent with the findings of Stiffler et al. that beneficial mutations for a new or less evolved function are likely to be deleterious for existing functions when the selections pressures are high. Furthermore, IB-beneficial variants showed essentially no correlation for fitness in the ACT and PR selections (r = 0.0617 and 0.164, respectively). This finding indicates that, at least for amiE, predicting hits based on known fitness effects for a given substrate cannot be accomplished through correlative analysis.

We next analyzed the relationship between beneficial (ζ>0.15) and specificity-determining (ζ>0.15 for one amide and ζ<0 for the other two substrates) mutations and their distance to the catalytic active site. Distance was measured by the minimum distance from the alpha-carbon of positions with beneficial mutations to any active site atom (six identical active sites in the functional homohexamer). The mutations were placed in 3 Å bins that were normalized to total available mutations in each distance shell. For beneficial mutations, we found that most were >15 Å from the active site for the ACT and PR variants, while the IB variants were mostly 9-21 Å away (**Figure 3.6a**). Strikingly, we found very few specificity-determining mutations for the ACT and PR selections (n = 6 and 14, respectively), with variants distanced by 6-15 Å for ACT and >14 Å for PR (**Figure 3.6b**). By contrast, we found 395 specificity-determining mutations for IB, which were distributed similarly to the set of all IB-beneficial mutations. Thus, beneficial and specificity-determinant positions are globally dispersed throughout the primary sequence and structure of amiE.

Biophysical Characterization of Beneficial Mutations

To understand the biophysical basis underlying beneficial mutations, we expressed, purified, and characterized a set of 11 variants chosen in part on their ability to predict larger sets of beneficial

variants (**Table 3.2, Supplementary Fig. 9**). For example, globally beneficial mutation S9A was chosen because it could potentially explain other N-terminal beneficial mutations. For all variants save one (see **Methods**), apparent melting temperatures ($T_{m,app}$) were within 7°C of the wild-type $T_{m,app}$ of 67.7 ± 0.1°C, indicating that differences in thermal stability are unlikely to explain *in vivo* beneficial fitness effects.

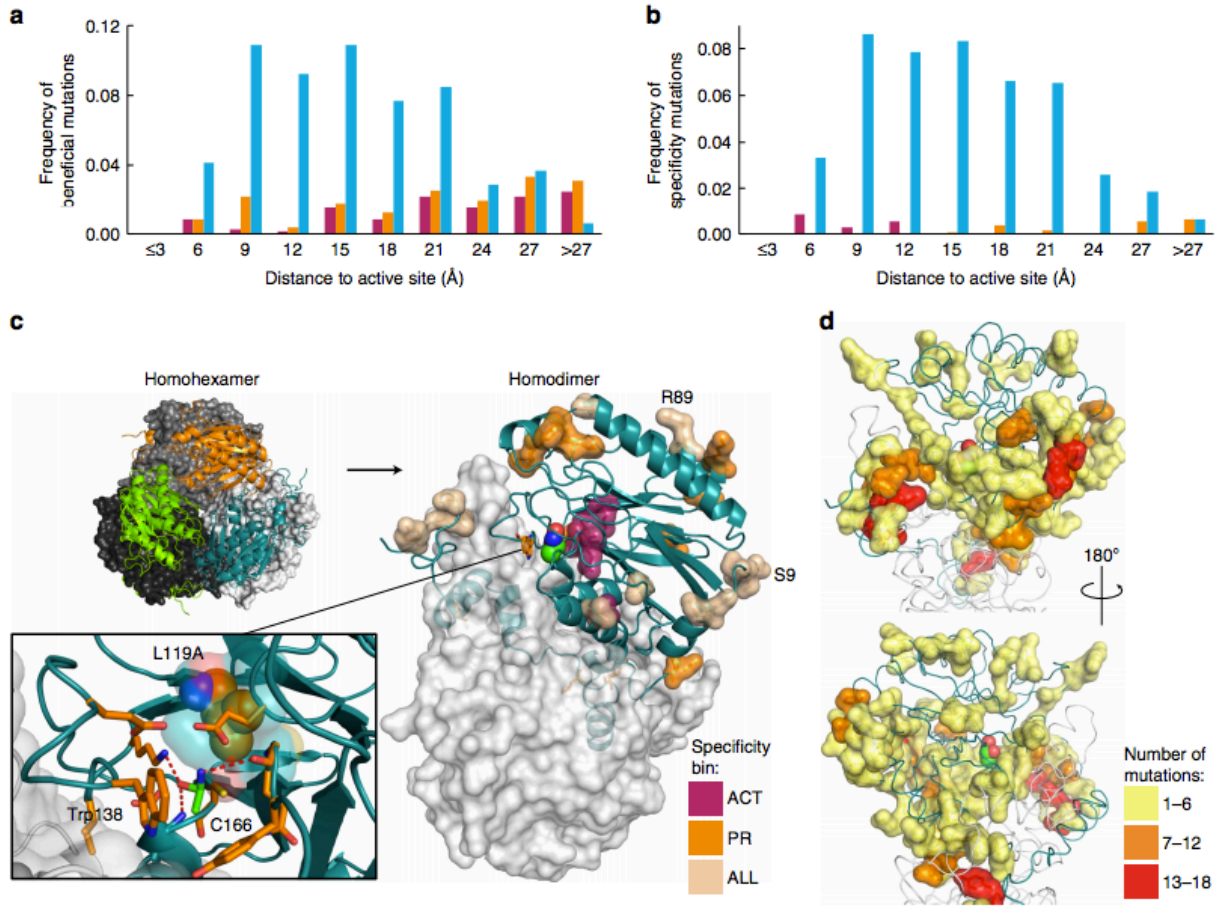


Figure 3.6: **Substrate specificity is globally encoded.** A-B.) Frequency of beneficial (A) and specificity-determining (B) mutations as a function of distance to active site. C.) Beneficial mutations for all selections and specificity-determinant mutations for ACT and PR selections mapped onto the structure of amiE. The inset illustrates the catalytic active site. D.) Specificity-determinant mutations for IB selection colored by number found at given position.

To evaluate commonalities between beneficial mutations, we sorted variants into seven possible bins for beneficial fitness metrics (ζ>0.15 for given selection(s) and ζ<0.15 in other

selection(s), **Figure 3.5c**). 21 of 26 beneficial mutations common to all three selections were found at extreme N- or C-terminal residues. Of the remaining five, we characterized R89E, a surface mutation located over 20 Å away from the active site that yielded an increase in relative k_{cat}/K_m of 1.96 ± 0.59 and 1.42 ± 0.42 for PR and IB substrates, respectively (**Figure 3.6c**). Alternatively, shared N-terminal mutation S9A had slightly reduced relative k_{cat}/K_m. Thus, even for a highly stable protein like amiE, we found few mutations like R89E that can generally increase k_{cat} or K_M and increase fitness.

Beneficial mutations shared in two of the three selections were scarce. 18/29 mutations shared between ACT and PR cluster at the extreme N- or C- termini. The 17 PR+IB specific mutations cluster at Q273, a 2nd shell residue that buttresses W138 at the active site, and at M202 located 14 Å to the active site. Variant M202H showed over 2.5-fold increase in relative k_{cat}/K_m for IB and PR, but Q273A did not show increased catalytic efficiency *in vitro*. We speculate the conditions required by the enzyme assay for sensitive ammonia detection prohibited the recapitulation of *in vivo* kinetics.

Four ACT-specific mutations encoded smaller substitutions (A/C/S/V) at position L119, a residue that supplies hydrophobic packing behind the catalytic nucleophile C166 10 Å from the active site (**Figure 3.6c**). L119A showed a 2.2 ± 0.1–fold increase in k_{cat} relative to wild-type with a compensatory increase in K_M.

In stark contrast to ACT, there were 435 IB-specific and 395 specificity-determining mutations for IB distributed throughout the protein structure (**Figure 3.6d**). Substitution W138A/G decreases van der Waals area in the vicinity of the amide transition state, allowing accommodation of the bulky isobutyrl group. However, most specificity-altering mutations were located far from the active site. Hot spots of positions where 5 or more specificity-determining mutations confer

Table 3.2: Wild-type and variant amiE biophysical data.

Variant	ACT ζ	PR ζ	IB ζ	K_m (mM) / $K_{m,wt}$ (mM)	k_{cat} (s^{-1}) / $k_{cat,wt}$ (s^{-1})	k_{cat}/K_m (M^{-1} s^{-1}) / $k_{cat,wt}/K_{m,wt}$ (M^{-1} s^{-1})	$T_{m,app}$ (°C)
Wildtype	0.00	0.00	0.00	4.7 ± 0.5 52.7 ± 8.3 297.2 ± 54.8	59.0 ± 2.0 144.7 ± 9.9 13.3 ± 1.1		67.7 ± 0.1
S9A	0.33	0.41	0.36	2.3 ± 0.6 4.4 ± 1.1 0.5 ± 0.1	0.6 ± 0.1 1.6 ± 0.2 0.4 ± 0.0	0.28 ± 0.1 0.36 ± 0.13 0.76 ± 0.27	63.1 ± 0.1
A28R	0.27	0.10	0.11	nd	nd	nd	67.7 ± 0.1
R89E	0.30	0.34	0.15	nd 0.7 ± 0.1 0.8 ± 0.2	nd 1.4 ± 0.1 1.1 ± 0.1	nd 1.96 ± 0.59 1.42 ± 0.42	66.7 ± 0.1
L119A	0.30	-0.80	-0.60	2.8 ± 0.4 not active 2.5 ± 0.6	2.2 ± 0.1 not active 1.3 ± 0.2	0.8 ± 0.16 not active 0.52 ± 0.18	67.4 ± 0.2
I165C	0.25	-0.27	-0.32	2.6 ± 0.4 0.7 ± 0.2 0.7 ± 0.2	1.7 ± 0.1 0.6 ± 0.1 0.8 ± 0.1	0.66 ± 0.14 0.79 ± 0.24 1.22 ± 0.65	65.7 ± 0.1
V201M	0.37	0.12	0.22	10.2 ± 2.0 0.8 ± 0.4 4.6 ±1.3	1.1 ± 0.1 0.1 ± 0.0 0.8 ± 0.1	0.11 ± 0.03 0.13 ± 0.10 0.18 ± 0.08	61.2 ± 0.1
V201T	0.20	0.34	0.25	1.6 ± 0.3 nd 3.8 ± 0.8	1.0 ± 0.1 nd 1.4 ± 0.1	0.65 ± 0.16 nd 0.37 ± 0.11	64.9 ± 0.2
M202H	-0.08	0.16	0.43	0.9 ± 0.1 0.7 ± 0.1 0.4 ± 0.1	1.3 ± 0.0 1.8 ± 0.1 1.4 ± 0.1	1.4 ± 0.23 2.55 ± 0.72 3.08 ± 1.05	63.0 ± 0.1
M203W	-1.30	-0.80	0.43	nd	nd	nd	nd
A234M	0.33	0.15	0.21	2.8 ± 0.4 0.5 ± 0.2 3.6 ± 0.9	1.0 ± 0.0 0.3 ± 0.0 0.8 ± 0.1	0.35 ± 0.07 0.58 ± 0.32 0.23 ± 0.08	64.4 ± 0.1
Q273A	-0.71	0.31	0.23	4.9 ± 2.0 2.3 ± 0.6 0.6 ± 0.2	0.2 ± 0.0 0.5 ± 0.1 0.4 ± 0.0	0.05 ± 0.03 0.20 ± 0.08 0.71 ± 0.28	69.7 ± 1.1

increased fitness occur at the N- and C-terminus (residues H3, S7, T323, R324, T327, V329, and C332-V334), as well as P50, C139, I174, A196, K197, V201, M202, W209, N212, F223, S228, G247-E249, G252, Q271, Q273-H275, and Y284. Interestingly, hot spot positions 197 through 212 are located on an alpha helix located at least 12 Å from the active site that contacts the dimeric interface. As these mutations do not benefit all substrates, we hypothesize that mutations at these positions cause rigid body motion of the helix to yield subtle geometric rearrangement, if not large-scale disruption, of the active site that favors IB catalysis. We tested V201M/T for activity on IB and, contrary to expectations, found a decrease in relative k_{cat}/K_m. We speculate that the mismatch between expected and measured catalytic efficiency results from hexamer dissociation caused by the low enzyme concentration required by the activity assay, as the lysate assays showed an increase in velocity for the V201T mutant.

DISCUSSION

In this contribution, we generated single-mutation protein fitness landscapes for an amidase on three different substrates. In contrast to studying protein-protein interactions, the application of deep mutational scanning to enzymes has been limited by the difficulty in developing generalizable high-throughput functional assays, as the nature of enzyme function is highly diverse. Regardless, exhaustive mutational studies permit a glimpse into how natural enzymes evolve for new functions. Our results show that, at least for amiE, mutations which are beneficial for only one substrate are 1.) not confined to vicinity of the active site and 2.) cannot be predicted based on known fitness for another substrate.

In terms of predicting fitness, we conclude that single-mutation fitness landscapes are highly substrate dependent, which is consistent with previous works[8,38,48,49]. However, this work

provides a unique perspective of comparing two structurally similar substrates, ACT and PR, to the dissimilar IB substrate. Not surprisingly, we found the IB single-mutation fitness landscapes to be the most divergent, signaling that at the biophysical level the requirement to accommodate the larger IB substrate significantly alters the mutational landscape. The percentage of beneficial mutations observed is consistent with previous deep mutational scanning experiments on enzymes[8,37,38,50]. These rates are significantly larger than that predicted for the natural adaptation of organisms[51] because in deep mutational scanning experiments a strong selection is imposed upon a gene that is, by experimental design, intended to influence only a single phenotypic trait[4]. This intention to mitigate pleiotropy is especially true with bulk growth competition experiments, and it should be noted that randomly drifting populations contain genes that do not ascribe to such constraints.

Our results strengthen the theoretical case that fitness for beneficial mutations is approximately exponentially distributed even though the percentage of beneficial mutations differs substantially between substrates. We note that this exponential distribution holds even for the IB selection which presumably causes large-scale rearrangements of the active site to allow better access to the branched chain IB. Other studies have explored the mechanics of multiple steps and epistasis[39,42,43,52,53]. In this work, we considered only single steps in the local fitness landscape. Thus, the generality of our observations for multiple steps remain to be seen.

For the design and engineering of substrate promiscuous or specific biocatalysts, knowledge of the sequence and spatial distribution of 'hits' is imperative. Our findings indicate that, at least for amiE, most substrate-determining mutations for new functions, in this case IB, localize approximately 9-24 Å from the active site. Together, these results have strong implications for design and engineering of substrate promiscuous biocatalysts because it suggests current

strategies of iterative site saturation mutagenesis near the active site are sub-optimal[1]. Additionally, computational design algorithms focused solely on the modifying 1st and 2nd shell mutations around the active site need to be revisited.

MATERIALS AND METHODS

Reagents

All chemicals were purchased from Sigma-Aldrich unless specified otherwise. All primers and mutagenic oligonucleotides were designed using the Agilent QuikChange Primer Design Program (www.agilent.com) and were ordered from Integrated DNA Technologies. Propionamide and isobutyramide solids were recrystallized from ethyl acetate and water, respectively.

Plasmid construction

pEDA6_amiE was renamed from pJK_proK14_amiE as described in Bienick et al.[33]. pEDA2_amiE was constructed by Kunkel mutagenesis[54] of pJK_proK17_amiE from Bienick et al.[33] to introduce a knockdown ribosome binding sequence (primer kRBS3). Protein expression constructs were made by subcloning the *amiE* gene from pEDA2_amiE into the pET-29b(+) (Novagen) backbone at the NdeI and XhoI restriction sites following standard protocols. amiE point mutants were created using Kunkel mutagenesis[54]. Primer sequences used in this work are listed in **Table B 3.5**.

Construction of mutagenesis libraries

Eight comprehensive single-site saturation mutagenesis libraries of amiE were constructed (residues 1-85, 86-170, 171-255, and 256-341 on plasmids pEDA2_amiE and pEDA6_amiE)

using PFunkel mutagenesis[35] with modifications as noted[34]. Library cell stocks of the selection strain, *E. coli* MG1655 rph+ [F- λ-] (Coli Genetic Stock Center, #7925), were made essentially as described in Klesmith et al.[50].

Growth selections

Starter cultures for growth selection were prepared as in Klesmith et al.[50], except 1X M9 minimal media lacking ammonium chloride (M9 N^-) was used to wash cell pellets prior to inoculation of selection media. 3 mL of selection media (M9 N^- + 10 mM acetamide, M9 N^- + 15 mM propionamide, and M9 N^- + 10 mM isobutyramide for ACT, PR, and IB selections, respectively) was inoculated to an initial OD_{600} of 0.02 at a volume of 3 mL (>$6x10^6$ cells). To ensure exponential growth during the entire selection experiment, after approximately four generations the cells were harvested, washed with M9 N^-, and a fresh 3 mL culture with selection media was inoculated to the same initial OD_{600} of 0.02 (to maintain the same initial population size). Growth selections were carried out and samples preserved for sequencing as described in Klesmith et al.[50]. Replicates were performed using the same pre-selection population. Based on the high correlation between replicates and the fact that a major source of error in deep sequencing measurements are counting errors (Poisson noise)[34], the fitness metrics used in subsequent analysis were computed by combining reads from the two replicates and repeating the analysis.

Sequencing

Libraries were amplified, barcoded, cleaned, and quantified following Method B as described in Kowalsky et al.[34]. Gene amplification primers are listed in **Table B 3.5**. Pre- and post-selection samples were pooled and sequenced with 300bp PE reads on an Illumina MiSeq available at the

Michigan State University sequencing core. Deep sequencing data was analyzed using Enrich software[55] with modifications as noted in Kowalsky et al.[34] and scripts freely available at Github (https://github.com/JKlesmith/Deep_Sequencing_Analysis).

Normalized fitness metrics for each variant, ζ_i, were determined according to the 'Normalization for Growth Rate Selections' section as outlined in Kowalsky et al.[34]. Briefly, deep sequencing was used to count each library member in the pre- and post-selection populations. For each single nonsynonymous mutation and wild-type an enrichment ratio was calculated by:

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

Where f_{fi} and f_{oi} represent the frequency of member i in the final (post-selection) and initial (pre-selection) populations. Normalized fitness metrics were calculated using the following equation:

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

Where ε_i is enrichment ratio for variant i, g_p is the number of population doublings, and ε_{WT} is the enrichment ratio for wild-type.

Beneficial mutations and lower bounds for fitness metrics

A beneficial mutation was defined as having at least 10% increase in growth rate (ζ>0.15) relative to wild-type. Weighted means for synonymous codon fitness metrics, where the weights were read counts (depth of coverage) for each mutation, were calculated to be 0.03 ± 0.09, -0.02 ± 0.11, and -0.01 ± 0.07 for the ACT, PR, and IB datasets, respectively. A fitness metric of 0.15 was found to be in >90% percentile for all three datasets.

To determine lower-bound fitness metrics for each selection, we first determined the half-median of read counts of the pre-selection library for each selection (63, 49, and 31 for the ACT, PR, and IB selections, respectively). This number was normalized by the ratio of post- to pre-selection read counts (2.97, 1.86, and 2.04 for ACT, PR, and IB selections, respectively). Next, a lower-bound enrichment ratio (ε_{LB}) based on 10 read counts in the post-selection population was calculated:

$$\varepsilon_{LB} = \log_2\left(\frac{10}{f_{LB}}\right) \quad (3)$$

Where f_{LB} represents the normalized half-median pre-selection reads. The lower-bound fitness metrics, ζ_{LB}, was then calculated using 8 population doublings (g_p) and the wild-type enrichment ratio (ε_{WT}):

$$\zeta_{LB} = \log_2\left(\frac{\frac{\varepsilon_{LB}}{g_p}+1}{\frac{\varepsilon_{WT}}{g_p}+1}\right) \quad (4)$$

Distribution fitting of beneficial DFE

Distribution fitting analysis was conducted using R statistical software[56]. Bootstrap goodness of fit and parameter estimation for the generalized Pareto distribution were done using the package gPdtest[57]. Model parameters were approximated and log likelihood values were determined using maximum likelihood estimation with package fitdistrplus[58]. Anderson-Darling tests were performed using the package kSamples[59]. Log-likelihood (LL) ratios were calculated as 2*[(LL H_A) – (LL H_0)], where H_0 = null hypothesis and H_A = alternative hypothesis. P-values were computed from a chi-squared distribution with one degree of freedom.

Protein characterization

Wild-type and variant amiE protein was expressed using Studier auto-induction[60] and purified according to Klesmith et al.[50]. The eluate was buffer exchanged into PBS buffer, pH 7.5 using GE disposable PD-10 desalting columns (GE Healthcare). Purified protein was stored in PBS at 4°C. Wild-type and variant amiE melting temperatures were measured using a SYPRO Orange thermal-shift assay[61,62] as described in Klesmith et al.[50], but in PBS buffer, pH 7.5. Catalytic parameters (K_m and k_{cat}) were assayed at 37°C in PBS buffer, pH 7.5 using a phenol and hypochlorite ammonia detection assay[63]. PCR plates containing 100 μL of 7 different concentrations of amide (highest concentrations were 40, 150, and 800 mM for ACT, PR, and IB activity assays, respectively, with 1:2 serial dilutions for remaining substrate concentrations) in PBS were incubated on a thermocyler block (Eppendorf) with the lid open at 37°C for 5 minutes. To begin the assay, 20 μL of 0.02 μM (ACT and PR assays) or 0.2 μM (IB assays) enzyme was added. At discrete time points, 100 μL of the reaction was removed and quenched by depositing into a clear 96-well plate containing 50 μL phenol nitroprusside solution held on ice. At the end of the last time point, 50 μL alkaline hypochlorite solution was added to all wells and the plate was covered and incubated in a metal bead heat bath for 10 minutes at 35°C. The plate was then transferred to a Synergy H1 spectrophotometer (BioTek) held at 35°C and A_{625} was measured every minute for 15 minutes. Non-linear regression was performed using GraphPad Prism version 6 for Mac OS X, GraphPad Software, La Jolla California USA, www.graphpad.com. All measurements were performed at least in duplicate. The IB specific variant M203W shows increased fitness in the deep sequencing selection but decreased lysate activity compared with wild type. M203W immediately precipitated out when we tried to purify this enzyme. Thus, for this case, lysate activity would not be representative of *in vivo* conditions. For PR variants the coefficient of variation for wild-type was

prohibitively high to calculate statistically significant ratios; note the variance of the other wild-types measurements.

Isogenic growth and lysate flux assays

Starter cultures were prepared by inoculating 2 mL of M9 minimal media + carbenicillin (50 μg/mL) with scrapings of MG1655 rph+ cell stocks harboring pEDA2_amiE or pEDA6_amiE variant plasmids and grown overnight at 37°C with 250 rpm shaking. In the morning, cells were pelleted, washed twice with M9 N^-, and resuspended in 1 mL M9 N^-. 3 mL of selection media + carbenicillin (50 μg/mL) in Hungate tubes was inoculated to a final OD_{600} of 0.02. Cultures were grown at 37°C with shaking at 250 rpm. For growth assays, OD_{600} was measured every 30-45 minutes until a final OD_{600} of approximately 0.5 was reached. All growth rate measurements represent at least 4 biological replicates collected on at least 2 separate dates. Lysate flux assays were adapted from Bienick et al.[33]. 2 mL of exponential phase culture (OD_{600} of approximately 0.15-0.3) was spun down at 15,000xg for 5 minutes. Cell pellets were washed twice and resuspended with 1 mL PBS, pH 7.5. Cells were lysed as described in Bienick et al.[33]. 0.5-0.9 mL of lysate was used in a 1 mL total volume assay containing 10 mM, 15 mM, or 10 mM acetamide, propionamide, or isobutyramide, respectively. The assay was conducted at 37°C. Every 5 minutes, 100 μL of the assay volume was removed and added to a 96-well plate containing 50 μL pre-chilled phenol nitroprusside. At the end of the last time point, 50 μL of alkaline hypochlorite was added to all wells. Absorbance at 625 nm was measured as in Bienick et al.[33].

Data Availability

Full datasets including normalized fitness metrics, pre- and post-selection read counts, and raw log base two enrichment scores for each variant can be obtained from Figshare (https://dx.doi.org/10.6084/m9.figshare.3505901.v2). Raw sequencing reads for this work have been deposited in the SRA (SAMN06237792-SAMN06237827).

APPENDIX

APPENDIX

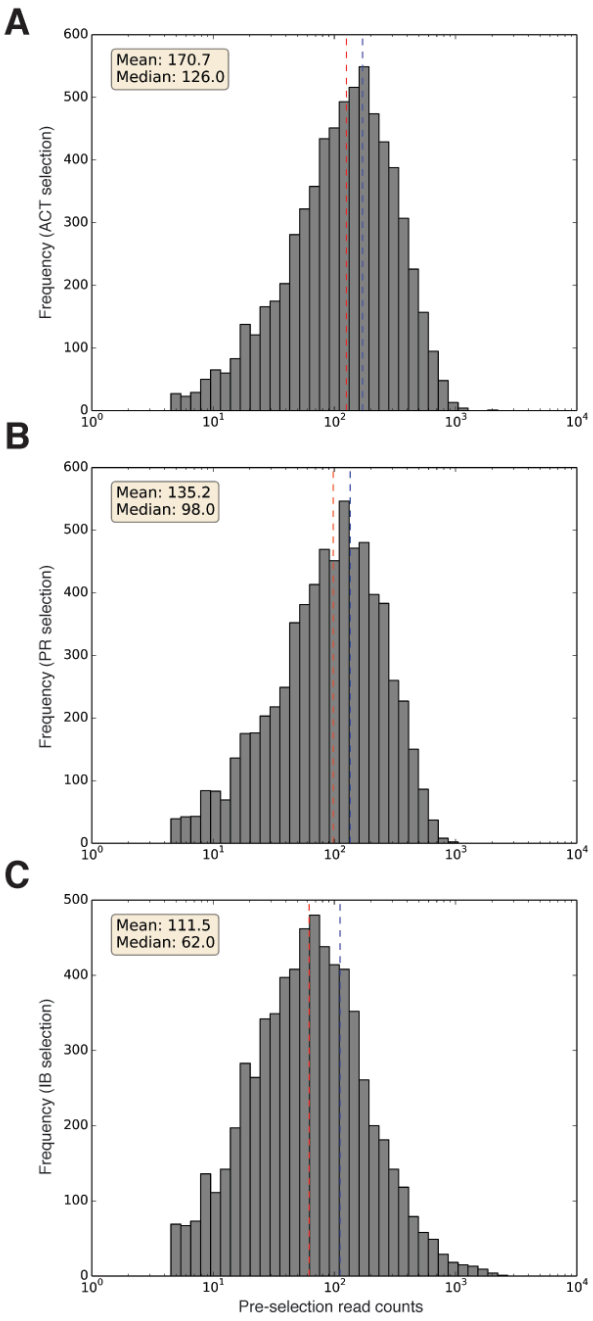


Figure B 3.1: Frequency distribution of library member counts. Panels are for reads in the pre-selection libraries for A.) acetamide B.) propionamide and C.) isobutyramide. Vertical lines indicate median (red) and mean (blue) read coverage.

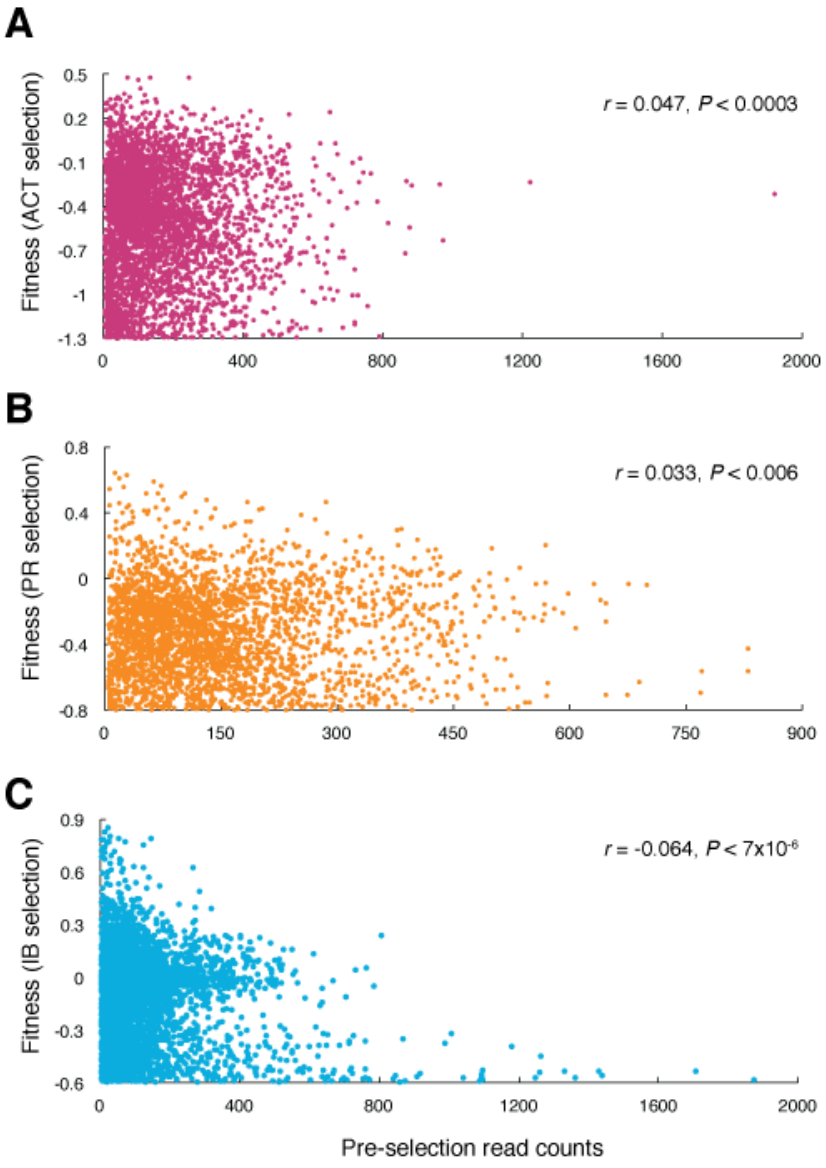


Figure B 3.2: Fitness versus pre-selection read counts. Panels are for each variant in the A.) acetamide B.) propionamide and C.) isobutyramide libraries. Variants with insignificant read counts ($n \leq 5$) and fitness metrics below the lower bounds were excluded from the analysis. Plots represent n = 4037, 3135, and 4969 variants. P-values for Pearson's product moment correlation coefficients were calculated using a two-tailed t-test.

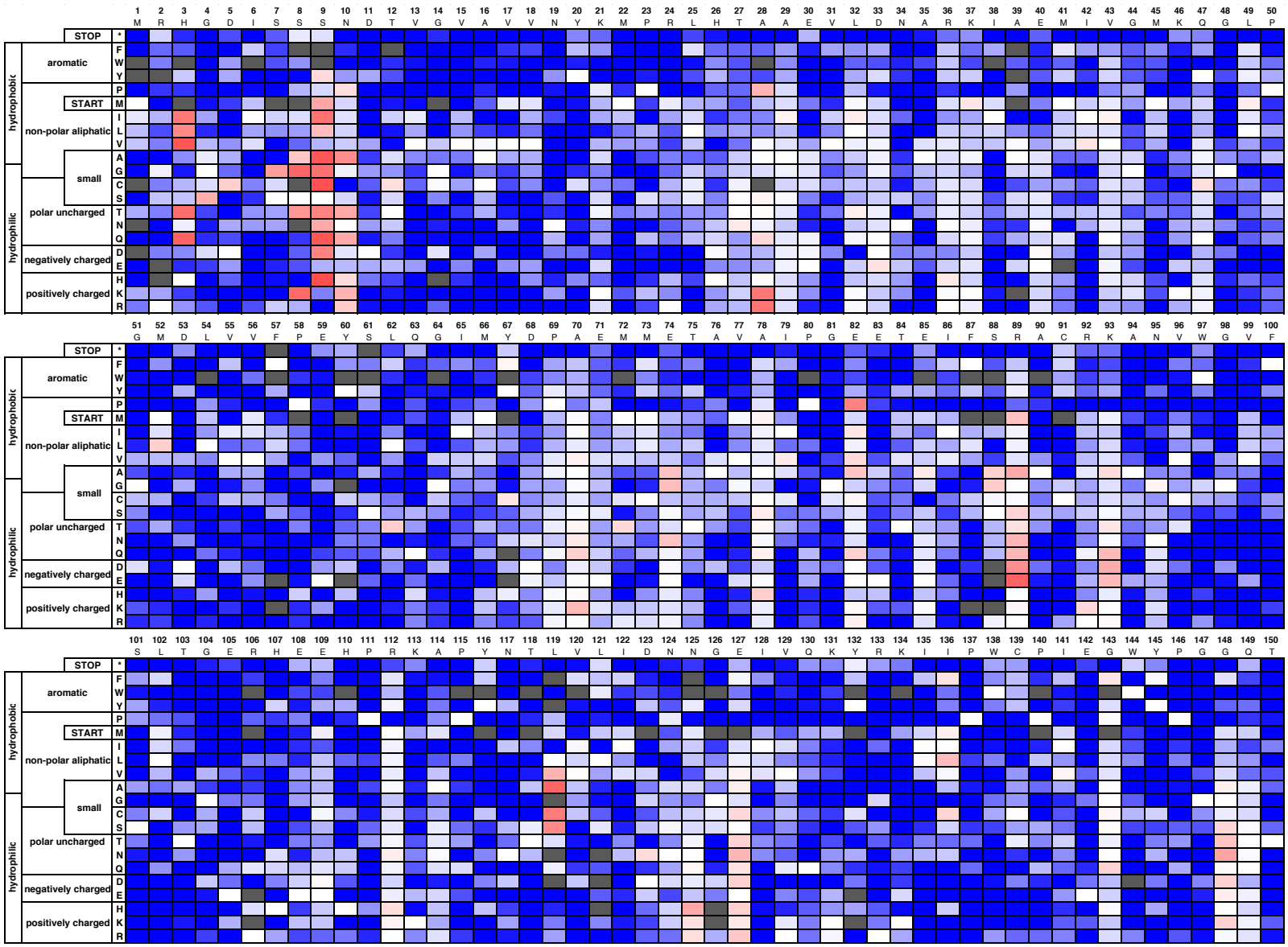


Figure B 3.3: Fitness landscape for acetamide selection.

Figure B 3.3 (cont'd)

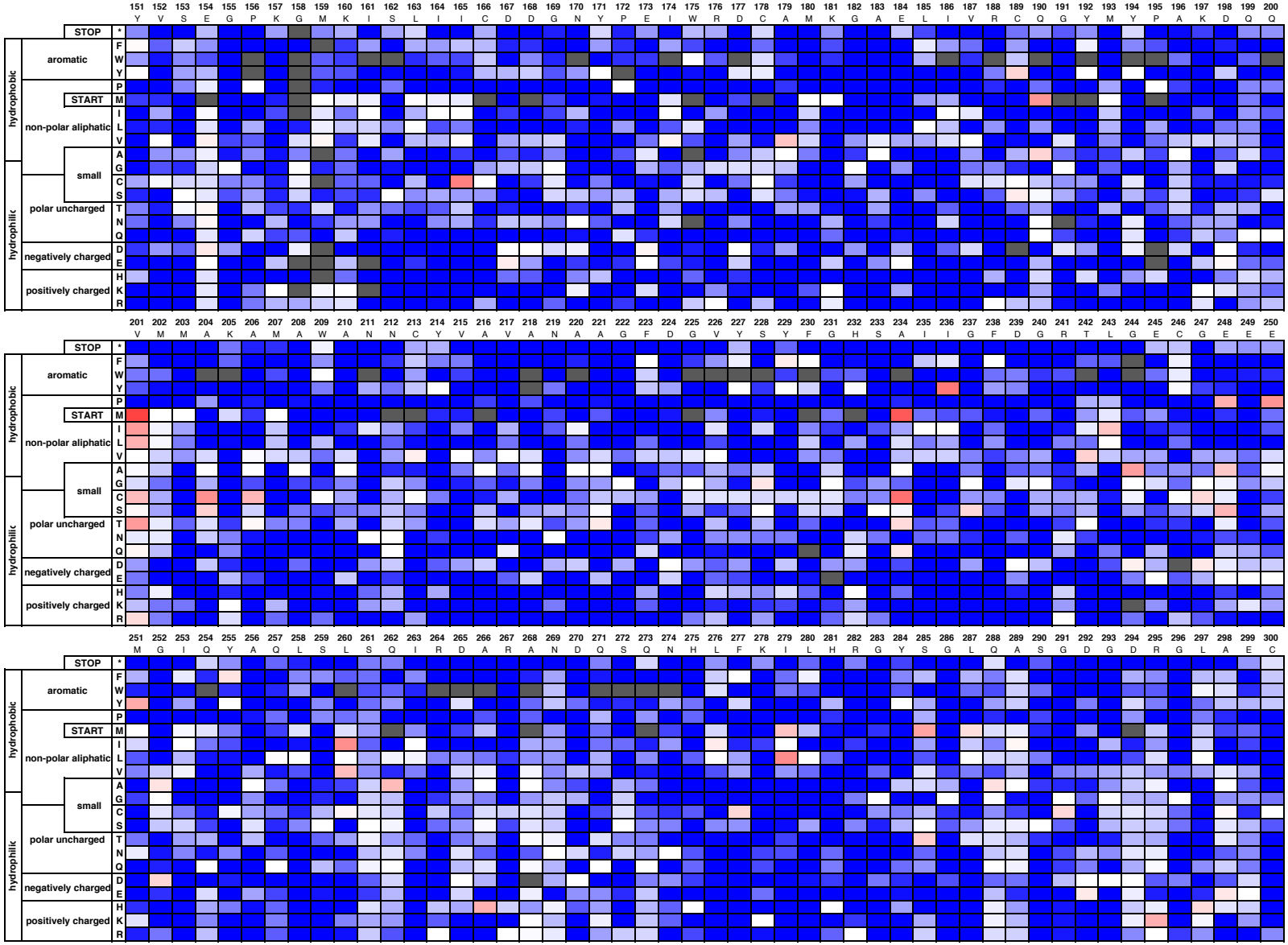
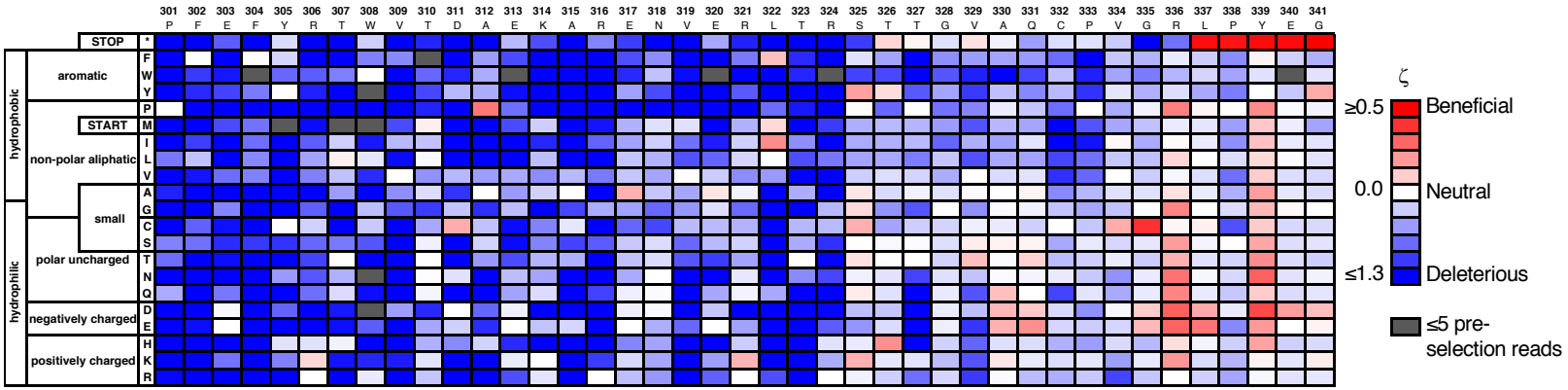


Figure B 3.3 (cont'd)



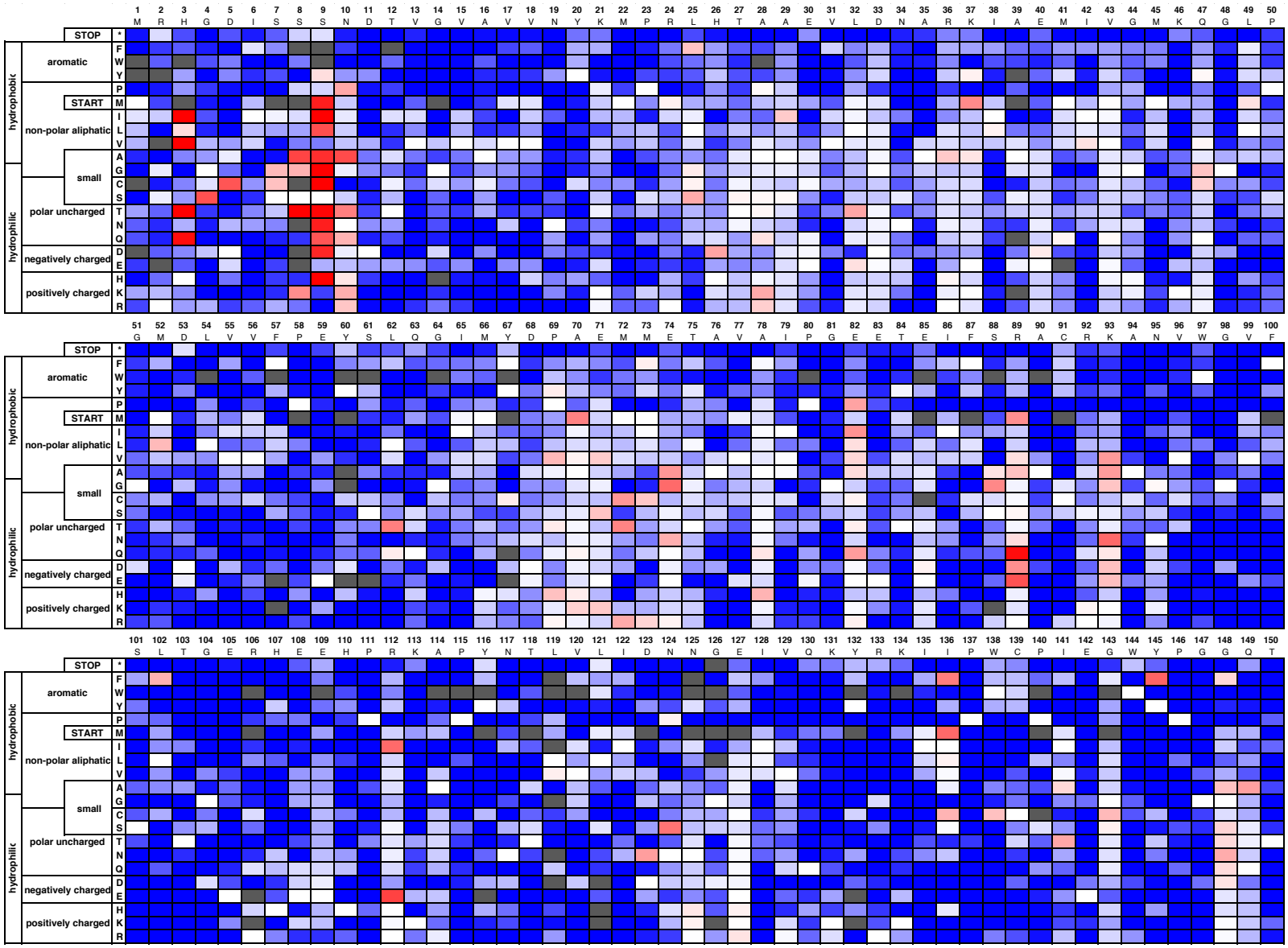


Figure B 3.4: Fitness landscape for propionamide selection.

Figure B 3.4 (cont'd)

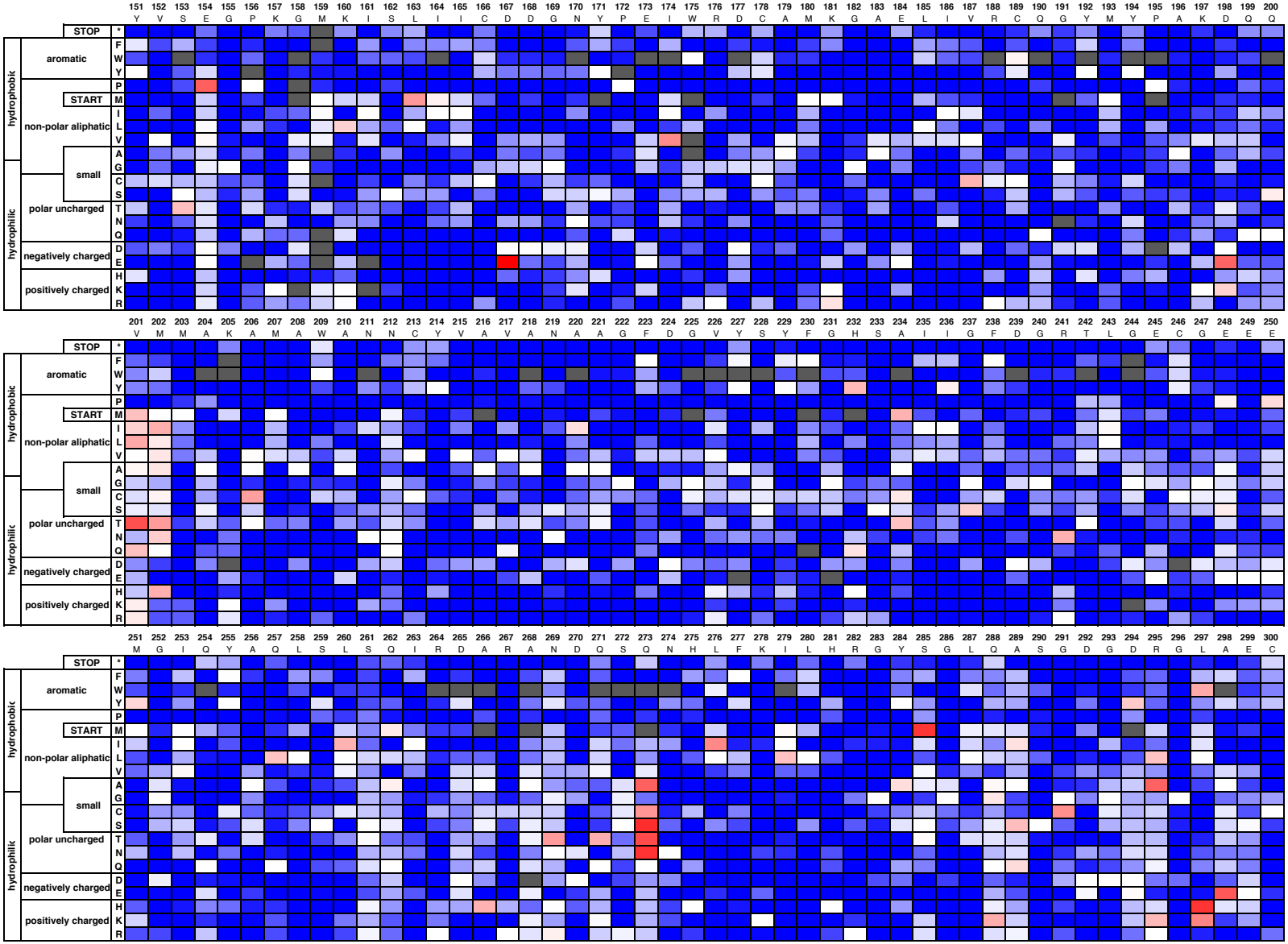
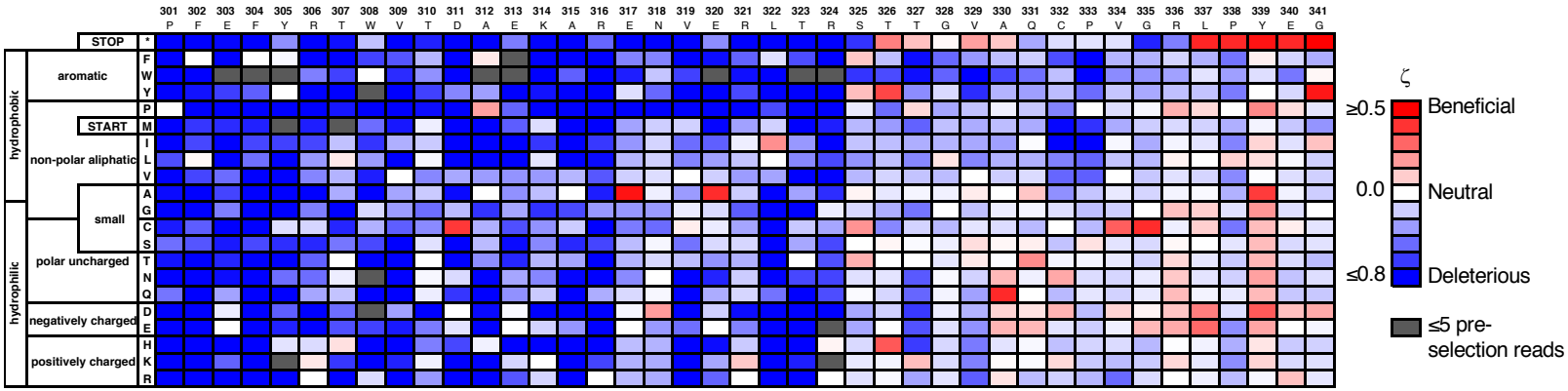


Figure B 3.4 (cont'd)



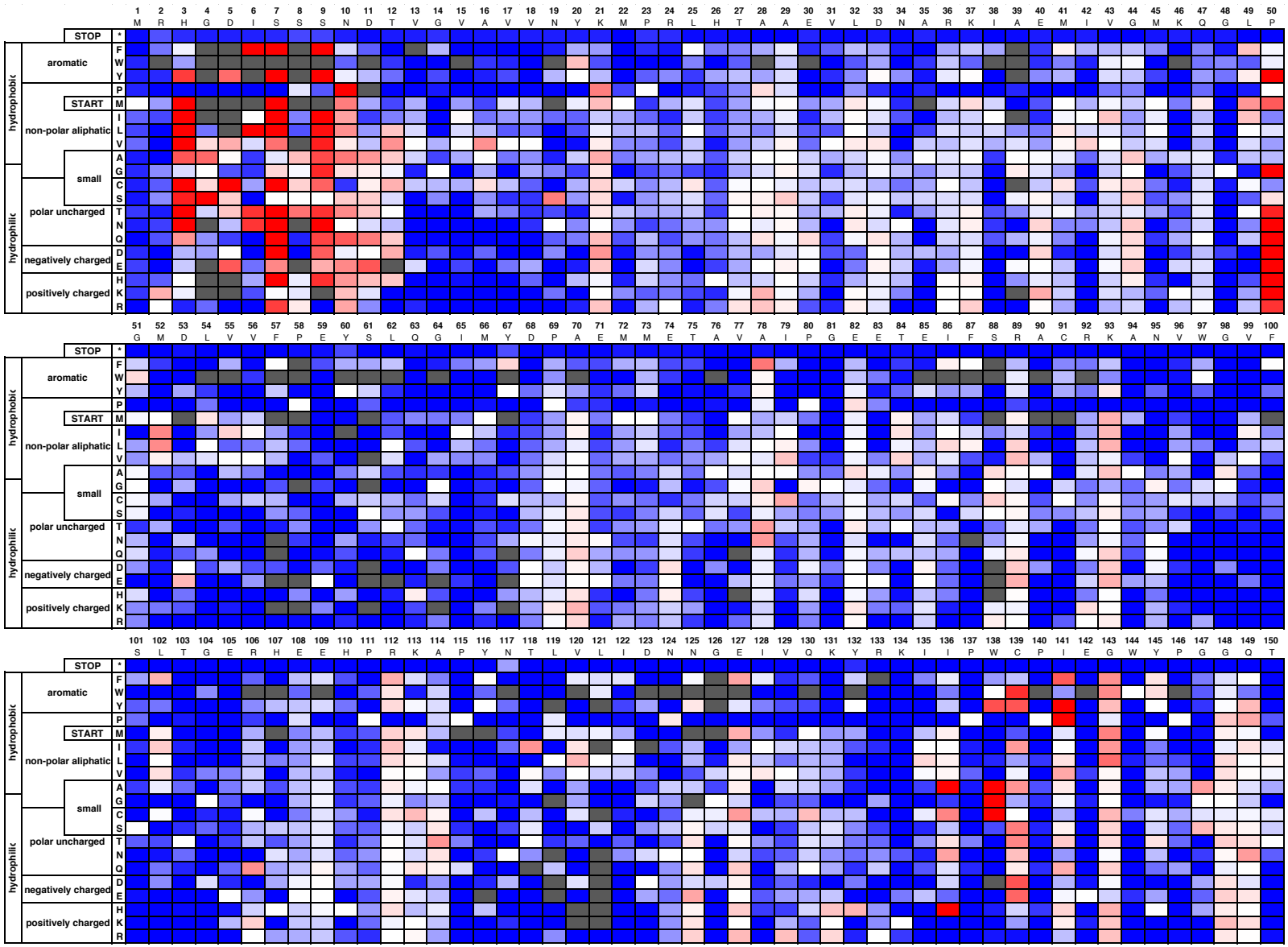


Figure B 3.5: Fitness landscape for isobutyramide selection.

Figure B 3.5 (cont'd)

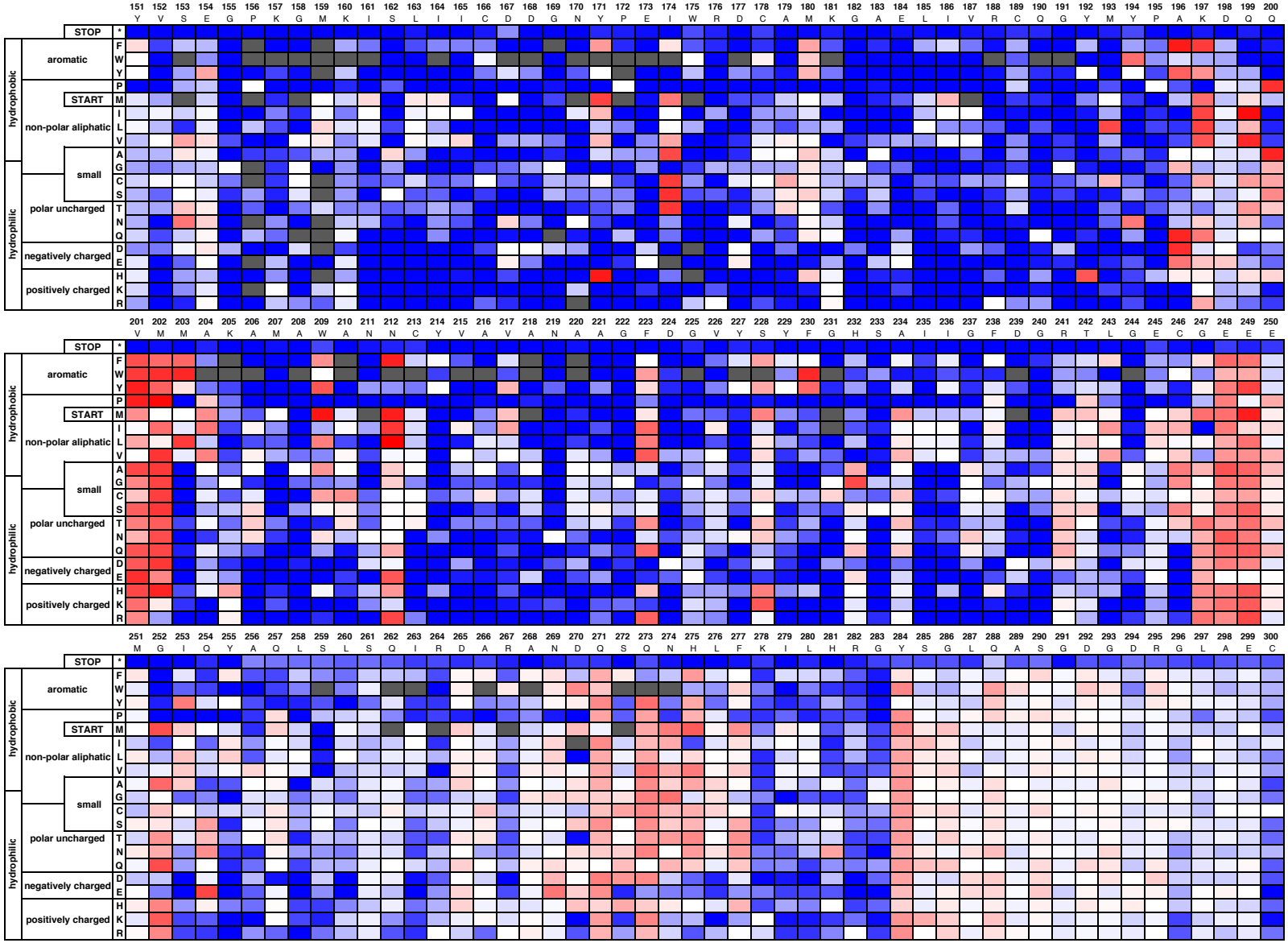
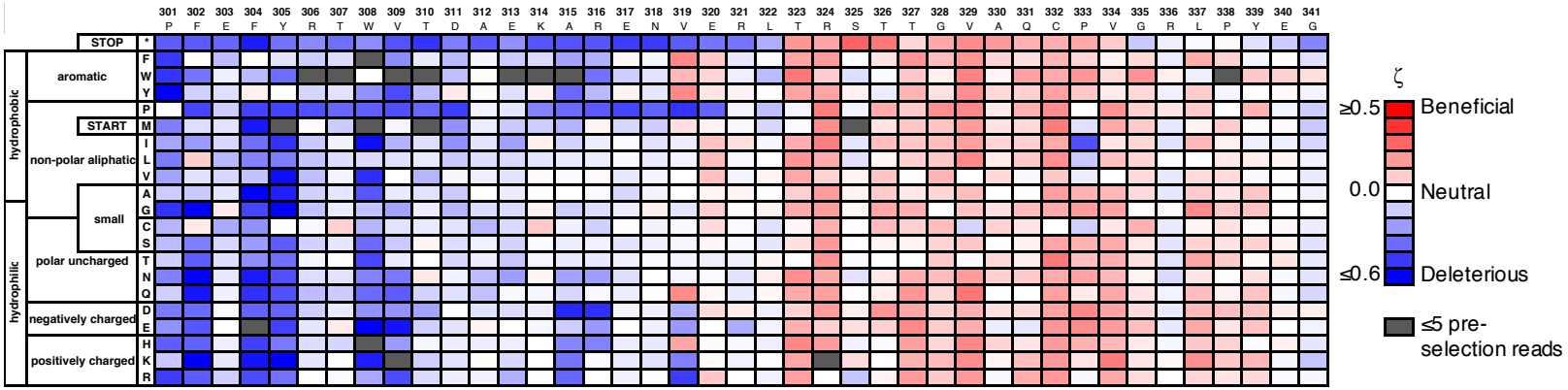


Figure B 3.5 (cont'd)



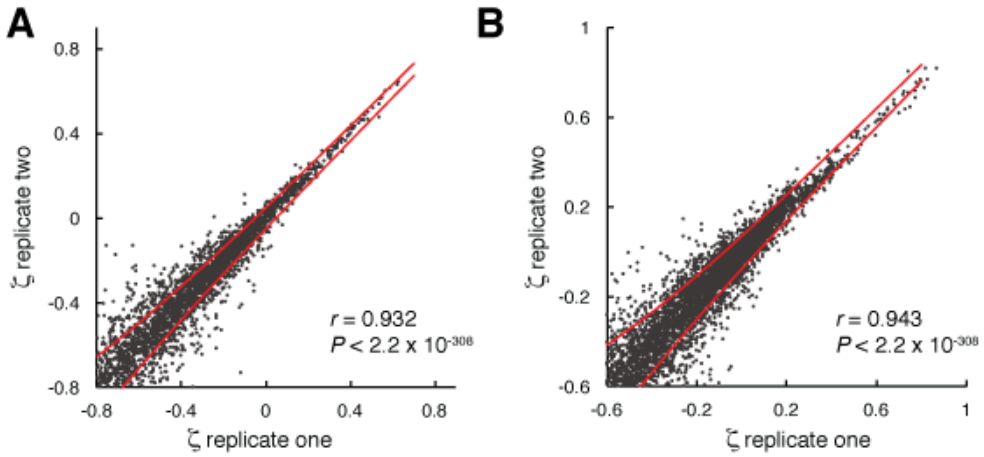


Figure B 3.6: Fitness metrics from biological replicate growth selection experiments. Panels represent replicates in A.) acetamide and B.) isobutyramide media. Plots represent n = 3834 and 4977 variants for panels A and B, respectively. Red lines indicate two standard deviations from theoretical error estimation[34]. P-values for Pearson's product moment correlation coefficients were calculated using a two-tailed t-test.

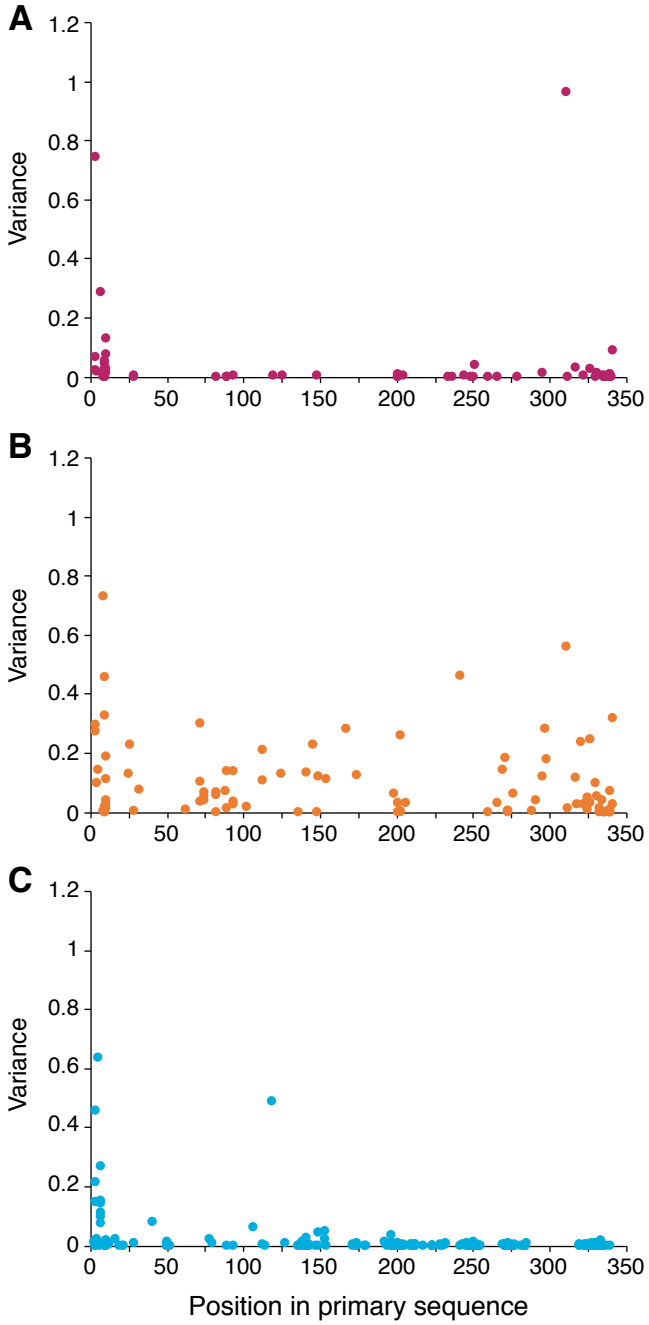


Figure B 3.7: Variance of fitness metrics for synonymous codons of beneficial mutations (ζ>0.15). Panel represent fitness metrics for the A.) acetamide B.) propionamide and C.) isobutyramide selections as a function of position in the primary sequence.

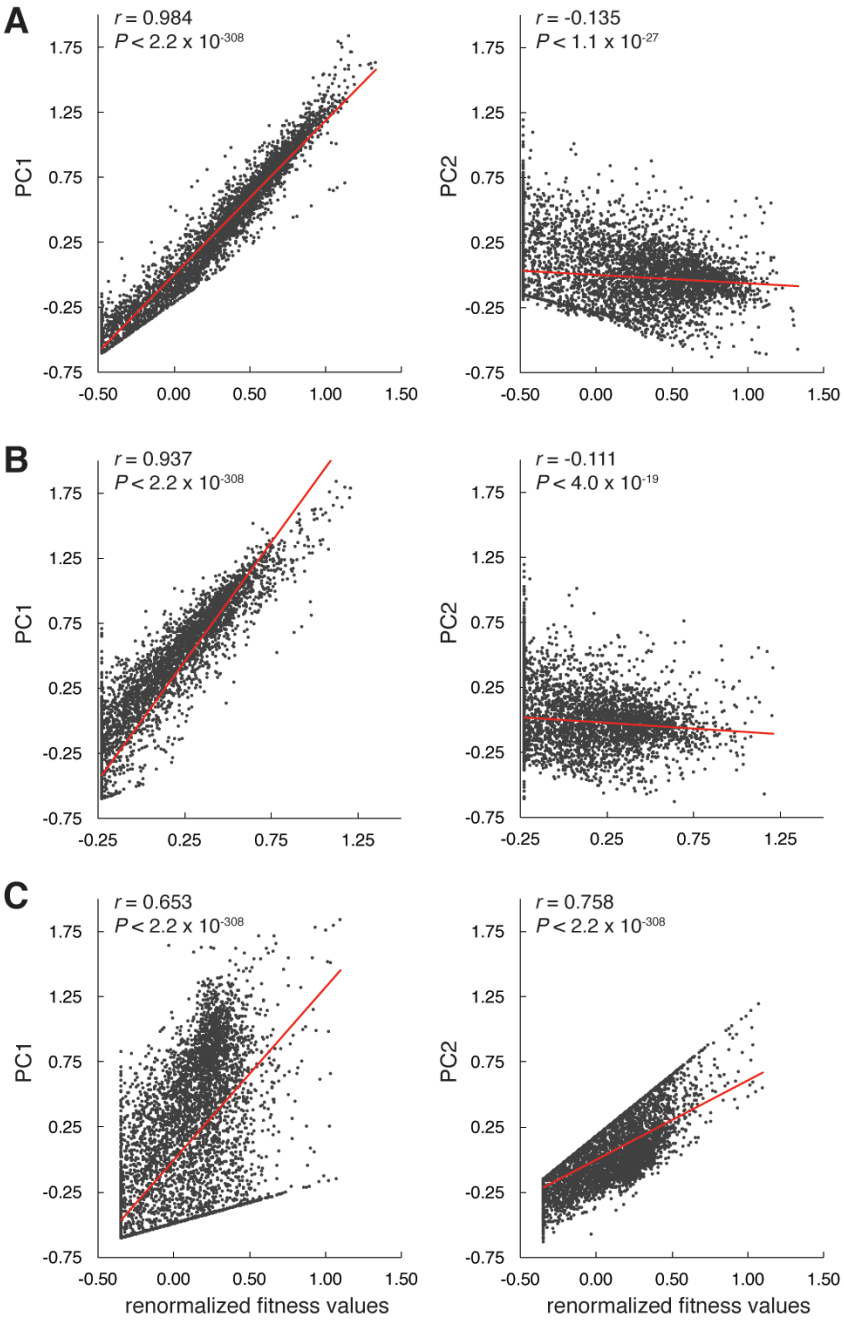


Figure B 3.8: Principle component analysis of renormalized fitness values. Panels are for A.) acetamide B.) propionamide and C.) isobutyramide selections. Fitness values were renormalized by subtracting the mean fitness (mean = -0.824, -0.575, -0.255 for acetamide, propionamide, and isobutyramide, respectively) from each variant. P-values for Pearson's product moment correlation coefficients were calculated using a two-tailed t-test.

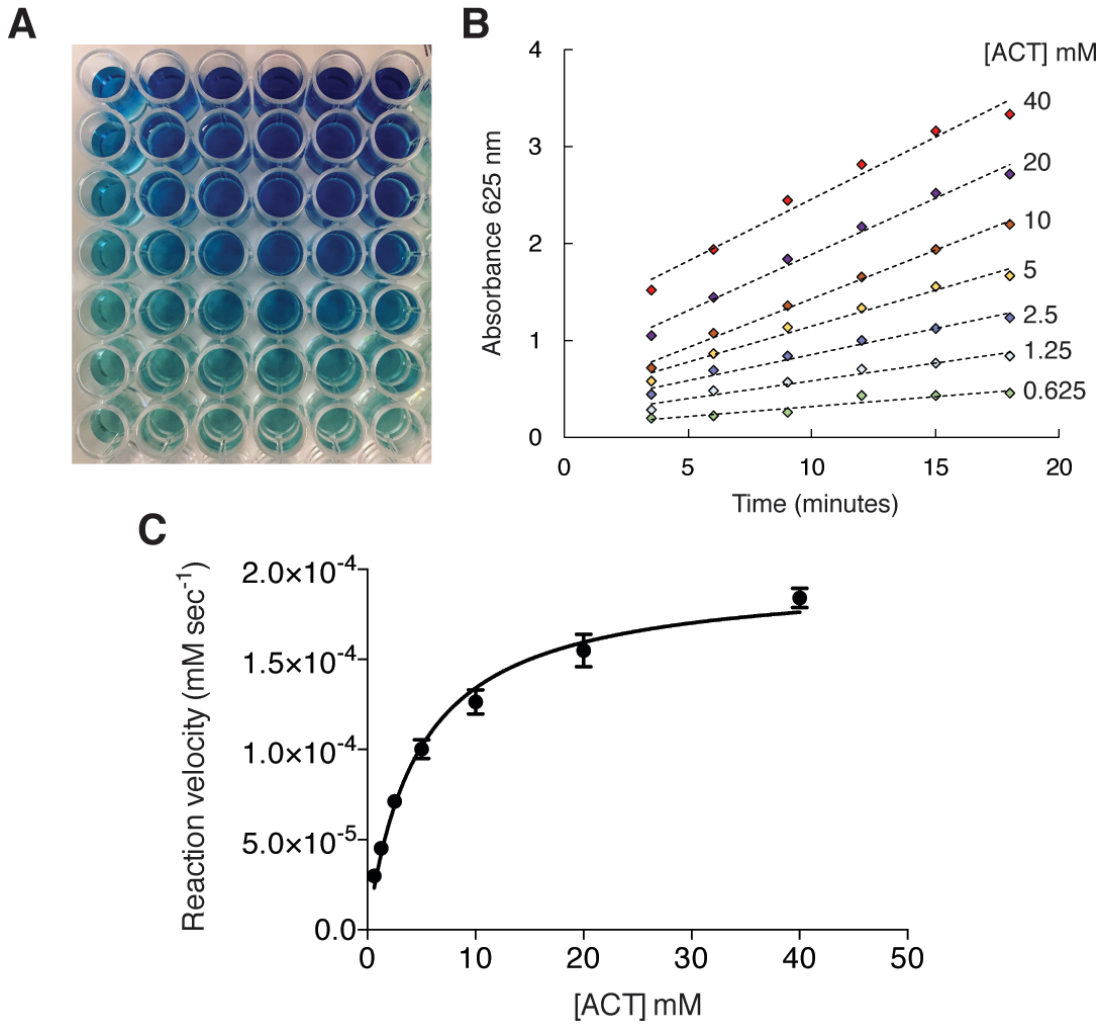


Figure B 3.9: amiE activity assay. A.) amiE activities were measured using a colorimetric Berthelot reaction[33,63] for ammonia detection with phenol nitroprusside and alkaline hypochlorite. B.) Representative data for the amiE activity assay. Absorbance at 625 nm was measured at discrete time intervals for reactions containing one of seven concentrations of acetamide (ACT) and purified wild-type amiE. Reaction velocities were calculated by obtaining the slopes of each line. C.) Michaelis-Menten plot of wild-type amiE activity on acetamide substrate. Plot represents four independent measurements. Non-linear regression was performed using GraphPad Prism version 6 for Mac OS X, GraphPad Software, La Jolla California USA, www.graphpad.com.

Table B 3.1: Constructs used in growth selections. ACT, PR, and IB = acetamide, propionamide, and isobutyramide selection media (see Methods). Promoters obtained from Bienick et al.[33].

plasmid	selection media	promoter	-35 hexamer	-10 hexamer	RBS name	RBS sequence	$\mu_{S,wt}$ (hr^{-1}) $\mu_{M9,wt}$ (hr^{-1}) $\mu_{S,wt}/\mu_{M9,wt}$
pJK_proK17_amiE	**ACT**	proK17	TTCCCG	TAATAT	t7RBS	AGGAGA	0.60 ± 0.02 0.65 ± 0.03 **0.92 ± 0.05**
pEDA2_amiE	**ACT**	proK17	TTCCCG	TAATAT	kRBS3	AGTTTT	0.44 ± 0.04 0.78 ± 0.03 **0.56 ± 0.06**
pEDA2_amiE	**PR**	proK17	TTCCCG	TAATAT	kRBS3	AGTTTT	0.29 ± 0.06 0.78 ± 0.03 **0.37 ± 0.08**
pEDA6_amiE	**IB**	proK14	TGTACG	TAATAT	t7RBS	AGGAGA	0.36 ± 0.07 0.66 ± 0.04 **0.54 ± 0.11**

Table B 3.2: Library coverage statistics for combined amiE libraries (replicate 1 and 2) used in the acetamide, propionamide, and isobutyramide selections. Raw sequencing reads were quality filtered using Enrich[55].

	Acetamide selection	Propionamide selection	Isobutyramide selection
Pre-selection population DNA reads post quality filter	1,935,216	1,735,919	1,390,991
Post-selection population DNA reads post quality filter	7,738,122	3,236,744	2,831,599
Percent of possible codon substitutions observed:			
1-base substitution	100.0	100.0	100.0
2-base substitution	97.6	97.6	98.5
3-base substitution	95.7	95.6	97.2
All substitutions	97.1	97.1	98.1
Percent of reads in pre-selection library with:			
No nonsynonymous mutations	40.0	39.6	38.3
One nonsynonymous mutation	52.0	51.8	52.6
Multiple nonsynonymous mutations	8.0	8.7	9.1
Coverage of possible single nonsynonymous mutations:	**97.2**	**97.2**	**96.3**

Table B 3.3: Isogenic growth and lysate flux data. Confidence intervals in error are given as 1 s.d. of at least 3 independent measurements.

variant	fitness metric (selection)	μ_i (hr^{-1})	μ_i/μ_{WT}	theoretical μ_i/μ_{WT}	lysate flux J_i/J_{WT} (mmol NH_3 $gDCW^{-1}$ hr^{-1})
wildtype	0.00 (ACT)	0.44 ± 0.04	1.00	1.00	0.15 ± 0.02
wildtype	0.00 (PR)	0.29 ± 0.06	1.00	1.00	*
wildtype	0.00 (IB)	0.36 ± 0.07	1.00	1.00	0.11 ± 0.01
S9A	0.33 (ACT)	0.68 ± 0.04	1.40 ± 0.08	1.26	nd
A28R	0.27 (ACT)	0.59 ± 0.03	1.44 ± 0.07	1.21	nd
L119A	0.30 (ACT)	0.66 ± 0.01	1.35 ± 0.03	1.23	1.98 ± 0.66
I136A	-0.07 (PR)	0.26 ± 0.00	0.78 ± 0.01	0.95	nd
I136A	0.52 (IB)	0.58 ± 0.00	1.31 ± 0.01	1.43	nd
I136H	-0.10 (PR)	0.33 ± 0.01	0.96 ± 0.02	0.94	nd
Q149A	0.19 (PR)	0.25 ± 0.00	0.88 ± 0.01	1.14	nd
I165C	0.25 (ACT)	0.62 ± 0.03	1.51 ± 0.08	1.19	nd
Y192V	-1.3 (ACT)	ng	nd		nd
V201M	0.12 (PR)	0.37 ± 0.01	1.09 ± 0.03	1.09	nd
V201M	0.22 (IB)	0.50 ± 0.00	1.29 ± 0.07	1.17	nd
V201T	0.20 (ACT)	0.52 ± 0.01	1.07 ± 0.02	1.15	2.08 ± 0.66
V201T	0.34 (PR)	0.39 ± 0.00	1.17 ± 0.01	1.27	nd
V201T	0.25 (IB)	0.56 ± 0.01	1.25 ± 0.02	1.19	1.9 ± 0.23
M203W	0.43 (IB)	0.61 ± 0.01	1.86 ± 0.07	1.34	0.69 ± 0.07
A234M	0.33 (ACT)	0.63 ± 0.01	1.29 ± 0.03	1.25	nd
A234M	0.15 (PR)	0.42 ± 0.01	1.25 ± 0.02	1.11	nd
A234M	0.21 (IB)	0.50 ± 0.01	1.14 ± 0.02	1.15	nd
I236Y	0.26 (ACT)	0.60 ± 0.00	1.22 ± 0.02	1.20	nd
Q273A	0.23 (IB)	0.59 ± 0.01	1.55 ± 0.09	1.17	nd

*error in measurements was prohibitively high for calculating ratios

nd = not determined

ng = no growth

Table B 3.4: mRNA effects on fitness. Pearson correlation analysis of mRNA model parameters calculated as in[47] with codon fitness metrics obtained in this work. Analysis was restricted to variants with ≥50 pre-selection read counts.

			All codons		Codons 2-16	
		Term	*r*	p-value	*r*	p-value
ACT		# of variants	6760		196	
	mRNA folding parameters	ΔG_{UH}	-0.022	0.073	-0.111	0.122
		a_H	0.008	0.492	0.069	0.338
		g_H	-0.003	0.814	-0.003	0.966
		u_{3H}	0.012	0.325	0.163	0.022
		$\pi(\theta_{wt})$	**-0.022**	**0.069**	**0.219**	**0.002**
	Codon influence parameters	$\Sigma\beta_c f_c$	-0.033	0.007	0.301	0.000
		$s_{7\text{-}16}$	0.031	0.010	0.247	0.000
		$s_{17\text{-}32}$	-0.020	0.092	-	-
		r	0.023	0.061	0.084	0.240
PR		# of variants	6193		161	
	mRNA folding parameters	ΔG_{UH}	-0.0293	0.021	-0.1421	0.071
		a_H	0.0036	0.777	0.0625	0.429
		g_H	-0.0131	0.303	-0.0641	0.418
		u_{3H}	-0.0061	0.630	0.1002	0.205
		$\pi(\theta_{wt})$	**-0.0131**	**0.303**	**0.2256**	**0.004**
	Codon influence parameters	$\Sigma\beta_c f_c$	-0.0087	0.496	0.2795	0.000
		$s_{7\text{-}16}$	0.0245	0.054	0.3097	0.000
		$s_{17\text{-}32}$	-0.0147	0.248	-	-
		r	0.0131	0.304	0.1038	0.189
IB		# of variants	2975		43	
	mRNA folding parameters	ΔG_{UH}	0.015	0.407	-0.025	0.874
		a_H	0.097	0.000	0.275	0.071
		g_H	-0.066	0.000	-0.161	0.297
		u_{3H}	0.088	0.000	0.162	0.295
		$\pi(\theta_{wt})$	**0.044**	**0.017**	**0.064**	**0.678**
	Codon influence parameters	$\Sigma\beta_c f_c$	-0.048	0.008	-0.224	0.143
		$s_{7\text{-}16}$	0.028	0.125	-0.303	0.045
		$s_{17\text{-}32}$	-0.016	0.392	-	-
		r	-0.015	0.402	-0.012	0.936

Table B 3.5: Gene amplification primers for preparing samples for deep sequencing.

Gene amplification: inner primers	
Fwd_Tile1_amiE	gttcagagttctacagtccgacgatcttaactttaagaagtttttatacat
Fwd_Tile1_amiE-2	gttcagagttctacagtccgacgatcttaactttaagaaggagatatacat
Fwd_Tile2_amiE	gttcagagttctacagtccgacgatcggcgaagaaacggaa
Fwd_Tile3_amiE	gttcagagttctacagtccgacgatcctgcgatgacggtaat
Fwd_Tile4_amiE	gttcagagttctacagtccgacgatcaagaaatgggcattcaatac
Rev_Tile1_amiE	ccttggcacccgagaattccaaagcacggctaaagat
Rev_Tile2_amiE	ccttggcacccgagaattccactctccaaatttccggata
Rev_Tile3_amiE	ccttggcacccgagaattccacagagacaactgcgc
Rev_Tile4_amiE	ccttggcacccgagaattccatggtggtgctcgag
blue = Illumina sequencing primer; black = gene overlap	
Gene amplification: outer primers	
Illumina_FWD	aatgatacggcgaccaccgagatctacacgttcagagttctacagtccga
Primer (selection, sample)	
RPI41 (ACT, T1-1)	caagcagaagacggcatacgagat**GTCGTC**gtgactggagttccttggcacccgagaattcca
RPI38 (ACT, T1-2)	caagcagaagacggcatacgagat**AGCTAG**gtgactggagttccttggcacccgagaattcca
RPI33 (ACT, T2-1)	caagcagaagacggcatacgagat**CGCCTG**gtgactggagttccttggcacccgagaattcca
RPI34 (ACT, T2-2)	caagcagaagacggcatacgagat**GCCATG**gtgactggagttccttggcacccgagaattcca
RPI43 (ACT, T3-1)	caagcagaagacggcatacgagat**GCTGTA**gtgactggagttccttggcacccgagaattcca
RPI40 (ACT, T3-2)	caagcagaagacggcatacgagat**TCTGAG**gtgactggagttccttggcacccgagaattcca
RPI44 (ACT, T4-1)	caagcagaagacggcatacgagat**ATTATA**gtgactggagttccttggcacccgagaattcca
RPI41 (ACT, T4-2)	caagcagaagacggcatacgagat**GTCGTC**gtgactggagttccttggcacccgagaattcca
RPI37 (ACT, T1U)	caagcagaagacggcatacgagat**ATTCCG**gtgactggagttccttggcacccgagaattcca
RPI22 (ACT, T2U)	caagcagaagacggcatacgagat**CGTACG**gtgactggagttccttggcacccgagaattcca
RPI39 (ACT, T3U)	caagcagaagacggcatacgagat**GTATAG**gtgactggagttccttggcacccgagaattcca
RPI40 (ACT, T4U)	caagcagaagacggcatacgagat**TCTGAG**gtgactggagttccttggcacccgagaattcca
RPI25 (PR, T1-1)	caagcagaagacggcatacgagat**ATCAGT**gtgactggagttccttggcacccgagaattcca
RPI26 (PR, T1-2)	caagcagaagacggcatacgagat**GCTCAT**gtgactggagttccttggcacccgagaattcca
RPI27 (PR, T2-1)	caagcagaagacggcatacgagat**AGGAAT**gtgactggagttccttggcacccgagaattcca
RPI28 (PR, T2-2)	caagcagaagacggcatacgagat**CTTTTG**gtgactggagttccttggcacccgagaattcca
RPI29 (PR, T3-1)	caagcagaagacggcatacgagat**TAGTTG**gtgactggagttccttggcacccgagaattcca
RPI30 (PR, T3-2)	caagcagaagacggcatacgagat**CCGGTG**gtgactggagttccttggcacccgagaattcca
RPI31 (PR, T4-1)	caagcagaagacggcatacgagat**ATCGTG**gtgactggagttccttggcacccgagaattcca
RPI32 (PR, T4-2)	caagcagaagacggcatacgagat**TGAGTG**gtgactggagttccttggcacccgagaattcca
RPI21 (PR, T1U)	caagcagaagacggcatacgagat**CGAAAC**gtgactggagttccttggcacccgagaattcca

Table B 3.5 (cont'd)

RPI22 (PR, T2U)	caagcagaagacggcatacgagat**CGTACG**gtgactggagttccttggcacccgagaattcca
RPI23 (PR, T3U)	caagcagaagacggcatacgagat**CCACTC**gtgactggagttccttggcacccgagaattcca
RPI24 (PR, T4U)	caagcagaagacggcatacgagat**GCTACC**gtgactggagttccttggcacccgagaattcca
RPI13 (IB, T1-1)	caagcagaagacggcatacgagat**TTGACT**gtgactggagttccttggcacccgagaattcca
RPI14 (IB, T1-2)	caagcagaagacggcatacgagat**GGAACT**gtgactggagttccttggcacccgagaattcca
RPI15 (IB, T2-1)	caagcagaagacggcatacgagat**TGACAT**gtgactggagttccttggcacccgagaattcca
RPI16 (IB, T2-2)	caagcagaagacggcatacgagat**GGACGG**gtgactggagttccttggcacccgagaattcca
RPI17 (IB, T3-1)	caagcagaagacggcatacgagat**CTCTAC**gtgactggagttccttggcacccgagaattcca
RPI18 (IB, T3-2)	caagcagaagacggcatacgagat**GCGGAC**gtgactggagttccttggcacccgagaattcca
RPI19 (IB, T4-1)	caagcagaagacggcatacgagat**TTTCAC**gtgactggagttccttggcacccgagaattcca
RPI20 (IB, T4-2)	caagcagaagacggcatacgagat**GGCCAC**gtgactggagttccttggcacccgagaattcca
RPI9 (IB, T1U)	caagcagaagacggcatacgagat**CTGATC**gtgactggagttccttggcacccgagaattcca
RPI10 (IB, T2U)	caagcagaagacggcatacgagat**AAGCTA**gtgactggagttccttggcacccgagaattcca
RPI11 (IB, T3U)	caagcagaagacggcatacgagat**GTAGCC**gtgactggagttccttggcacccgagaattcca
RPI12 (IB, T4U)	caagcagaagacggcatacgagat**TACAAG**gtgactggagttccttggcacccgagaattcca
red = Illumina adapter sequence; **BOLD** = barcode; blue = Illumina sequencing primer	

REFERENCES

REFERENCES

1. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485,** 185–194 (2012).

2. Liberles, D. A. *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **21,** 769–785 (2012).

3. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79,** 471–505 (2010).

4. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **106 Suppl,** 9995–10000 (2009).

5. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6,** 119–127 (2005).

6. Orr, H. A. The Population Genetics of Adaptation: The Distribution of Factors Fixed During Adaptive Evolution. *Evolution (N. Y).* **52,** 935–949 (1998).

7. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev.* **8,** (2007).

8. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** gku511 (2014).

9. Arnold, F. H., Wintrode, P. L., Miyazaki, K. & Gershenson, A. How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **26,** 100–106 (2001).

10. Currin, A., Swainston, N., Day, P. J. & Kell, D. B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* **44,** 1172–1239 (2015).

11. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2,** 17 (2007).

12. Dalby, P. A. Strategy and success for the directed evolution of enzymes. *Curr. Opin. Struct. Biol.* **21,** 473–480 (2011).

13. Morley, K. L. & Kazlauskas, R. J. Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.* **23,** 231–237 (2005).

14. Schmidt, M. *et al.* Directed Evolution of an Esterase from Pseudomonas fluorescens Yields a Mutant with Excellent Enantioselectivity and Activity for the Kinetic Resolution of a Chiral Building Block. *ChemBioChem* **7,** 805–809 (2006).

15. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37,** 73–76 (2005).

16. Lee, J. & Goodey, N. M. Catalytic Contributions from Remote Regions of Enzyme Structure. *Chem. Rev.* 7595–7624 (2011).

17. Omari, K. El, Liekens, S., Bird, L. E., Balzarini, J. & Stammers, D. K. Mutations Distal to the Substrate Site Can Affect Varicella Zoster Virus Thymidine Kinase Activity: Implications for Drug Design. *Mol. Pharmacol.* **69,** 1891–1896 (2006).

18. Tomatis, P. E., Rasia, R. M., Segovia, L., Vila, A. J. & Gray, H. B. Mimicking Natural Evolution in Metallo-beta-Lactamases through Second-Shell Ligand Mutations. *Proc. Natl. Acad. Sci.* **102,** 13761–13766 (2005).

19. Yang, G., Hong, N., Baier, F., Jackson, C. J. & Tokuriki, N. Conformational Tinkering Drives Evolution of a Promiscuous Activity through Indirect Mutational Effects. *Biochemistry* **55,** 4583–4593 (2016).

20. Oue, S., Okamoto, A., Yano, T. & Kagamiyama, H. Redesigning the Substrate Specificity of an Enzyme by Cumulative Effects of the Mutations of Non-active Site Residues. *J. Biol. Chem.* **274,** 2344–2349 (1999).

21. Mavor, D. *et al.* Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife* **5,** e15802 (2016).

22. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *J. Mol. Biol.* **425,** 1363–1377 (2013).

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

24. Gillespie, J. H. Molecular Evolution Over the Mutational Landscape. *Evolution (N. Y).* **38,** 1116–1129 (1984).

25. Orr, H. A. The population genetics of beneficial mutations. *Philos. Trans. R. Soc. London B Biol. Sci.* **365,** 1195–1201 (2010).

26. Kelly, M. & Clarke, P. H. An Inducible Amidase Produced by a Strain of Pseudomonas aeruginosa. *Microbiology* **27,** 305–316 (1962).

27. Andrade, J., Karmali, A., Carrondo, M. A. & Frazao, C. Structure of Amidase from Pseudomonas aeruginosa Showing a Trapped Acyl Transfer Reaction Intermediate State. *J. Biol. Chem.* **282,** 19598–19605 (2007).

28. Schmid, A. *et al.* Industrial biocatalysis today and tomorrow. *Nature* **409,** 258–268 (2001).

29. Buchholz, K., Kasche, V. & Bornscheuer, U. T. *Biocatalysts and Enzyme Technology*. (John Wiley & Sons, 2012).

30. Kim, M. *et al.* Need-based activation of ammonium uptake in Escherichia coli. *Mol. Syst. Biol.* **8,** 616 (2012).

31. Reitzer, L. Nitrogen assimilation and global regulation in Escherichia coli. *Annu. Rev. Microbiol.* **57,** 155–176 (2003).

32. Davis, J. H., Rubin, A. J. & Sauer, R. T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* **39,** 1131–1141 (2011).

33. Bienick, M. S. *et al.* The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate. *PLoS One* **9,** e109105 (2014).

34. Kowalsky, C. A. *et al.* High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PLoS One* **10,** e0118193 (2015).

35. Firnberg, E. & Ostermeier, M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One* **7,** e52031 (2012).

36. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat. Publ. Gr.* **11,** 572–582 (2010).

37. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

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

39. Steinberg, B. & Ostermeier, M. Environmental changes bridge evolutionary valleys. *Sci. Adv.* **2,** e1500921 (2016).

40. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nat. Genet.* **38,** 484–488 (2006).

41. Serohijos, A. W. R. & Shakhnovich, E. I. Contribution of Selection for Protein Folding Stability in Shaping the Patterns of Polymorphisms in Coding Regions. *Mol. Biol. Evol.* **31,** 165–176 (2014).

42. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–932 (2006).

43. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci.* **110,** 13067–13072 (2013).

44. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics* **170,** 989–999 (2005).

45. Rokyta, D. R. *et al.* Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* **67,** 368–376 (2008).

46. Sanjuán, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8396–401 (2004).

47. Boël, G. *et al.* Codon influence on protein expression in E . coli correlates with mRNA levels. *Nature* **529,** 358–363 (2016).

48. van der Meer, J. Y. *et al.* Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective Michaelases. *Nat. Commun.* **7,** 1–16 (2016).

49. Chevereau, G. *et al.* Quantifying the Determinants of Evolutionary Dynamics Leading to Drug Resistance. *PLos Biol.* **13,** e1002299 (2015).

50. Klesmith, J. R., Bacik, J., Michalczyk, R. & Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth. Biol.* **4,** 1235–1243 (2015).

51. Sniegowski, P. D. & Gerrish, P. J. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philos. Trans. R. Soc. London B Biol. Sci.* **365,** 1255–1263 (2010).

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

53. Gong, L. I. & Bloom, J. D. Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS One* **10,** (2014).

54. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci.* **82,** 488–492 (1985).

55. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27,** 3430–3431 (2011).

56. R Core Team, R. R: A language and environment for statistical computing. (2015). at <https://www.r-project.org/>

57. Villaseñor-Alva, J. A. & González-Estrada, E. A bootstrap goodness of fit test for the generalized Pareto distribution. *Comput. Stat. Data Anal.* **53,** 3835–3841 (2009).

58. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **64,** 1–34 (2015).

59. Scholz, F. & Zhu, A. K-Sample Rank Tests and their Combinations. (2016). at <http://cran.r-project.org/package=kSamples>

60. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

61. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357,** 289–298 (2006).

62. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. High-throughput thermal scanning: a general, rapid dye-binding thermal shift screen for protein engineering. *J. Am. Chem. Soc.* **131,** 3794–3795 (2009).

63. Searle, P. L. The Berthelot or indophenol reaction and its use in the analytical chemistry of nitrogen. A review. *Analyst* **109,** 549–568 (1984).

CHAPTER FOUR

Computational design of destabilized proteins for assessing theories on adaptive molecular evolution

ABSTRACT

Understanding and predicting the mechanisms of adaptive evolution is a key challenge for theoretical and experimental biologists. Of interest to protein engineers is knowledge of how beneficial mutations arise in a population. More specifically, how many beneficial mutations are available given a starting amino acid sequence, and what is the nature of their distribution of fitness effects (DFE)? Among other biophysical properties, does the stability of a given protein (thermodynamic, colloidal, etc.) have an impact on this distribution? While theoretical frameworks have been developed, generating empirical data to rigorously test these theories has been a challenge. Deep sequencing mutational studies provide data on thousands of mutations in a single experiment and have proven useful in testing adaptive molecular evolution theories. In this chapter, a computational methodology was developed and applied to design functional yet destabilized proteins for use in future work to test hypotheses on the DFE for beneficial mutations.

INTRODUCTION

Understanding the mechanisms of molecular evolution is important to a broad range of scientists including molecular biologists, virologists, evolutionary biologists, and protein engineers. Researchers interested in evolving natural proteins or designing proteins *de novo* must wrestle with the implicit evolutionary limitations set forth by nature. The challenge, then, is to first define these mechanisms. Of interest to most protein engineers are beneficial mutations or 'hits' – variants that achieve some engineering goal. Given the amino acid sequence of an enzyme, how many 'hits' are available? What are the mechanisms and constraints that govern the evolvability of a protein for new functions and, more importantly, can we leverage knowledge of these mechanisms to improve the design process?

Since Ronald Fisher's seminal work in the 1930's in developing a 'geometric' evolutionary model[1,2], several theoretical developments have followed that attempt to mathematically describe the adaptive behavior of stochastic mutations. One phenomena noted early on was that beneficial mutations are rare; deleterious and neutral mutations are far more probable than ones providing a selective advantage. Understanding the rare nature of beneficial mutations, Gillespie reasoned that extreme value theorem mathematics could be utilized to model the extreme tails of these distributions[3]. Orr later matured these theories and predicted that the DFE for beneficial mutations should be approximately exponentially distributed[4,5]. Several experimental works have attempted to test these theories, yet the results have been conflicting[5]. A significant challenge is identifying enough beneficial mutations for a given protein so that rigorous distribution model fitting can be carried out.

Deep mutational scanning experiments provide a wealth of mutational data that can be used address questions in molecular evolution[6]. In **Chapter 3**, deep mutational scanning was used to

study how an enzyme encodes substrate specificity by generating comprehensive single-mutation fitness landscapes on multiple substrates[7]. Using the set of beneficial mutations - variants with functional scores above wild-type – we were able to address the shape of the DFE for beneficial mutations with high statistical power. As was predicted by Orr, all three datasets could essentially be described as exponential. As these results are contingent upon a laboratory experiment and not 'true' adaptive evolution, one has to consider the details of the selection. Importantly, the thermal stability of the enzyme under study was favorable to the selection conditions; the melting temperature of wild-type amiE is >65°C and the selections were performed at 37°C. Since proteins are generally only marginally stable in their native environments, an interesting question one could ask is, then, how does the shape of the DFE for beneficial mutations behave when the protein under study is not stable in the selection conditions?

In growth selections, fitness of an enzyme is determined by its catalytic activity and the amount of folded, functional protein. The probability that a protein will fold to a native 'lowest energy' functional confirmation is a complex computation governed by both the thermodynamic and kinetic stability constraints in the present environment[8,9]. However, in the context of a growth selection, this can be simplified and modeled as a two-state system: folded (functional) and unfolded (inactive). What is the relationship between folding probability and fitness outcomes? Previous works have aimed to address this question[10–13]. In a classic paper, Bloom et al. showed that a stabilized cytochrome P450 had a greater probability of accepting mutations conferring new functions than the wild-type enzyme[14]. The fitness effect of a mutation then, is a function of the protein's mutational 'robustness' – random mutations to a less stable protein are more likely to be deleterious. Counterintuitive to this finding and a point of frustration for protein engineers, is that stabilizing mutations often come with trade-off to activity[15–19].

The design of experiment for this project is simple: compare DFE of a destabilized protein against the known DFE of the stable wild-type protein. Using the experiment framework developed previously for the amiE enzyme as described in **Chapter 3**[7], the experimental objectives of this project are to 1.) design and generate destabilized variants of amiE while maintaining wild-type catalytic function, 2.) establish and perform growth selections on the acetamide substrate, and 3.) compare the beneficial DFE of the destabilized protein to the wild-type. The null hypothesis is that the shape of the DFE is independent of thermodynamic stability. In the following chapter I will describe a computational approach to address objective 1 using the Rosetta Design Software. We were able to identify several variants of amiE that impacted the relative purification yield (a proxy for enzyme stability). In the **Discussion and Outlook** section I will discuss the implications for use of computational design in altering protein stability and outline future work necessary to accomplish the remaining objectives of this project.

RESULTS

We first sought to generate destabilized variants of amiE using computational methods. The objective was to identify mutations that would decrease the stability of the protein while maintaining wild-type catalytic function. In effort to computationally identify destabilizing mutations, we modified the Rosetta protocols used in the PROSS pipeline developed by Goldenzweig et al.[20]. Briefly, the objective of PROSS is to suggest designs given an input protein sequence and structure that will have improved stability and heterologous expression yields from bacterial hosts such as *E coli*. As our experimental objective is essentially the inverse problem, we modified the protocols used in PROSS and then selected mutations with worse energy scores relative to wild-type. The computational pipeline that was used is outlined in **Figure 4.1**. The input structure for amiE (PDB code 2UXY)[21] was first refined by iterative rounds of sidechain and backbone repacking to obtain the lowest energy structure. Next, we reduced the number of positions to test by removing any position that was 1.) within 8 Å of the active site or 2.) on the surface of the protein. The former is to reduce the likelihood of choosing a mutation that impacts catalytic function, as proximity to active site correlates with activity[15,17,22]. The removal of positions on the surface of the protein from our search space was to avoid any potential effects to protein oligomerization (amiE is a homohexamer). The FilterScan protocol[23] was then run on the remaining positions to test and score all possible single amino acid substitutions.

Figure 4.1: Process flow for computational design of destabilized proteins.

Using the output scores from the FilterScan module along with the point-mutant fitness metric data previously generated[7], six destabilized amiE designs harboring 1-3 mutations each were expressed and purified from *E coli* BL21* along with wild-type amiE (**Table C 4.1**). For 4/6 designs we observed essentially no protein yield from our purification, and the remaining two designs (ED2 and LA3) had a significant reduction in the amount of protein yielded relative to wild-type (**Figure 4.2a** and **Table C 4.1**).

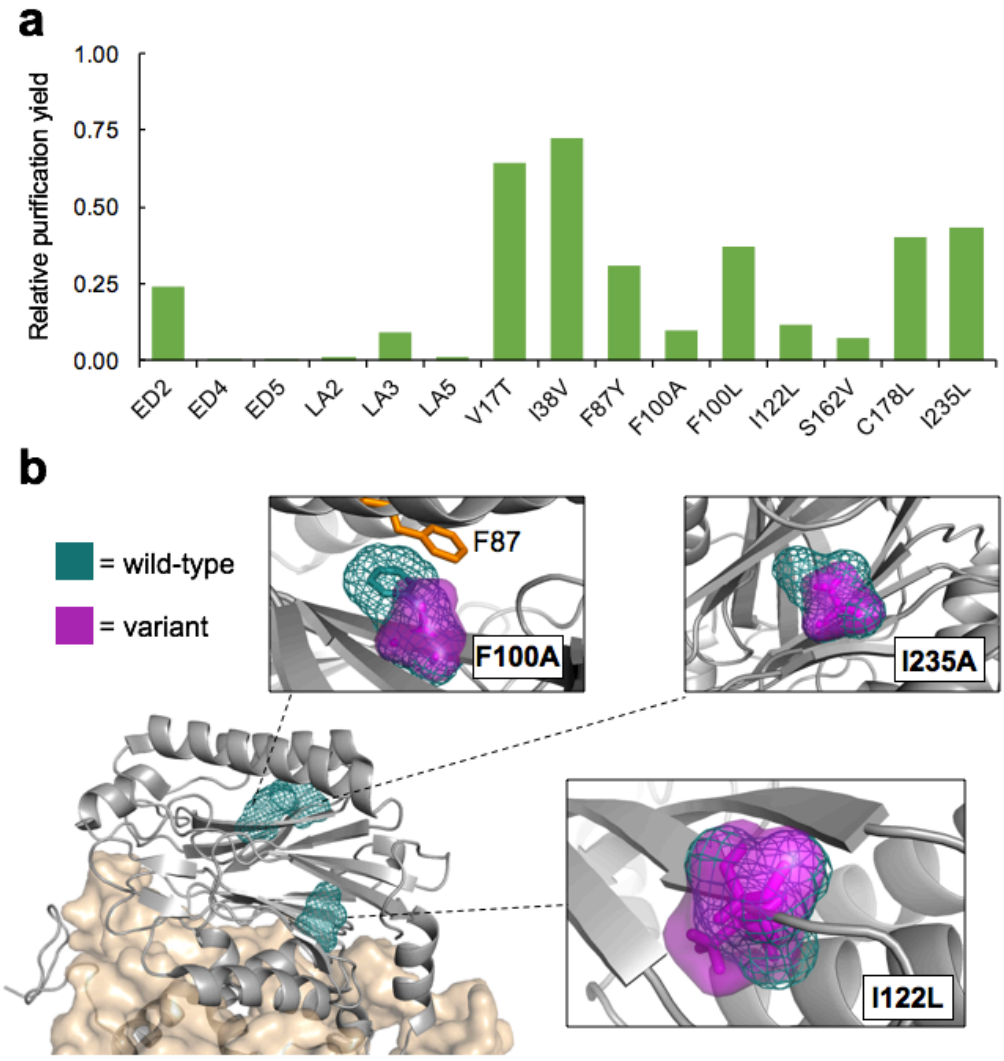


Figure 4.2: Destabilized amiE variants. a.) Wild-type and variant amiE proteins were expressed, purified, and quantified using the A_{280} method and the yield of each variant protein relative to wild-type was computed. b.) The 'large-to-small' mutations introduce voids into the core of the

(**Figure 4.2 cont'd**) protein (F100A and I235A) while other hydrophobic to hydrophobic mutations alter the local packing.

To obtain an unambiguous view of the contributions of each mutation, we made and purified some of the point-mutations contained within the six designs and found that all had decreased expression yields relative to wild-type (**Figure 4.2a** and **Table C 4.1**). The majority of mutations introduce a void into the core of the structure. For example, F100A and I235A are both 'large-to-small' mutations, with the former also eliminating a stabilizing pi-pi stacking interaction with F87 (**Figure 4.2b**). Other mutations such as I122L are less aggressive in terms of void introduction, yet change the local packing dynamics in the vicinity of the mutation (**Figure 4.2b**). Interestingly, when we measured the apparent melting temperatures of the point-mutants we found that most were close to the wild-type T_m of 67.7 ± 0.1ºC (wild-type T_m previously reported in Wrenbeck et al.[7]). However, this could be explained by the fact that the homohexameric biological assembly of amiE complicates measuring the true melting temperature of the monomer with the irreversible thermal shift assay used. Following dissociation of the homohexameric complex and further (likely immediate) unfolding of each monomer unit, one cannot reverse the process to refold the monomers and thus obtain the true T_m of monomeric amiE.

DISCUSSION AND OUTLOOK

Altering the stability of a protein while maintaining functionality is a significant challenge in protein science. Generally, the objective is to improve stability: thermodynamic, colloidal, solvent tolerance, etc. Important examples include efforts to stabilize enzymes in biocatalytic or *in vivo* process[15,22,24–27] and improved shelf-life and/or heat-tolerance of protein therapeutics[28]. Computational methods, though imperfect, are becoming increasingly better at predicting the

thermodynamic effect (ΔΔG) that a mutation will have on the folding stability of a protein. However, it has long been understood that there is an inherent tradeoff between stability and activity, especially with enzymes[15,17,24]. These effects will be discussed at length in **Chapter 5**, but in the context of this project the objective was to utilize computational predictions of free energy change upon mutation to select less stable structures.

The next objective of this project is to identify which of the destabilized variants maintain catalytic activity. A fellow graduate student in the Whitehead lab, Matthew Faber, has expressed, purified, and performed activity assays as described in Wrenbeck et al.[7] on the variants described in this chapter (unpublished data). Two point-mutants, I38V and I122L, retained near wild-type activity. Future work involves establishing growth selection for one or both of these variants, performing a growth selection on a comprehensive single-mutational library using the variant as the parental enzyme, deep sequencing the pre- and post-selection populations, and then analyzing the data as in the previous study to observe the shape of the DFE for beneficial mutations.

MATERIALS AND METHODS

Reagents

All chemicals were purchased from Sigma-Aldrich unless otherwise noted. Mutagenic oligonucleotides were ordered from Integrated DNA Technologies and are listed in **Table C 4.2**.

Plasmid construction

pET29b_amiE was constructed as previously described[7]. Briefly, the amiE coding sequence was subcloned into the pET-29b(+) backbone (Novagen). All variants were generated using Kunkel mutagenesis[29]. Mutagenic primers are listed in **Table C 4.2**.

Protein purification and characterization

amiE proteins were expressed and purified following the exact protocols in Wrenbeck et al.[7]. Briefly, proteins were expressed using the auto-induction method[30] and purified on Ni-NTA column according to Klesmith et al.[25]. Purified protein solutions were quantified by measuring absorbance at 280 nm on a BioTek Synergy H1 plate reader in 96-well UV-transparent plates using an extinction coefficient of 5.883×10^{-2} $uM^{-1}cm^{-1}$ for amiE. Apparent melting temperatures were measured using a SYPRO Orange thermal-shift assay[31] as detailed in Klesmith et al.[25] and Wrenbeck et al.[7].

Computational point-mutant scan

Rosetta scripts and command lines used in this work are listed in **Notes C 4.1** and **4.2**. The crystal structure for amiE was obtained from the Protein Data Bank (PDB code 2UXY)[21] and cleaned for use in Rosetta with the 'clean_pdb_keep_ligand.py' script as part of the Rosetta 3

release. The structure was refined using the refine.xml script (without alteration) included in Data S1 from Goldenzweig et al.[20]. Residues within 8 Å of the C3Y ligand (substrate transition state analogue crystalized with amiE structure) and those comprising the C-terminus were fixed during refinement. A list of fixed residues is included in **Note C 4.1**.

Distance to the catalytic active site was calculated by finding the minimum distance of a position's alpha carbon to any active site atom (six identical active sites in the homohexamer amiE). Residues with 8 Å or less distance to the active site were excluded from the FilterScan. Surface residues were identified with the Python script findSurfaceResidues.py (https://pymolwiki.org/index.php/FindSurfaceResidues) and were also excluded from the FilterScan. The filterscan.xml script from Goldenzweig et al.[20] was modified to exclude PSSM input. The modified script can be found in **Note C 4.2**.

APPENDIX

APPENDIX

Table C 4.1: Characterization of destabilized amiE variants.

Design	Mutation(s)	FilterScan score	T_{mapp} (°C)	Relative purification yield
ED2	F87Y, I235L	4.927	68.6 ± 1.05	0.24
ED4	V17T, I38V, I122L, L258A	8.130	nd	0.00
ED5	V56I, F100H, L258A	9.449	nd	0.01
LA2	I235A	5.801	nd	0.01
LA3	F100L, C178L	8.777	60.28 ± 0.10	0.09
LA5	F100A, S162V	11.188	nd	0.01
	V17T	2.354	65.8 ± 0.20	0.64
	I38V	2.023	67.3 ± 0.24	0.72
	F87Y	1.645	73.1 ± 4.33	0.31
	F100A	7.469	67.4 ± 0.31	0.10
	F100L	4.583	79.8 ± 4.30	0.37
	I122L	0.395	nd	0.11
	S162V	3.719	nd	0.07
	C178L	4.195	64.4 ± 0.15	0.40
	I235L	3.281	79.2 ± 2.30	0.43

Table C 4.2: Primers for generating destabilized amiE variants.

Primer	Sequence
AmiE_C178L	gcacccttcatggctaaatctctccaaatttccggataattac
AmiE_F100A	gttcgccggtcagggaggccacaccccaaacatttg
AmiE_F100H	gttcgccggtcagggagtgcacaccccaaacatttg
AmiE_F100L	cgccggtcagggataacacaccccaaacatt
AmiE_F87Y	tacaagcacggctatagatttccgtttcttcgcctg
AmiE_I122L	gatttcaccgttgttatcgagcaagaccagagtgttgtatg
AmiE_I235A	cggccgtcaaaaccgatagcggcactatgaccgaagta
AmiE_I235L	gccgtcaaaaccgataagggcactatgaccgaagt
AmiE_I38V	atcatttccgcaactttgcgggcattatccaggac
AmiE_L258A	cggatttgtgacagagacgcctgcgcgtattgaatgcc
AmiE_S162V	ccgtcatcgcagataattaatacgatcttcatgcctttcggacc
AmiE_V17T	ggcatcttgtaattcaccgtggctacgcctacggtatc
AmiE_V56I	ctgtaaagaatattcaggaaatataaccagatccatgcccggca

Note C 4.1: Command lines and supporting files for Rosetta computational design

Prepare coordinate constraints file for amiE

make_csts.sh infile.pdb > outfile.cst

Flags file used in refinement

-ex1
-ex2
-use_input_sc
-extrachi_cutoff 5
-ignore_unrecognized_res
-use_occurrence_data
#-corrections::correct
#-corrections::score:no_his_his_pairE
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation SpecialRotamer VirtualBB ShoveBB VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals pro_hydroxylated_case1 pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated tyr_phosphorylated tyr_sulfated lys_dimethylated lys_monomethylated lys_trimethylated lys_acetylated glu_carboxylated cys_acetylated tyr_diiodinated N_acetylated C_methylamidated MethylatedProteinCterm
#-output_virtual
-linmem_ig 10
-ignore_zero_occupancy false

#-out:path:pdb pdbs/
#-out:path:score scores/

Refinement command line

./path/to/rosetta/scripts -database ./path/to/rosetta/database/ -in:file:s infile.pdb -parser:protocol refine.xml -parser:script_vars res_to_fix=
22A,59A,60A,65A,103A,117A,119A,132A,134A,136A,137A,138A,139A,142A,144A,163A,164A,165A,166A,167A,168A,169A,170A,171A,174A,175A,187A,188A,189A,190A,191A,192A,193A,200A,203A,217A,218A,219A,227A,229A,230A,260A,261A,262A,263A,264A,265A,266A,267A,268A,269A,270A,271A,272A,273A,274A,275A,276A,277A,278A,279A,280A,281A,282A,283A,284A,285A,286A,287A,288A,289A,290A,291A,292A,293A,294A,295A,296A,297A,298A,299A,300A,301A,302A,303A,304A,305A,306A,307A,308A,309A,310A,311A,312A,313A,314A,315A,316A,317A,318A,319A,320A,321A,322A,323A,324A,325A,326A,327A,328A,329A,330A,331A,332A,333A,334A,335A,336A,337A,338A,339A,340A,341A -parser:script_vars pdb_reference=infile.pdb -parser:script_vars cst_full_path=infile.cst -parser:script_vars cst_value=0.4 @flags -overwrite

Command line for FilterScan

for i in {245..341}; do ./path/to/rosetta/scripts -database ./path/to/rosetta/database/ -in:file:s refinedinfile.pdb -parser:protocol filterscan.xml -parser:script_vars res_to_fix=22A,59A,60A,65A,103A,117A,119A,132A,134A,136A,137A,138A,139A,142A,144A,163A,164A,165A,166A,167A,168A,169A,170A,171A,174A,175A,187A,188A,189A,190A,191A,192A,193A,200A,203A,217A,218A,219A,227A,229A,230A,260A,261A,262A,263A,264A,265A,266A,267A,268A,269A,270A,271A,272A,273A,274A,275A,276A,277A,278A,279A,280A,281A,282A,283A,284A,285A,286A,287A,288A,289A,290A,291A,292A,293A,294A,295A,296A,297A,298A,299A,300A,301A,302A,303A,304A,305A,306A,307A,308A,309A,310A,311A,312A,313A,314A,315A,316A,317A,318A,319A,320A,321A,322A,323A,324A,325A,326A,327A,328A,329A,330A,331A,332A,333A,334A,335A,336A,337A,338A,339A,340A,341A -parser:script_vars pdb_reference=refinedinfile.pdb -parser:script_vars res_to_restrict=22A,59A,60A,65A,103A,117A,119A,132A,134A,136A,137A,138A,139A,142A,144A,163A,164A,165A,166A,167A,168A,169A,170A,171A,174A,175A,187A,188A,189A,190A,191A,192A,193A,200A,203A,217A,218A,219A,227A,229A,230A,260A,261A,262A,263A,264A,265A,266A,267A,268A,269A,270A,271A,272A,273A,274A,275A,276A,277A,278A,279A,280A,281A,282A,283A,284A,285A,286A,287A,288A,289A,290A,291A,292A,293A,294A,295A,296A,297A,298A,299A,300A,301A,302A,303A,304A,305A,306A,307A,308A,309A,310A,311A,312A,313A,314A,315A,316A,317A,318A,319A,320A,321A,322A,323A,324A,325A,326A,327A,328A,329A,330A,331A,332A,333A,334A,335A,336A,337A,338A,339A,340A,341A -parser:script_vars cst_full_path=infile.cst -parser:script_vars cst_value=0.4 -parser:script_vars scores_path=scores/ -parser:script_vars resfiles_path=resfiles/ @flags_delay -parser:script_vars current_res=${i} -overwrite; done

Note C 4.2: Rosetta scripts used in this work

Modified FilterScan script excluding PSSM input

```
<ROSETTASCRIPTS>

<SCOREFXNS>
<talaris_full weights=talaris2014>
<Reweight scoretype="coordinate_constraint" weight=%%cst_value%%/>
<Reweight scoretype="res_type_constraint" weight=0.4/>
</talaris_full>
</SCOREFXNS>

 <TASKOPERATIONS>
<InitializeFromCommandline name=init/>
<DesignAround name=des_around design_shell=0.1 resnums="%%current_res%%"
repack_shell=8.0/>
<RestrictResiduesToRepacking name=restrict_res residues="%%res_to_restrict%%"/>
<PreventResiduesFromRepacking name=fix_res reference_pdb_id="%%pdb_reference%%"
residues="%%res_to_fix%%"/>
 </TASKOPERATIONS>

<MOVERS>
<ConstraintSetMover name=add_CA_cst cst_file="%%cst_full_path%%"/>
<MinMover name=min_all scorefxn=talaris_full chi=1 bb=1 jump=0/>
</MOVERS>

<FILTERS>
<ScoreType name=stability_score_full scorefxn=talaris_full score_type=total_score
threshold=0.0/>
<Delta name=delta_score_full filter=stability_score_full upper=1 lower=0 range=0.5/>
#upper and lower are booleans. Delta filters out all the mutations that are worse or better by less
than -0.5R.E.U
<Time name=timer/>
<FilterScan name=filter_scan scorefxn=talaris_full relax_mover=min_all keep_native=1
task_operations=init,des_around,fix_res,restrict_res delta_filters=delta_score_full delta=true
resfile_name="%%resfiles_path%%designable_aa_resfile" report_all=1
delta_filter_thresholds=0.45,0.75,1.0,1.25,1.5,1.8,2.0
score_log_file="%%scores_path%%res%%current_res%%_score_full.log" dump_pdb=0/>
 </FILTERS>

<PROTOCOLS>
<Add filter=timer/>
<Add mover_name=add_CA_cst/>
<Add filter=filter_scan/>
```

<Add filter=timer/>
</PROTOCOLS>

</ROSETTASCRIPTS>

REFERENCES

REFERENCES

1. Fisher, R. A. *The Genetical Theory of Natural Selection*. (Oxford University Press, 1930).

2. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6,** 119–127 (2005).

3. Gillespie, J. H. Molecular Evolution Over the Mutational Landscape. *Evolution (N. Y).* **38,** 1116–1129 (1984).

4. Orr, H. A. The Population Genetics of Adaptation: The Distribution of Factors Fixed During Adaptive Evolution. *Evolution (N. Y).* **52,** 935–949 (1998).

5. Orr, H. A. The population genetics of beneficial mutations. *Philos. Trans. R. Soc. London B Biol. Sci.* **365,** 1195–1201 (2010).

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

7. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8,** 15695 (2017).

8. Baker, D. & Agard, D. A. Kinetics Thermodynamics in Protein Folding. *Biochemistry* **33,** 7505–7509 (1994).

9. Shakhnovich, E. I. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7,** 29–40 (1997).

10. Serohijos, A. W. R. & Shakhnovich, E. I. Contribution of Selection for Protein Folding Stability in Shaping the Patterns of Polymorphisms in Coding Regions. *Mol. Biol. Evol.* **31,** 165–176 (2014).

11. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596–604 (2009).

12. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science (80-.).* **324,** 203–208 (2009).

13. Zeldovich, K. B., Chen, P. & Shakhnovich, E. I. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl. Acad. Sci.* **104,** 16152–16157 (2007).

14. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.* **103,** 5869–5874 (2006).

15. Klesmith, J. R., Bacik, J., Wrenbeck, E. E., Michalczyk, R. & Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci.* **114,** 2265–2270 (2017).

16. Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S. & Babbitt, P. C. Stability for Function Trade-Offs in the Enolase Superfamily 'Catalytic Module'. *Biochemistry* **46,** 6688–6695 (2007).

17. Tokuriki, N. *et al.* Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nat. Commun.* **3,** 1257 (2012).

18. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459,** 668–675 (2009).

19. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4,** e1000002 (2008).

20. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression. *Mol. Cell* **63,** 337–346 (2016).

21. Andrade, J., Karmali, A., Carrondo, M. A. & Frazao, C. Structure of Amidase from Pseudomonas aeruginosa Showing a Trapped Acyl Transfer Reaction Intermediate State *. *J. Biol. Chem.* **282,** 19598–19605 (2007).

22. Arnold, F. H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409,** 253–257 (2001).

23. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–8 (2012).

24. Morley, K. L. & Kazlauskas, R. J. Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.* **23,** 231–237 (2005).

25. Klesmith, J. R., Bacik, J., Michalczyk, R. & Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth. Biol.* **4,** 1235–1243 (2015).

26. Polizzi, K. M., Bommarius, A. S., Broering, J. M. & Chaparro-Riggers, J. F. Stability of biocatalysts. *Curr. Opin. Chem. Biol.* **11,** 220–225 (2007).

27. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485,** 185–194 (2012).

28. Frokjaer, S. & Otzen, D. E. Protein Drug Stability: A Formulation Challenge. *Nat. Rev.* **4,** 298–306 (2005).

29. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci.* **82,** 488–492 (1985).

30. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

31. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357,** 289–298 (2006).

CHAPTER FIVE

Improving the expression of a polyketide synthase in a biosynthetic pathway using deep mutational scanning and GFP-fusion screening

ABSTRACT

Engineering organisms for the production of various chemicals, therapeutics, and fuels is a promising and sustainable alternative to other sources. To this end, enzymes, the entities responsible for *in vivo* conversion of feedstocks into desired products, are heterologously expressed in chassis organisms such yeast or bacteria. However, poor expression of heterologous proteins can create bottlenecks in pathways. These efforts are complicated by the low probability of proteins to fold and function in non-native environments, often at concentrations far past their solubility limit. Further, mutations that improve the stability and solubility of a protein often come with a tradeoff for activity. In this work, we sought to improve the stability and solubility of a Type III polyketide synthase (PKS) using GFP-fusion high-throughput screening coupled to deep mutational scanning. This PKS comes from the tropane alkaloids biosynthetic pathway from *Atropa belladonna*, of which we aim to reconstruct a portion of the pathway, including the PKS, in *Saccharomyces cerevisiae*. Hits from the screen were filtered for their probability to be catalytically neutral, and combinatorial libraries were prepared and screened for improved expression. Stabilized variants containing ≥12 total mutations were identified, with the best variant providing a 6.2-fold improvement in expression in *E. coli*. However, all initial designs had negligible activity. Analysis of individual PKS point mutants revealed that at least two of the included mutations significantly reduce activity. Future work includes generating backcrosses of these inactivating mutations on the stabilized designs and engineering the tropane alkaloids pathway in yeast.

INTRODUCTION

Biomanufacturing – the production of molecules by engineered microorganisms – is a viable and sustainable alternative to traditional chemical synthetic routes[1]. Key factors influencing the rapid advancement of this field include the dramatic increase of available gene coding sequences, improved accuracy and reduced cost of synthetic DNA synthesis and assembly[2], and improved tools for engineering biology[3–7]. However, the economization of biomanufacturing processes to compete with alternative sources (i.e. fluctuating crude oil prices) remains a grand challenge. For example, reported titers from the production of plant secondary metabolites in microbes are generally miniscule (microgram per liter quantities), necessitating significant optimization efforts to compete with plant-derived or traditional chemical synthetic routes.

At the heart of these processes are the enzymes responsible for chemical transformation; a typical engineered organism will express at least five heterologous proteins. While other factors are certainly important (toxicity, metabolic flux balancing, localization, etc.), the efficiency of each enzyme step is a fundamental determinant for specific productivity and production titer. Flux through an enzyme, J_E, can be modeled as a function of total active enzyme, $[E]_t$, its catalytic efficiency (K_M and k_{cat}), and the thermodynamic reversibility of a reaction[8]:

$$J_E = [E]_t \times \left(k_{cat} \frac{\frac{[S]}{K_M}}{1 + \frac{[S]}{K_M}} \right) \times \left(1 - e^{\frac{\Delta G_{rxn}}{RT}} \right) \quad (1)$$

where $\Delta G_{rxn} = -RTK'_{eq} + RT\ln(p/s)$, with p = concentration of product and s = concentration of substrate. At equilibrium, ΔG_{rxn} will go to zero and thus thermodynamic reversibility term will

become zero. Provided that an enzyme is not limited by catalytic efficiency, poor functional expression can significantly hinder productivity and efficiency. Indeed, a review of current literature reveals that several recently engineered pathways are limited by poor expression of at least one enzyme (**Table D 5.1**)[9–20]. Native proteins are marginally stable, and their native expression levels are often at their solubility limit. A common strategy to overcome a bottleneck enzyme is to overexpress the protein. However, overexpression of heterologous genes can lead to poor solubility and aggregation.

Recently, Klesmith et al. demonstrated that by improving the apparent melting temperature of a *Lipomyces starkeyi* levoglucosan kinase by 5.1°C while maintaining near wild-type activity, growth rates of levoglucosan kinase expressing *E. coli* fed levoglucosan as a sole carbon source were improved by 15-fold from the wild-type enzyme[21]. Similarly, Xie et al. engineered improved solubility of Simvastatin synthase in *E. coli* and achieved approximately 50% increase in whole cell activity and solubility[22]. While these examples are demonstrative, this strategy has not been widely adopted amongst pathway engineers. Stability and solubility engineering of enzymes is complicated by the need to maintain functional enzymatic activity, and mutations with stabilizing effects often come with a tradeoff for protein function.

To address this challenge, we recently published on identifying stabilizing 'hits' using high-throughput screens for stability and solubility in deep mutational scanning experiments[23]. Existing comprehensive single-mutation functional datasets for two enzymes were compared against datasets generated with the stability screens. In short, we found a 90% probability of choosing a catalytically neutral mutation by filtering out mutations that were near the active site, not evolutionarily conserved, or buried in the protein core.

As a rigorous test and application of this method, in this work we sought to improve the expression of an enzyme without an existing functional dataset or crystal structure. As a model system, we performed a GFP-fusion stability scan on a Type III polyketide synthase (PKS) from *Atropa belladonna* (Ab) that expresses very poorly in both bacterial and yeast systems. This enzyme is part of the Tropane Alkaloids (TA) pathway from Ab that was recently elucidated by the Barry Lab at Michigan State University (**Figure 5.1**, unpublished data).

RESULTS

As a model system to test our hypothesis that improving protein expression boosts the productivity of engineered metabolic pathways, we sought to reconstruct a portion of the TA pathway in *Saccharomyces cerevisiae*. The engineered pathway begins with the pathway precursor putrescine followed by chemical transformations to the TA pharmacore tropine. This transformation is catalyzed by five enzymes: putrescine methyltransferase (PMT), methyl-putrescine amine oxidase (MPO), a Type III polyketide synthase (PKS), tropinone synthase (TS), and tropinone reductase (TRI) (**Figure 5.1a**). We first evaluated the localization and expression of native Ab genes in yeast by generating EGFP-tagged fusions of each gene under the control of a galactose inducible promoter. Fluorescence microscopy of *S. cerevisiae* strain BY4710[24] expressing EGFP-fusions revealed that all genes except for MPO were expressed in the cytosol (**Figure 5.1b**). The native MPO contains a canonical Ala-Lys-Leu C-terminal peroxisomal targeting sequence (PTS). Co-expression of an RFP-tagged peroxisomal protein[25] and EGFP-MPO confirmed that the MPO localizes to the yeast peroxisome (**Figure 5.1b**).

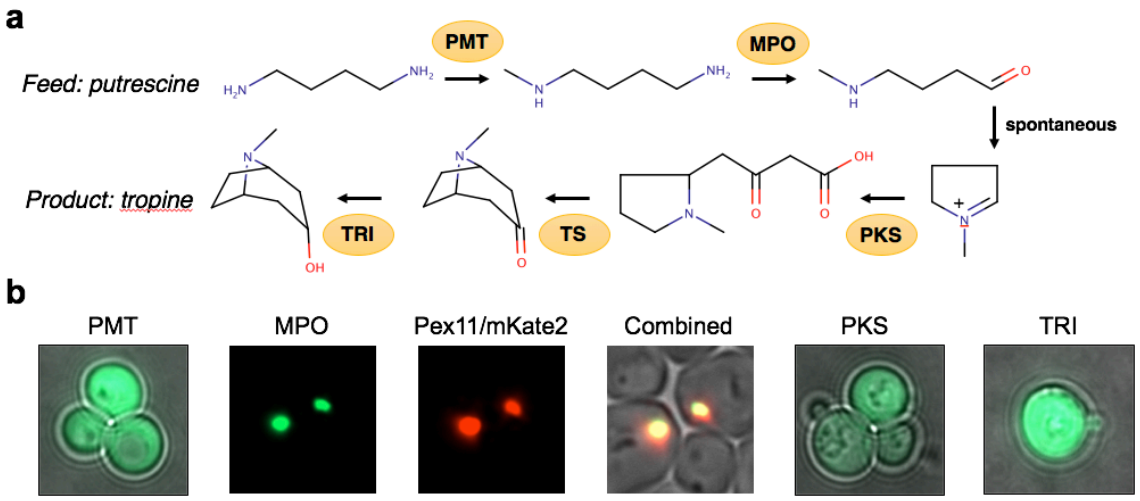


Figure 5.1: Overview of the Tropane Alkaloids (TA) pathway enzymes. a.) The pathway precursor, putrescine, will be fed in the growth medium of yeast cells expressing five enzymes:

Figure 5.1 (cont'd) PMT, MPO, TRI, TS, and PKS. The final products, tropinone and tropine will be detected from the cultures using LC/MS methods developed by the Barry Lab. b.) Fluorescence microscopy images of EGFP-tagged Ab genes. PMT, PKS, and TRI indicate cytosolic expression, whereas MPO localizes to the yeast peroxisome as confirmed by co-localization with Pex11-mKate2 fusion protein.

We quantified the mean fluorescence of the EGFP-tagged Ab gene products by flow cytometry and found that the PKS expressing cells were significantly less fluorescent compared to the other genes, indicating poor expression (**Figure 5.2**). Attempts by the Barry Lab to express and purify PKS for characterization yielded extremely low levels of active protein (~5 mg/L yield from auto-induction cultures, unpublished data). They found that essentially all of the protein was insoluble and that activity sharply declines at temperatures in excess of 25°C. Together, these data indicate that the PKS expresses poorly in both bacterial and yeast hosts and support engineering the enzyme to find variants with improved stability and solubility.

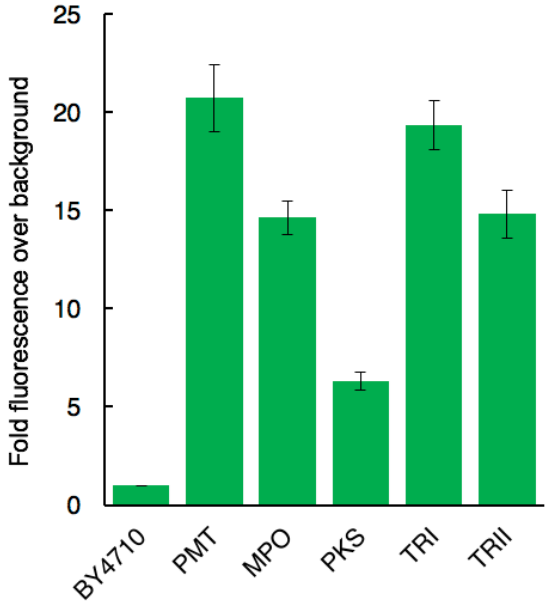


Figure 5.2: Relative fluorescence of EGFP-tagged Ab genes in yeast. Fluorescence of *S. cerevisiae* strain BY4710[24] cells expressing EGFP-tagged Ab genes under galactose induction was quantified using flow cytometry. Error bars represent the standard deviation of at least two independent measurements. TRII is a homolog of TRI that produces pseudo-tropine.

In effort to improve the expression of the PKS, we sought to use deep mutational scanning coupled to a high-throughput screen for stability and solubility. Due to its experimental ease, we first explored the use of yeast surface display coupled to FACS[26]. The PKS coding sequence was cloned into the pETConNK backbone[23,27] and expressed with galactose induction at 22°C in EBY100 cells. Initial tests for display proved futile (**Figure D 5.1**). We then tested several alternate induction temperatures (18-30°C) as well as mutating a potential glycosylation site at Asn339. We hypothesized that glycosylation of this asparagine could disrupt folding and thus ability to display, so we made and tested PKS_N339A using nicking mutagenesis[27]. However, despite these troubleshooting efforts, we were unable to successfully display the PKS on the yeast surface.

We next attempted use of a GFP-fusion stability screen. The concept of GFP-fusion is to genetically encode a gene of interest linked to GFP generating a fusion protein (**Figure D 5.2a**), such that the folding probability of GFP is tied to the folding probability of the gene of interest. Expression of a protein library can then be screened by fluorescence intensity using FACS (**Figure D 5.2b**). Recently, the use of a GFP-fusion stability screen[28] in deep mutational scanning experiments was validated in our lab using a levoglucosan kinase (LGK) from *Lipomyces starkeyi* as a model system (unpublished data). The screen was able to correctly identify 9/12 known stabilizing LGK mutations (P-value < 0.0001, Fisher's exact test). Thus, GFP-fusion has been validated in a deep mutational scanning pipeline.

The objective of this project was to perform a GFP-fusion deep mutational scan on PKS, filter the resulting hits for probability of maintaining catalytic activity using recently published classification methods[23], and then generate a combinatorial library to identify active variants with improved expression (**Figure 5.3a**). We generated a comprehensive single-site saturation

mutagenesis library of PKS using nicking mutagenesis[27]. Plasmid DNA libraries were transformed into BL21*(DE3) and protein expression was induced with IPTG (see **MATERIALS AND METHODS**). Individual cells were sorted using FACS and two populations were collected: a reference population and the top 8-10% of cells based on GFP fluorescence intensity. The resulting samples were deep sequenced and the population counts and enrichment ratios of wild-type and each variant were calculated using Enrich[29] (see **Table D 5.2** for library statistics). A normalized stability metric, ζ, for each variant was calculated as outlined in Kowalsky et al.[30], where a stability metric of zero corresponds to wild-type, above zero are beneficial mutations (more stable), and below zero are deleterious. Unfortunately, the reference population for the gene tile covering residues 157-234 did not grow after FACS, thus we chose to omit these positions from further analysis.

Deep sequencing of the reference populations revealed 84.3% coverage of single nonsynonymous (NS) mutations (5107/6060). Nonsense mutations had a mean stability metric of -0.653 ± 0.48 (1 s.d.), which was significantly lower than the mean of -0.0561 ± 0.54 for missense mutations (*P-value* < 0.0001, two-tailed unpaired Student's T-test). To evaluate the reproducibility of the method, we performed replicate sorting, deep sequencing, and analysis for one gene tile. The Pearson's correlation coefficient between replicates was found to be 0.72, which is low compared to previous deep mutational scanning experiments (coefficients of 0.85[23] and 0.93[31] have been previously reported from our lab). As reproducibility generally improves with increasing depth of sequencing coverage, we calculated Pearson's correlation coefficients for mutations with at least 100 read counts in the reference population and found the coefficient improves to 0.83. Thus, the relatively low depth of coverage in this experiment partially but not completely explains the relatively high variance between replicates.

Since we are interested in 'improved' variants, we next asked how correlation scales with coverage for variants with at least 20% improvement in stability metric (ζ>0.15). We found that variants with ≥50 average selected read counts had a Pearson's of 0.84, which we deemed a reasonable threshold for reliability of the deep sequencing experiment to yield predictive results on a mutation's stability effect (**Table 5.1**).

Table 5.1: Correlation of stability metrics for beneficial mutations from replicate GFP-fusion experiments based on depth of sequencing coverage. Average selected read counts between replicates were calculated and Pearson's product moment correlation coefficients were determined above the indicated read count thresholds. N indicates the number of mutations at each threshold.

Average selected read count threshold	N	Pearson's
≥10	298	0.70
≥30	280	0.71
≥50	247	0.84
≥100	193	0.90

The GFP-fusion experiment identified an astounding 1,115 beneficial missense mutations (ζ>0.15) with ≥50 selected read counts (19.4% of total tested). As the objective was to generate a stabilized PKS variant with wild-type catalytic activity, hits were filtered using a multiple filter approach as validated in Klesmith et al.[23] for their probability to be catalytically neutral[23]. In brief, hits at positions within 15 Å of the active site or with a PSSM score <3 were removed. Mutations with PSSM scores ≥3 represent mutations that have been evolutionarily conserved, and thus are less likely to impact catalytic function (see **MATERIALS AND METHODS** for details of PSSM generation). Proximity to the active site was calculated using a homology structure for PKS generated with I-TASSER with default options[32]. The resulting set of hits post-filter was comprised of 38 mutations at 35 unique positions (**Table D 5.3**).

Using the homology structure, we selected 23 mutations to include in a combinatorial library[27]. Included in this set were 5 mutations that did not pass the stringent filtering criteria, but were included due to having either relatively high stability metric scores and/or PSSM scores of ≥0 (**Table D 5.4**). Initially, BL21*(DE3) cells expressing the combinatorial library were cultured, induced with IPTG, plated and grown at 30°C, and visually screened for fluorescence intensity. 20 variants were picked and their fluorescence was quantified using flow cytometry (**Figure D 5.3**). The best variants from this initial screening effort only provided approximately 3-fold improvement over wild-type. Because the number of library members one can reasonably screen on plates is low (~10^2), we performed a FACS sort to enrich the library for the top 3% of cells based on GFP fluorescence intensity. Sorted cells were plated and visually screened as before, and 10 additional variants were picked for isogenic characterization. The best variant, PKS.21, provided a 6.2 ± 0.01-fold improvement in fluorescence over wild-type.

Next, we selected the top four variants (PKS.4, PKS.20, PKS.21, PKS.23), sequenced and cloned them into a GST-tagged expression vector, and characterized them for relative catalytic activity. Each design had ≥12 mutations, with 8 mutations shared amongst all designs (V12I, S37A, P106A, M115R, A121G, A143V, T245A, and S284K, **Figure 5.3a**). Based on the homology structure, these mutations generally appear to alter surface charge characteristics, core packing, loop flexibility, or dimeric interface contacts (**Figure. 5.3b**). For example, N64E/D introduce a negative charge to a patch on the surface that is otherwise positive. V282I is a hydrophobic-hydrophobic mutation in the core of the protein that presumably improves hydrophobic packing. Lastly, A121G likely improves loop flexibility.

Interestingly, when we screened these hits for activity with a lysate assay we found that all had no detectable activity except for PKS.20, which had approximately 0.1% of the activity

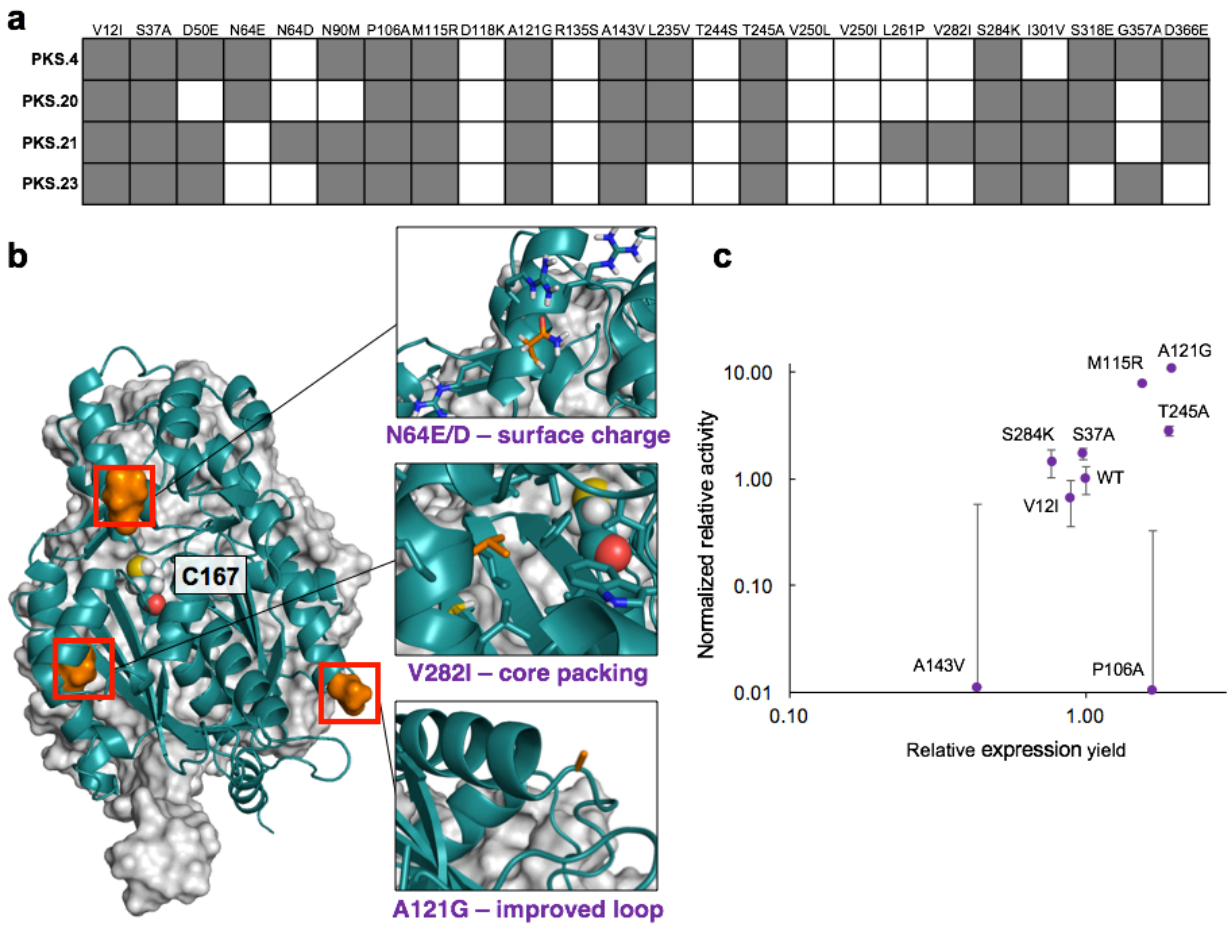


Figure 5.3: Combinatorial PKS hits. a.) Sequencing of the top four combinatorial hits revealed several shared mutations. Grey shading indicates the mutation is present. b.) Structural analysis of beneficial mutations indicates that many improve surface charge characteristics, hydrophobic core packing, and secondary structure like loops. The grey surface representation is the dimer subunit. C167 is the putative catalytic nucleophile. c.) Point-mutant analysis of the 8 shared mutations. Lysates of GFP-fused PKS mutants were tested for their activity and fluorescence intensity (expression yield). Error bars represent 1 s.d. of technical replicates. Axis are both logarithmic scale (base 10).

compared with wild-type (data not shown). We hypothesized that one or more of the 8 shared mutations were responsible for destroying enzyme activity. To test this hypothesis, we generated the 8 point-mutations in the GFP-fusion background using nicking mutagenesis[27] and performed lysate activity assays. We found two mutations, P106A and A143V, that reduced activity to 0.93% and 0.02% of wild-type activity (**Figure 5.3c**). Not surprisingly, these two mutations were

ones that did not pass the filtering criteria because they had low PSSM scores of -2 (P106A) and 0 (A143V), however were included in the library as they had stability scores of 0.641 and 0.257, respectively. This result underlines the importance of the filtering criteria: mutations that improve stability but are not evolutionarily conserved are significantly more probable of being deleterious for catalytic function.

DISCUSSION AND FUTURE WORK

In this work, we performed a high-throughput screen for stability and solubility to test thousands of mutations on a protein sequence. Deep sequencing driven protein science enables the generation of previously unthinkable amounts of mutational data[33]. Applications stretch as far as studying enzyme function[21,34–38,31,39,40], probing mechanics of evolution[31,41,42], and antibody engineering[43,44].The ability of deep mutational scanning – counting DNA sequences from unselected and selected populations – to recapitulate various biological, chemical, and physical phenomena is wholly dependent on the quality of the screen or selection method used. Replication of experiment is an important metric for all methods, and the correlations observed in this work indicate that GFP-fusion, at least for PKS, is not as sensitive as other high-throughput screening technologies. Nevertheless, computational methods for filtering hits can aid in generating stabilized protein designs. However, if given the option a more robust stability screen such as yeast surface display[23,26] is more desirable.

There are two important takeaways from this project. First, this work clearly validates the filtering method previously developed by Klesmith et al.[23]. Results from the point-mutation analysis indicate that although certain mutations provide stabilizing effects, if they were not conserved in nature they are likely to be deleterious for function. Indeed, proline 106 is a

canonical example of this stability/function trade-off. P106 lies in the middle of a helix, where prolines are generally disfavored, and the solubility screen indicates that several other residues at this position improve overall stability of the protein. However, the PSSM indicates that proline is highly conserved and thus important to catalytic function. The P106A variant increased the solubility of the PKS-EGFP fusion but almost completely ablated activity. Since all characterized PKS hits (PKS.4, PKS.20, PKS.21, and PKS.23) contained P106A and A143V that did not pass the filter, immediate next steps include backcrossing these mutations and testing the resulting enzyme variants for activity.

The second important takeaway is now that the filtering method has been validated on 3 different enzymes, library size from the outset can be significantly decreased; any mutation that does not pass the filtering criteria need not be included in the library. In the absence of stability metric data, for the PKS removing mutations with PSSM scores <3 reduces the library size from 7,448 (392 positions with 19 amino acid substitutions) down to 154, and removing positions within 15 Å of the active site reduces the library size down to 123. Testing approximately one hundred mutations versus thousands (comprehensive scan of a gene) is certainly more practical and economical. Notably, nicking mutagenesis developed in **Chapter 2** enables such efficient library generation.

A remaining objective for this project is to reconstruct the TA pathway in yeast. This will be accomplished using the hierarchal MoClo cloning strategy and a kit of characterized parts for yeast[45] developed by John Dueber and company at University of California, Berkeley. Multi-gene cassettes will be integrated into the chromosome of yeast strain BY4742[24], as there are several available options for auxotrophic selection. Initial designs will feature genes placed under the control of medium strength constitutive promoters with the objective of detecting

tropinone and/or tropine in 72hr cultures. Once confirmed that this is the minimum set of genes to produce tropane alkaloids in yeast, optimization of expression elements (promoters, copy number, integration location, etc.) will be carried out using a modular approach. In brief, optimization of the PMT+MPO, then PMT+MPO+PKS, and lastly PMT+MPO+PKS+TS+TRI will be done. Additionally, three putative cytochrome P450 reductases (CPR) for TS will be screened for activity and their expression level will be optimized with respect to TS. Ultimately, the objective is to test stabilized active PKS designs versus the wild-type in the context of the engineered pathway to see the effects on pathway productivity.

MATERIALS AND METHODS

Reagents

Chemicals were sourced from Sigma-Aldrich unless otherwise noted. Mutagenic oligos were designed using the QuikChange Primer Design Program (Agilent, Santa Clara, CA). All oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA). All minipreps were done using the Monarch Plasmid Miniprep Kit (New England Biolabs).

Plasmid construction

Yeast gateway expression constructs used for fluorescence microscopy were made using plasmids from the Yeast Gateway Kit from the Lindquist Lab (available from www.addgene.com). PMT, PKS, TRI, and TRII *Atropa belladonna* gene sequences were cloned from pENTR/D-TOPO entry vectors into the pAG424GAL_EGFP_ccdB plasmid using LR cloning kit (Thermo Fisher Scientific). MPO was cloned as above but into a gateway plasmid modified to harbor an RFP peroxisomal marker. The Pex11-mKate2 transcriptional unit from

pWCD2520 (gifted from the Dueber Lab) was PCR amplified attaching BsmBI sites on either side. The resulting amplicon was subcloned into the pAG424GAL-EGFP-ccdB plasmid between the two BsmBI sites, generating construct pAG424GAL-EGFP-MPO_Pex11/mKate2.

The pET29NK_<gene_of_interest>/mGFPmut3 vector was constructed by Justin Klesmith as follows. Overhang PCR was used attaching a 5' XhoI site and a 3' His6x, stop codon, BbvCI site to mGFPmut3 from a plasmid based from pJK_proB_GFP (45). Similarly, overhang PCR was used to add a BbvCI site to pET29b just after the stop codon in the plasmid. The mGFPmut3 construct was cloned between the XhoI and BbvCI sites using standard techniques to make the fusion construct <gene_of_interest>-Leu-Glu-mGFPmut3-His6x. pET29NK-PKS-RD/mGFPmut3 was constructed by overhang PCR of the PKS coding sequence attaching NdeI and XhoI sites for ligation into the pET29NK-GOI/mGFPmut3 construct.

pGEX expression vectors were made by subcloning PKS genes between the BamHI and SmaI sites of the pGEX-4T1 backbone. Wild-type and variant PKS sequences were amplified by overhang PCR attaching 5' BamHI and 3' SmaI restriction sites and then cloned into the pGEX backbone following standard restriction digest and ligation protocols.

PKS comprehensive point-mutant library construction

Nicking mutagenesis was used to generate a comprehensive single-site mutagenesis library on the pET29NK_PKS-RD/mGFPmut3 plasmid[27]. Degenerate NNK primers covering residues Lys8 to Arg388 were used in 5 separate mutagenesis reactions to generate 5 gene tiles: T1 (K8-E78), T2 (I79-S156), T3 (V157-G234), T4 (L235-I312), T5 (V313-R388) (see **Table D 5.2**). Mutagenesis reaction products were transformed into XL1-Blue Electrocompetent Cells

(Agilent) and plated on 245mm x 245 mm Large Bioassay Dishes (Sigma). The following day, cells were scraped and plasmids harvested with a miniprep.

30 ng of library plasmid DNA was transformed into electrocompetent E. coli BL21*(DE3) cells and plated as above. The following day, cells were scraped and used to inoculate a 100 mL LB culture at an initial OD_{600} of 0.05 and grown at 30°C, 250 rpm. Once the cultures reached an OD_{600}=0.4-0.6, DMSO was added (7% v/v) and 1 mL aliquots were flash frozen in liquid nitrogen.

FACS screening of GFP-fusion libraries

Kanamycin was used at a final concentration of 50 μg/mL in all cultures. Library cell stocks were thawed on ice for 30-45 minutes and washed twice in TB media. For each library, 3 mL TB cultures were inoculated to an initial OD_{600}=0.05 in Hungate tubes and grown at 30°C with 250 rpm shaking. Once the cultures reached an OD_{600}=0.8-1.6, cultures were diluted into a fresh 3 mL TB culture in Hungate tubes to an initial OD_{600}=0.0025. After approximately 4-5 generations (OD_{600}=0.05-0.08) IPTG was added to a final concentration of 250 μM. Once the cultures reached an OD_{600}=0.25-0.3, 1 mL was pelleted and washed with cold sterile PBS twice.

Cells were sorted on a BD Influx cell sorter. 700,000 cells each from two populations were collected for each sample: a reference population (FSC vs. SSC gate), and a selected population (intersection of FSC vs. SSC gate and top 8-10% based on GFP fluorescence intensity with a 530/40 nm filter [488 nm]). The collected cells were added to 10 mL TB media and grown at 25°C with 250 rpm shaking until they reached an OD_{600}=0.3-0.6. Cells were pelleted and stored at -20°C until the DNA was extracted with a miniprep.

DNA deep sequencing and analysis

Library DNA was prepared for deep sequencing using Method B PCR amplification as described in Kowalsky et al.[30] (PCR primers listed in **Table D 5.5**). Amplicons were cleaned using Agencourt AMPure XP Beads (Beckman Coulter) and quantified using Quant-iT PicoGreen reagent (Life Technologies). Deep sequencing was performed on an Illumina MiSeq with 250 bp paired-end reads. The resulting data was processed using Enrich[29] and custom scripts freely available from GitHub (user JKlesmith). Stability scores for each mutant were calculated exactly as described in Klesmith et al.[23].

PKS PSSM generation

A PSSM for the Ab PKS gene was generating following similar methodologies outlined in Goldenzweig et al.[46] and Klesmith et al.[23]. A BLASTp search[47] of the nonredundant protein database was done for the PKS sequence with an e-value cutoff of 10^{-4}, excluding synthetic and engineered items from the search. The top 20,000 results were saved and sequences with less than 30% sequence identity and/or 60% coverage of the query sequence were removed. Hits were clustered using Cd-hit[48] with a 98% threshold and the top 500 clusters were aligned using MUSCLE[49]. The alignment was split into 20 amino acid sections (to reduce gap penalty) and then PSI-BLAST[50] was used to generate a PSSM.

Combinatorial library generation and screening methods

The single and multi-site nicking mutagenesis protocol[27] was used to generate combinatorial PKS libraries. Two separate reactions were performed at a primer:template molar ratio of 3:1 and 10:1 in effort to obtain mutants with a range in number of total mutations. The

two reactions were transformed into XL1-Blue Electrocompetent Cells (Agilent) and plated on 245 mm x 245 mm Large Bioassay Dishes (Sigma). Cells were scraped and plasmids were harvested with a miniprep. 10 ng of library DNA was transformed into electrocompetent BL21*(DE3) and plated as above. The following day, cell stocks were prepared from cell scrapings following the same methods used for the comprehensive single-mutation PKS libraries.

For the first round of plate screening, library cell stocks were thawed on ice for 30-45 minutes and then washed twice in TB. An overnight TB cultures was started at an initial OD_{600}=0.05. In the morning, a fresh 3 mL TB culture was inoculated to an OD_{600}=0.02 in Hungate tubes. Once the cultures reached an OD_{600}=0.2-0.3, IPTG was added to a final concentration of 250 μM. Cells were grown for 2.5 hours and then plated on plain LB-Agar plates and grown at 30°C overnight. Individual colonies were visually screened for fluorescence intensity and 'winners' were picked to be grown in isogenic 2 mL TB cultures overnight. PKS.1-PKS.10 and PKS.11-PKS.20 originated from the 3:1 and 10:1 primer:template ratio libraries, respectively. The next day, glycerol stocks were made of each variant for downstream characterization.

In effort to enrich for the best variants, a second round of plate screening was performed following an initial FACS sort. Cell stocks were thawed and cultured and induced as above, and then washed twice in cold filtered PBS. Libraries were sorted on a BD Influx sorter and 20,000 cells from the intersection of the FSC vs. SSC gate and the top 3% based on GFP fluorescence intensity (530/40 nm filter [488 nm]) were collected. 2 mL of TB media was added to the collected cells and plated on 245 mm x 245 mm Large Bioassay Dishes (Sigma) with LB-Agar and Kanamycin at a final concentration of 50 μg/mL. Plates were visually screened as before, and 'winner' colonies picked for growth in isogenic cultures.

Characterization of combinatorial hits and point-mutations

Quantification of the mean fluorescence intensity of GFP-fused PKS hits was performed as follows. Overnight cultures of wild-type and variant PKS were started from isogenic glycerol stocks in 2 mL TB. In the morning, 3 mL TB cultures were prepared from overnight cultures at an initial OD_{600}=0.02 in Hungate tubes. Cells were grown until the OD_{600} reached 0.2-0.3 and then IPTG was added at a final concentration of 250 μM. After 2.5 hours, 1 mL of culture was removed and washed twice in filtered PBS. Cells were diluted with PBS to an OD600=0.1 and fluorescence intensity was measured on a BD Acuri C6 Flow Cytometer. The sort was run for 50,000 events in a polygon gate drawn on the FSC vs. SSC plot. Mean fluorescence was obtained on the plot of the intersection of the FSC vs. SSC and an FL1 vs. Count plot, where FL1 represents fluorescence intensity using a 510/15 nm filter.

Auto-induction cultures of wild-type and variant PKS proteins for initial activity assays were prepared as follows. Isogenic overnight cultures in 2 mL TB were started from glycerol stocks of BL21*(DE3) cells harboring the pGEX-PKS plasmids. In the morning, 1 mL of overnight culture was removed, spun down at 8,000xg for 3 minutes, and resuspended in 1 mL of standard auto-induction media[37]. The OD_{600} was measured and 1 mL solutions of cells at an OD_{600}=0.5 in auto-induction media were prepared. 500 μL of these solutions were added to 25 mL pre-warmed auto-induction media in 250 mL Erlenmeyer flasks and were grown at 30°C with 250 rpm shaking for 6 hours. Cultures were then switched to grow at 18°C with 250 rpm shaking for 20 hours. Cultures were spin down at 8000xg for 20 minutes at 4°C and the wet cell weight recorded. Cell pellets were washed once and then resuspended in PBS to a final volume of 10 mL at an OD_{600}=10. Samples were pelleted and stored at -80°C for future analysis. To prepare lysates, cell pellets were thawed and resuspended to an OD=2.5 in resuspension buffer

(10% glycerol, 147mM NaCl, 4.5 KCl, 100 mM HEPES pH 8.0, 1x Sigma Protease Inhibitor Cocktail, 5 mM DTT, 1 mg/mL lysozyme). 1 mL of resuspension was lysed in a 1.5 mL microfuge tube on ice using a sonicator fitted with a 1/8" horn with the following settings: 3 sec on, 10 sec off, 60 sec total on time, 37% amplitude.

Lysate activity assays were performed by Matt Bedewitz in the Barry Lab at Michigan State University. Standard activity assays for AbPKS were performed using 25 mM potassium phosphate buffer pH 8.0, 50 μM *N*-methyl-Δ^1-pyrrolinium hydrochloride, and 100 μM malonyl-coenzyme A lithium salt (Santa Cruz Biotechnology Cat. No. sc-215286) in 50 μL, using 5 μL of crude lysate. Reactions were stopped using 100 μL 2% formic acid, 200 mM ammonium formate, 5% methanol, and 2 μM Telmisartan as an internal standard. Reaction products were analyzed using a Waters Acquity TQ-D Mass Spectrometer coupled to a Waters Acquity UPLC system. Parameters for electrospray ionization in positive-ion mode were as follows: 2.99 kV capillary voltage, source temperature of 130°C, desolvation temperature of 350°C, and desolvation gas flow of 700 L/h, with MS/MS transitions as provided in **Table D 5.6**, using the gradient described in **Table D 5.7** with a flow rate of 0.3 mL/min. Chromatography was performed using an Ascentis Express PFPP column (2.1 × 100 mm with 2.7-μm particle size) with an oven temperature of 50°C and an injection volume of 10 μL.

Where reaction products were quantified, 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid was quantified using a standard generated the same day via alkaline hydrolysis of 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid methyl ester. This hydrolysis was performed as follows: 12 μL 25 mM 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid methyl ester in THF was added to 138 μL of THF in a glass vial. The hydrolysis was begun by addition of 150 μL of 0.335 M ammonium hydroxide. Hydrolyses were performed for 4 h at 37° C and quenched by addition of

300 μL 0.26 M ammonium formate and 1% formic acid. A standard curve of 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid was determined from this solution by subtraction of unreacted 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid methyl ester and spontaneous decarboxylation product hygrine from the calculated concentration of standard. Cuscohygrine in reactions was quantified using a cuscohygroline standard. All other compounds were quantified using authentic standards. *N*-methyl-Δ^1-pyrrolinium hydrochloride, 4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid methyl ester, and cuscohygroline were kindly provided by John d'Auria, Texas Tech.

Lysates for PKS point-mutant activity assays were prepared as follows. 3 mL TB cultures were inoculated to an initial OD_{600}=0.05 and grown until OD_{600}=0.15. IPTG was added to a final concentration of 250 μM, and cells were grown for 3 hours. Cultures were transferred to ice, spun down at 8,000xg for 5 minutes, washed twice in cold filtered PBS, and resuspended in 1 mL resuspension buffer. Cells were lysed by sonication as above. GFP fluorescence was quantified on a BioTek Hybrid plate reader in 96-well black round-bottom plates. 200 μL lysate was quantified with the following parameters: excitation=485 nm, emission=507 nm, gain=50, height=0.7. Lysate activity assays were performed as described above.

APPENDIX

APPENDIX

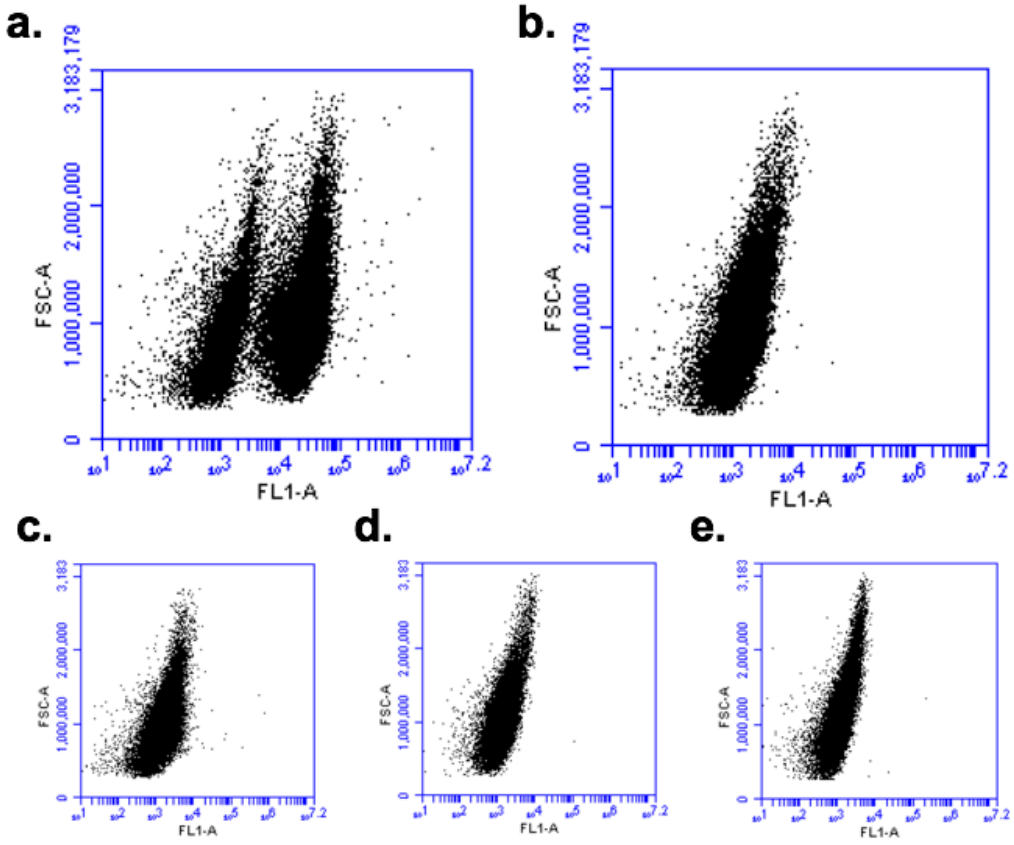


Figure D 5.1: Display of PKS on the surface of yeast proves unsuccessful. a.) A positive control for display yields two distinct populations: a non-displaying (leftmost) and a displaying (rightmost). B.) Initial attempts to display the PKS at 22°C with galactose induction were unsuccessful. Several induction conditions were tested, including temperatures ranging from 18°C (c) up to 30°C (d). We also mutated a potential glycosylation site, Asn339, to alanine, however this also was unsuccessful (e).

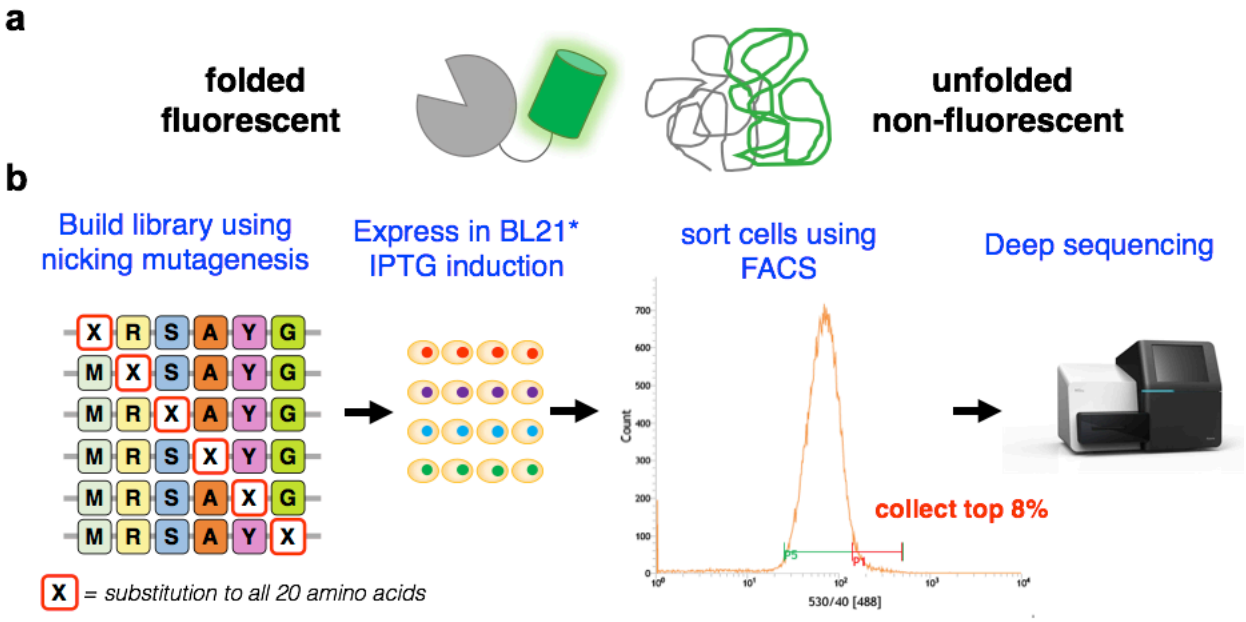


Figure D 5.2: Overview of GFP-fusion deep mutational scanning experiment. a.) A protein of interest is genetically encoded as a fusion to GFP. Upon expression, folded proteins will permit the folding and subsequent chromophore formation of GFP, while unfolded proteins will be non-fluorescent. b.) In the GFP-fusion deep mutational scanning experiment, a comprehensive site-saturation library of PKS was generated using nicking mutagenesis[27], expressed in BL21*(DE3) with IPTG induction, and sorted using FACS. The resulting libraries were then deep sequenced.

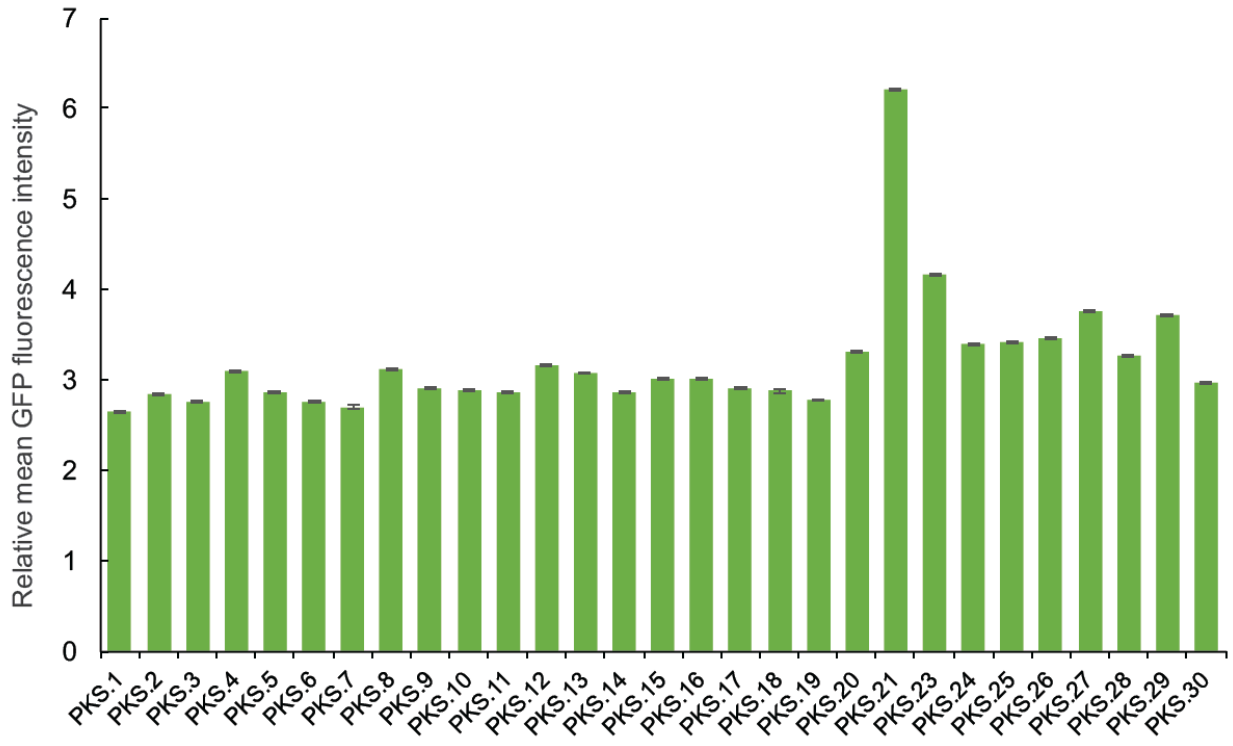


Figure D 5.3: Relative fluorescence intensity of combinatorial PKS hits. Mean fluorescence intensity of *E. coli* expressing GFP-fusions of the above hits obtained from plate screens was quantified using flow cytometry and normalized to the fluorescence of wild-type PKS expressing cells.

Table D 5.1: Literature examples of engineered biosynthetic pathways in microbes with limiting enzymes.

Product	Limiting Enzyme(s)	Publication(s)
Reticuline	norcoclaurine synthase, tyrosine hydroxylase	DeLoache et al. 2015[9]
Thebaine	salutaridine synthase	Galanie et al. 2015[10]
Synthetic Biodiesel	wax-ester synthase	Steen et al. 2010[11]
Glucaric Acid	myo-inositol oxygenase	Shiue and Prather 2014[12]
Ethyl Ester	alcohol-O-acetyltransferase	Zhu et al. 2015[13]
n-butanol, 1-butanol	butyryl-CoA dehydrogenase from Streptomyces collinus	Steen et al. 2008[14], Atsumi et al. 2008[15]
MEP (DXP) pathway	Several	Zhou et al. 2012[16]
Isobutanol	alcohol dehydrogenase	Atsumi et al. 2010[17]
Etoposide	Several	Lau and Sattely 2015[18]
Serotonin	monooxidase	Ehrenworth et al. 2015[19]
Flavonoids	cytochrome P450 reductase	Kim et al. 2009[20]

Table D 5.2: Library statistics for PKS comprehensive single-mutation libraries. NS = nonsynonymous.

	Residues 8-78	**Residues 79-156**	**Residues 235-312**	**Residues 313-388**
Number of initial library transformants	7E+04	9E+04	8E+04	1.8E+05
Reference population DNA reads post quality filter	391,403	144,646	814,066	1,605,888
Post-selection population DNA reads post quality filter	230,172	932,646	323,027 (rep1) 711,320 (rep2)	486,129
Percent of reads in reference population with:				
No NS mutations	46.4	49.2	39.4	45.4
One NS mutation	49.6	47.0	57.5	49.4
Multiple NS mutations	4.0	3.8	3.2	5.2
Coverage of possible single NS mutations:	**80.4**	**71.9**	**91.5**	**93.2**

Table D 5.3: Filtered beneficial mutations from the GFP-fusion experiment.

Wild-type residue	Position	Mutation	Stability Metric	Reference read counts	Selected read counts	Enrichment ratio	CA distance to active site (Å)	PSSM score
V	12	I	0.206	54	59	0.89	39.1	3
F	31	P	1.254	12	50	2.82	22.8	5
W	33	C	0.532	28	52	1.66	21.0	8
S	37	A	0.967	25	80	2.44	23.5	3
N	48	K	0.461	70	117	1.51	27.2	3
D	50	E	0.589	64	129	1.78	28.4	6
D	54	E	0.164	80	81	0.78	27.7	4
N	64	D	0.297	64	82	1.12	20.6	5
N	64	E	0.320	217	289	1.18	20.6	3
K	82	E	0.240	29	326	0.80	28.0	6
H	89	Y	0.157	8	77	0.58	20.1	8
N	90	M	0.323	21	274	1.02	20.5	7
F	94	L	0.220	46	498	0.75	19.4	5
V	96	A	0.246	51	580	0.82	22.3	4
I	100	M	0.169	18	177	0.61	20.3	5
V	113	A	0.497	346	6028	1.43	21.9	5
M	115	K	1.603	10	555	3.11	25.0	5
M	115	R	0.871	104	3034	2.18	25.0	3
D	118	K	0.390	51	746	1.18	29.1	5
A	121	G	0.756	49	1238	1.97	32.1	6
I	127	V	0.228	26	286	0.77	28.3	3
L	235	I	0.504	37	126	1.42	35.9	3
L	235	V	1.043	115	791	2.44	35.9	4
F	240	Y	0.216	82	170	0.71	24.0	3
T	244	S	0.666	549	2380	1.77	21.9	4
T	245	A	0.529	261	925	1.48	18.5	3
V	250	L	0.268	413	942	0.844	15.0	3
V	250	I	0.240	77	167	0.771	15.0	3
N	252	D	0.649	61	258	1.73	18.1	6
N	281	H	0.771	12	60	1.98	17.3	3
V	282	I	0.430	40	121	1.25	16.7	7
S	284	K	0.681	28	124	1.80	20.9	6
I	287	V	0.510	48	165	1.44	23.8	3
I	304	V	0.289	544	1290	0.90	15.5	3
Q	328	K	0.321	143	81	0.90	20.5	3
K	355	R	0.221	185	87	0.64	26.0	3
G	357	A	0.744	68	75	1.87	29.8	3
S	364	T	0.251	490	244	0.72	26.5	5
D	366	E	0.387	1099	699	1.07	26.4	5
V	372	C	0.660	305	299	1.70	18.6	4

Table D 5.4: Mutations included in the combinatorial PKS library. Highlighted mutations are ones that did not pass the filtering criteria.

Wild-type residue	Position	Mutation	Stability Metric	Reference read counts	Selected read counts	Enrichment ratio	CA distance to active site (Å)	PSSM score
V	12	I	0.206	54	59	0.894	39.1	3
S	37	A	0.967	25	80	2.444	23.5	3
D	50	E	0.589	64	129	1.777	28.4	6
N	64	E	0.320	217	289	1.179	20.6	3
N	64	D	0.297	64	82	1.123	20.6	5
N	90	M	0.323	21	274	1.017	20.5	7
P	106	A	0.641	66	1425	1.744	19.4	-2
M	115	R	0.871	104	3034	2.178	25.0	3
D	118	K	0.390	51	746	1.182	29.1	5
A	121	G	0.756	49	1238	1.970	32.1	6
R	135	S	0.530	40	733	1.507	8.4	1
A	143	V	0.257	97	1125	0.847	19.0	0
L	235	V	1.043	115	791	2.437	35.9	4
T	244	S	0.666	549	2380	1.771	21.9	4
T	245	A	0.529	261	925	1.480	18.5	3
V	250	L	0.268	413	942	0.844	15.0	3
V	250	I	0.240	77	167	0.771	15.0	3
V	282	I	0.430	40	121	1.251	16.7	7
S	284	K	0.681	28	124	1.801	20.9	6
I	301	V	0.565	154	576	1.558	24.4	2
S	318	E	0.358	86	52	0.998	22.0	2
G	357	A	0.744	68	75	1.865	29.8	3
D	366	E	0.387	1099	699	1.071	26.4	5

Table D 5.5: Primer sequences used in this work.

Deep sequencing inner primers	
PKS_T1R8_FWD	gttcagagttctacagtccgacgatcgttggaaaatggtcaa
PKS_T2_FWD	gttcagagttctacagtccgacgatctgtttttgacagaggaa
PKS_T3_FWD	gttcagagttctacagtccgacgatcgcctaagcccatca
PKS_T4_FWD	gttcagagttctacagtccgacgatcgaccctaagatgggc
PKS_T5_FWD	gttcagagttctacagtccgacgatccaggaggtaatgcaatt
PKS_T1_REV	ccttggcacccgagaattccagggatttttctgtaatat
PKS_T2_REV	ccttggcacccgagaattccacatcattacacgttgaac
PKS_T3_REV	ccttggcacccgagaattccagatgggcctctctag
PKS_T4_REV	ccttggcacccgagaattccactcgacttggtccac
PKS_T5R388_REV	ccttggcacccgagaattccagagaatgggcacact
blue = Illumina sequencing primer; black = gene overlap	
Combinatorial library mutant primers	
PKS_V12I	ggtcaaaaatttgggaggattcatgagagagctgaag
PKS_S37A	caacacctttccattgggttgatcaagcctcctatcctgatt
PKS_D50E	cagggttacaaatagtgagcatttggtggacctcaa
PKS_N64E	ggacctcaaagaaaaatttagacgtatctgtgagagaacaatgattagcaa
PKS_N64D	gacctcaaagaaaaatttagacgtatctgtgacagaacaatgattag
PKS_N90M	cccaatttgtgctctcacatggagccatcctttgatgtca
PKS_P106A	tcaggcaggacattttagtttcagaaatagccaaac
PKS_M115R	ttggaaaagaggctgtccttagggccattgatgaatg
PKS_D118K	ctgtccttatggccattaaggaatgggcccagcccaa
PKS_A121G	gccattgatgaatggggccagcccaaatccaaa
PKS_M115R+A121G	ggctgtccttagggccattgatgaatggggccagcccaaatcca
PKS_D118K+A121G	gaggctgtccttatggccattaaggaatggggccagcccaaat
PKS_R135S	tttagtcttttgcacaagcagtggtgttgacatgccc
PKS_A143V	ggtgttgacatgcccggtgtagattaccaattaattaagc
PKS_L235V	accctaagatgggcgtagagaggccc
PKS_T244S	atctttgagatagtctcaacggcccaaacattt
PKS_T245A	ggcccatctttgagatagtcacagcggcccaaacat
PKS_V250L	cacaacggcccaaacatttctccctaacgggg
PKS_V250I	cacaacggcccaaacatttatccctaacgggg
PKS_V282I	ggatgtaccaccaactattgcgaaaaatattgagagttgcttaa
PKS_S284K	tgtaccaccaactattgcgaaaaatgttgagaagtgcttaataaaggcttt
PKS_V282I+S284K	ccaaggatgtaccaccaactattgcgaaaaatattgagaagtgcttaataaaggcttttgaac
PKS_I301V	ggaatatcagattggaactcggtcttttggattcttcatccag
PKS_S318E	caattgtggaccaagtcgaggagacattgggcctagagcccaa
PKS_G357A	gagattagaaagaaatctgctagagaagggctgaagact
PKS_D366E	ggctgaagacttcaggagaggggctggact

Table D 5.6: Multiple reaction monitoring parameters utilized for LC-MS/MS analyses of AbPKS products.

Compound	Precursor ion > product ion (*m*/*z*)	Cone voltage (V)	Collision voltage (V)	Retention time (min)
N-methyl-Δ^1-pyrrolinium	84 > 42	34	16	1.28
Tropinone	140.1 > 98	40	22	1.44
Hygrine	142.1 > 84	28	16	1.78
4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid	186.1 > 84	28	16	1.42
4-(2-*N*-methylpyrrolidine)-3-oxobutanoic acid methyl ester	200.1 > 84	28	16	1.96
Cuscohygrine	225.2 > 84	28	16	1.50
Cuscohygroline	227.2 > 84	28	16	1.62
Telmisartan[a]	515.2 > 276.1	42	52	5.07

Data was analyzed in positive ion mode using a Waters Acquity TQ-D mass spectrometer.
[a]1 μM Telmisartan is included as an internal standard.

Table D 5.7: UPLC Mobile Phase Gradients Utilized for LC-MS/MS analyses of PKS products using a Waters Acquity TQ-D mass spectrometer.

Time (min)	Mobile phase A (%)	Mobile phase B (%)
0.00	99	1
0.50	62.5	37.5
2.00	50	50
4.00	0	100
5.00	0	100
5.01	99	1
6.00	99	1

Mobile phase A = 100 mM ammonium formate + 1% formic acid in water. Mobile phase B = 100 mM ammonium formate + 1% formic acid in 80% methanol.

REFERENCES

REFERENCES

1. Clomburg, J. M., Crumbley, A. A. & Gonzalez, R. Industrial biomanufacturing: The future of chemical production. *Science (80-.).* **355,** eaag0804 (2017).

2. Hughes, R. A. & Ellington, A. D. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb. Perspect. Biol.* **9,** a023812 (2017).

3. Billingsley, J. M., Denicola, A. B. & Tang, Y. Technology development for natural product biosynthesis in Saccharomyces cerevisiae. *Curr. Opin. Biotechnol.* **42,** 74–83 (2016).

4. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci.* **102,** 12678–12683 (2005).

5. Keasling, J. D. Synthetic biology and the development of tools for metabolic engineering. *Metab. Eng.* **14,** 189–195 (2012).

6. Yadav, V. G., Mey, M. De, Giaw, C., Kumaran, P. & Stephanopoulos, G. The future of metabolic engineering and synthetic biology: Towards a systematic practice. *Metab. Eng.* **14,** 233–241 (2012).

7. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10,** 957–963 (2013).

8. Noor, E., Flamholz, A., Liebermeister, W., Bar-Even, A. & Milo, R. A note on the kinetics of enzyme action: A decomposition that highlights thermodynamic effects. *FEBS Lett.* **587,** 2772–2777 (2013).

9. DeLoache, W. C. *et al.* An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose. *Nat. Chem. Biol.* **11,** 465–471 (2015).

10. Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F. & Smolke, C. D. Complete biosynthesis of opiods in yeast. *Science (80-.).* **349,** 1095–1100 (2015).

11. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463,** 559–562 (2010).

12. Shiue, E. & Prather, K. L. J. Improving d-glucaric acid production from myo-inositol in E. coli by increasing MIOX stability and myo-inositol transport. *Metab. Eng.* **22,** 22–31 (2014).

13. Zhu, J., Lin, J.-L., Palomec, L. & Wheeldon, I. Microbial host selection affects

intracellular localization and activity of alcohol-O-acetyltransferase. *Microb. Cell Fact.* **14,** 1–10 (2015).

14. Steen, E. J. *et al.* Metabolic engineering of Saccharomyces cerevisiae for the production of isobutanol and 3-methyl-1-butanol. *Appl. Microbiol. Biotechnol.* **98,** 9139–9147 (2008).

15. Atsumi, S. *et al.* Metabolic engineering of Escherichia coli for 1-butanol production. *Metab. Eng.* **10,** 305–311 (2008).

16. Zhou, K., Zou, R., Stephanopoulos, G. & Too, H. P. Enhancing solubility of deoxyxylulose phosphate pathway enzymes for microbial isoprenoid production. *Microb. Cell Fact.* **11,** 1–8 (2012).

17. Atsumi, S. *et al.* Engineering the isobutanol biosynthetic pathway in Escherichia coli by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Appl. Microbiol. Biotechnol.* **85,** 651–657 (2010).

18. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science (80-.).* **349,** 1224–1228 (2015).

19. Ehrenworth, A. M., Sarria, S. & Peralta-Yahya, P. Pterin-Dependent Mono-oxidation for the Microbial Synthesis of a Modified Monoterpene Indole Alkaloid. *ACS Synth. Biol.* **4,** 1295–1307 (2015).

20. Kim, D. H., Kim, B. G., Jung, N. R. & Ahn, J. H. Production of genistein from naringenin using Escherichia coli containing isoflavone synthase-cytochrome P450 reductase fusion protein. *J. Microbiol. Biotechnol.* **19,** 1612–1616 (2009).

21. Klesmith, J. R., Bacik, J., Michalczyk, R. & Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth. Biol.* **4,** 1235–1243 (2015).

22. Xie, X. *et al.* Rational Improvement of Simvastatin Synthase Solubility in Escherichia coli Leads to Higher Whole-cell Biocatalytic Activity. *Biotechnol. Bioeng.* **102,** 20–28 (2009).

23. Klesmith, J. R., Bacik, J., Wrenbeck, E. E., Michalczyk, R. & Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci.* **114,** 2265–2270 (2017).

24. Brachmann, C. B., Davies, A., Cost, G. J. & Caputo, E. Designer Deletion Strains derived from Saccharomyces cerevisiae S288C : a Useful set of Strains and Plasmids for PCR-mediated Gene Disruption and Other Applications. **132,** 115–132 (1998).

25. Deloache, W. C., Russ, Z. N. & Dueber, J. E. Towards repurposing the yeast peroxisome

for compartmentalizing heterologous metabolic pathways. *Nat. Commun.* **7,** 11152 (2016).

26. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17,** 467–73 (2007).

27. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* (2016). doi:10.1038/nmeth.4029

28. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17,** 691–695 (1999).

29. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27,** 3430–3431 (2011).

30. Kowalsky, C. A. *et al.* High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PLoS One* **10,** e0118193 (2015).

31. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8,** 15695 (2017).

32. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** (2008).

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

34. Deng, Z. *et al.* Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* **424,** 150–67 (2012).

35. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

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

37. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** gku511 (2014).

38. Thyme, S. B. *et al.* Massively parallel determination and modeling of endonuclease substrate specificity. **42,** 13839–13852 (2014).

39. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-

based deep mutational scanning. *Proc. Natl. Acad. Sci.* **112,** 7159–7164 (2015).

40. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci.* **110,** E1263–E1272 (2013).

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

42. Gong, L. I. & Bloom, J. D. Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS One* **10,** (2014).

43. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1,** 755–68 (2006).

44. Kowalsky, C. A. *et al.* Rapid Fine Conformational Epitope Mapping Using Comprehensive Mutagenesis and Deep Sequencing. *J. Biol. Chem.* **290,** 26457–26470 (2015).

45. Lee, M. E., Deloache, W. C., Cervantes, B. & Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. (2015). doi:10.1021/sb500366v

46. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression. *Mol. Cell* **63,** 337–346 (2016).

47. Altschup, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).

48. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

49. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

50. Altschul, S. F., Gertz, E. M., Agarwala, R., Scha, A. A. & Yu, Y.-K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* **37,** 815–824 (2009).

CHAPTER SIX

Summary and future work

SUMMARY AND OUTLOOK

In this thesis, deep sequencing technology was utilized in a standardized research pipeline developed by the Whitehead lab to study and engineer enzymes[1]. Deep mutational scanning, the testing of all possible single amino acid substitutions on the function of a protein, provides information rich datasets that address a variety of aims relevant to numerous fields[2]. The novelty of this dissertation is the application of this style of protein science to probe fundamental questions relating to the intricacies of enzyme function.

In **Chapter 2**, a novel comprehensive saturation mutagenesis method, Nicking Mutagenesis, was developed[3]. Analogous to popular commercial kits available for site-directed mutagenesis (Agilent's QuikChange or New England Biolabs Q5 Site-Directed Mutagenesis Kit), nicking mutagenesis conveniently requires routinely prepped dsDNA as input substrate. This solves the accessibility challenge presented with its best competing method, PFunkel, that requires a dU-ssDNA template that must be prepared from phage[4,5]. Until there is a significant decrease in the cost of DNA synthesis, methods such as nicking mutagenesis will be imperative for labs desiring to analyze comprehensive point-mutant libraries.

In **Chapter 3**, the deep sequencing technology pipeline was applied to address the question of how enzymes encode specificity through their primary sequence of amino acids[6]. Using growth-based selections that I developed, I was able to assess the effect of >6,000 single-amino acid substitutions on the function of a protein with three different substrates. Comparison of datasets between selections of multiple substrates provided an unprecedented look at the differential effects of mutations between substrates. Mutations benefiting only one substrate were spread throughout the protein sequence and structure, and did not correlate with the other selections.

The datasets obtained from deep mutational scanning of amiE provided a fortuitous opportunity to test theories on adaptive molecular evolution. Specifically, distribution model fitting of the DFE for beneficial mutations could be performed with high statistical power, as hundreds of beneficial mutations were identified. The DFE for beneficial mutations was found to be approximately exponentially distributed as predicted, however the relationship between protein biophysics – namely stability – and beneficial DFE has yet to be explored. To address this, in **Chapter 4** destabilized variants of wild-type amiE were designed using the Rosetta Design Software. I was able to successfully identify several variants that had decreased expression yields (a measure of stability) that have been shown to maintain wild-type catalytic function. Future work includes developing and performing growth-based selections and analyzing the resulting DFE in comparison to the existing datasets for wild-type amiE.

Natural product synthesis in workhorse organisms such as bacteria or yeast is an attractive technology, however plug-and-play of non-native enzymes as part of designed biosynthetic pathways often leads to poor protein expression. In **Chapter 5**, the deep sequencing pipeline was applied to test a generalizable method for improving protein expression while maintaining catalytic activity[7]. As a model system, a poorly expressing Type III PKS from *Atropa belladonna* was scanned using a high-throughput GFP-fusion folding reporter assay and resulting hits were combined to generate stabilized variants. However, two included mutations caused inactivation. Future work includes generating backcrosses of these two mutations and testing for activity. Additional future work includes reconstitution of a portion of the Tropane Alkaloids pathway from *Atropa belladonna* (from which the PKS originates) in *Saccharomyces cerevisiae*, and testing the effect stabilized PKS variants have on pathway productivity.

REFERENCES

REFERENCES

1. Kowalsky, C. A. *et al.* High-Resolution Sequence-Function Mapping of Full-Length Proteins. *PLoS One* **10,** e0118193 (2015).

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

3. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* (2016). doi:10.1038/nmeth.4029

4. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci.* **82,** 488–492 (1985).

5. Firnberg, E. & Ostermeier, M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One* **7,** e52031 (2012).

6. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8,** 15695 (2017).

7. Klesmith, J. R., Bacik, J., Wrenbeck, E. E., Michalczyk, R. & Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci.* **114,** 2265–2270 (2017).