EVALUATION OF ENZYME FUNCTION VIA HIGH-THROUGHPUT SEQUENCE TO FUNCTION MAPPING

By

Justin Ryan Klesmith

A DISSERTATION

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

Biochemistry and Molecular Biology—Doctor of Philosophy

2016

ABSTRACT

EVALUATION OF ENZYME FUNCTION VIA HIGH-THROUGHPUT SEQUENCE TO FUNCTION MAPPING

By

Justin Ryan Klesmith

The motivation of this work is to comprehend and overcome challenges in understanding and in design of optimal heterologous metabolic pathways that lead to production of biofuels and other valued biochemicals in microbial hosts. I specifically address the problem that within a designed pathway, introduced enzymes are often inefficient leading to a reduction of metabolic flux and consequently product yield. This enzymatic underperformance can be attributed to either poor catalytic fitness or poor soluble expression in the host. To help develop technologies that remedy these inefficiencies, the field of metabolic engineering was surveyed for current approaches that identify an optimal pathway variant and the limitations thereof. I identified numerous inadequacies in current isogenic and high-throughput pathway screening and optimization methods. Specifically, the amount of time and the number of unique variants tested in current methods is limiting. With the advent of high-throughput deep sequencing technologies, large population-based studies are now feasible which reduce the amount of time and increase the total number of unique variants tested. Therefore, this work set out to utilize this promising new approach to test unique enzyme variants in a pathway.

I developed a new deep sequencing approach to study the enzyme levoglucosan kinase (LGK) from *L. starkeyi* that was introduced into *E. coli*. LGK converts levoglucosan into glucose-6-phosphate which is then used for microorganism growth. A growth selection was developed such that growth on levoglucosan as a sole carbon source was dependent on an active LGK enzyme, and the change in growth was correlated to the change in enzymatic activity. This

method was able to quantify the effect of over 8,000 single point mutations on specific levoglucosan flux. The datasets were able to predict whether a beneficial mutation improved stability or catalytic efficiency. Combining computational modelling with these datasets aided the creation of nine enzyme designs. One enzyme design incorporating 38 mutations was crystallized to learn the structural basis of the beneficial mutations. The best enzyme design had a 15-fold improvement in growth rate and 24-fold improvement in pathway activity.

Developing this deep sequencing method illuminated a number of problems and opportunities: 1) growth selections are difficult to design and may not be feasible for enzymes in secondary metabolism, 2) improving the soluble expression of an enzyme is potentially an easy avenue to increase specific flux however, 3) stabilizing mutations often have small trade-offs in catalytic fitness. Therefore, the second project set out to extend the original deep sequencing method to improve soluble expression of enzymes without trading-off catalytic fitness in the absence of a growth selection. Using three solubility screens: yeast surface display, GFP fusion, and Tat export, I screened two enzymes, TEM-1 beta-lactamase and LGK. Deep sequencing was used to quantify the effect of all single point mutations on soluble enzyme production. Classifiers were developed to identify solubility-enhancing mutations from these datasets that maintain wild-type catalytic fitness with an accuracy of 90%.

The final project was a small extension of the solubility work where I developed analytical equations for converting the enrichment of a variant to a fitness metric for plate-based screens like the Tat export pathway. Using isogenic and mixed cultures I show that growth rates and survival percentages correlate for plate selections. This will help further deep sequencing-based studies for interpretation of the datasets.

I dedicate this dissertation to my parents Steve and Rebecca for their unending love and support, and for that I owe them the world. My wife Stephanie who still entertains my playful attitude even after all of these years together. My friend and undergrad advisor Dr. Tom Zamis who I credit my choice to become a biochemist.

ACKNOWLEDGEMENTS

I would like to thank my advisor and role-model, Tim Whitehead, for his support, insight, guidance, patience, and for assisting me to develop into an independent scientist with numerous publications and fellowship awards. His impact on my life far exceeds what is possible to be stated on this page or my CV. I thank my wife, Stephanie, for her patience with my dedication to my studies. I thank my family for their love and support in helping me with my life and career. I thank my lab family and life-long friends Caitlin Stein, Jim Stapleton, Emily Wrenbeck, Carolyn Haarmeyer, and Matthew Faber for allowing me to be part of their lives and me of theirs. My collaborators John-Paul Bacik and Ryszard Michalczyk for the structures of all of my proteins. My undergraduate advisees Sarah Thorwall, Vince Kelly, and Matt Smith who provided support on numerous projects and offered me the opportunity to become a mentor. Finally, I offer thanks to the USDA NIFA pre-doctoral fellowship for providing financial support in my final year.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

bla,beta-lactamase

DFE, Distribution of Fitness Effects

DMS, Deep Mutational Scanning

FACS, Fluorescence Activated Cell Sorting

LG, Levoglucosan

LGK, Levoglucosan Kinase

PC, Principal Component

PSSM, Position-Specific Scoring Matrix

RBS, Ribosome Binding Site

SSM, Single-site Saturation Mutagenesis

TAT, Twin-arginine translocation

TEM-1, TEM-1 beta-lactamase (*bla*)

YSD, Yeast Surface Display

CHAPTER 1

Introduction

Sections of the chapter were adapted from the publication “High-throughput evaluation of synthetic metabolic pathways” in *Technology* 4:9-14 by Justin R. Klesmith and Timothy A. Whitehead.

ABSTRACT

A central challenge in the field of metabolic engineering is the efficient identification of a metabolic pathway genotype that maximizes specific productivity over a robust range of process conditions. Here I review current methods for optimizing specific productivity of metabolic pathways in living cells. New tools for library generation, computational analysis of pathway sequence-flux space, and high-throughput screening and selection techniques are discussed.

INTRODUCTION

Microorganisms have the potential to produce many chemicals of use to society[1, 2]. In some cases, production from heterologous microorganisms is more sustainable than purifying the chemical from natural sources. Examples include harvesting Pacific yew trees or Chinese wormwood for taxol[3] or the anti-malarial artemisinin[4], respectively. Additionally, the ability to create renewable and sustainable biofuels and biochemicals is increasingly attractive given concerns about climate change and peak oil[5].

An organism producing a desired product may not exist, or a given strain may not be suitable for required economical processing conditions. Because of this, reconstructed pathways are often implanted into chassis microorganisms[5]. Some of these pathways include those specific for biofuels (ethanol[5], isobutanol[6], 1-butanol[7], 1,4-butanediol[8]), polymer monomers (polylactic acid[9], isoprene[10], 3-hydroxypropionic acid[11]), and pharmaceutically active ingredients (precursors for taxol[3] or opioids[12]). However, in many cases product toxicity or transport limits end titers, product recovery from aqueous fermentation broths is inefficient, or the volumetric productivity is below that required for a cost-effective process. Combined, these limitations temper the promise of sustainable replacement of the palette of petrochemicals and naturally extracted specialty chemicals currently in use by society.

In particular, the specific productivities of most engineered metabolic pathways are far below what is needed for industrial production. Some implanted pathways have limited flux because of substantial thermodynamic reversibility at key steps[13]. Additionally, pathway enzymes transplanted into heterologous hosts often have poor performance because of weak catalytic efficiency[14], poor protein solubility, or membrane targeting issues[12, 15]. Host-specific problems include cofactor accessibility[16], siphoning of pathway intermediates, intermediate

toxicity, and post-translational flux regulation of key precursors[12, 17, 18]. Furthermore, the performance of a specific engineered metabolic pathway may differ between host strains[19], media formulations[12], temperatures, and oxygen conditions[20].

A grand challenge in the field of metabolic engineering is the accurate and efficient identification of a pathway genotype that maximizes specific productivity over a robust range of process conditions. Attempts to improve specific productivity have largely focused on screening individual pathway enzymes for activity or balancing gene expression by testing libraries of elements like promoters and ribosome binding sites (RBS)[21-24]. Error-prone PCR mutagenesis of pathway enzymes has also been used to find activity-improving mutations[17, 18]. However, pathway optimization by total enumeration becomes unwieldy, as balancing activity at multiple nodes leads to a combinatorial explosion. Consider a plasmid-encoded pathway composed of a series of expression elements (e.g. promoters, ribosome binding sites (RBS), terminators) and pathway gene variants (**Figure 1.1a)**. A pathway library comprised of a single enzyme of average length[25] driven by ten alternative promoters and ten alternative RBS sequences can be covered by testing 10^2 variants. A library containing the above gene expression genotypes with all possible single non-synonymous mutations to the enzyme now contains $6x10^5$ variants. Testing the same number of variants using a two-enzyme pathway requires a theoretical coverage of $3.6x10^{11}$ variants, which is too large a sequence space to cover under most conditions. The combinatorial problem only becomes more acute with more pathway enzymes. To partially circumvent this combinatorial intractability, modular pathway design has been used to partition individual enzymatic steps into reaction groups. Then the expression of the resulting enzyme groups is balanced[12, 26-28] (**Figure 1.1b**). Alternative ways to explore sequence-flux

space include computational predictions from small training sets[29-31] (**Figure 1.1c**) or high-throughput screening or selection techniques[32-36] (**Figure 1.1d**).

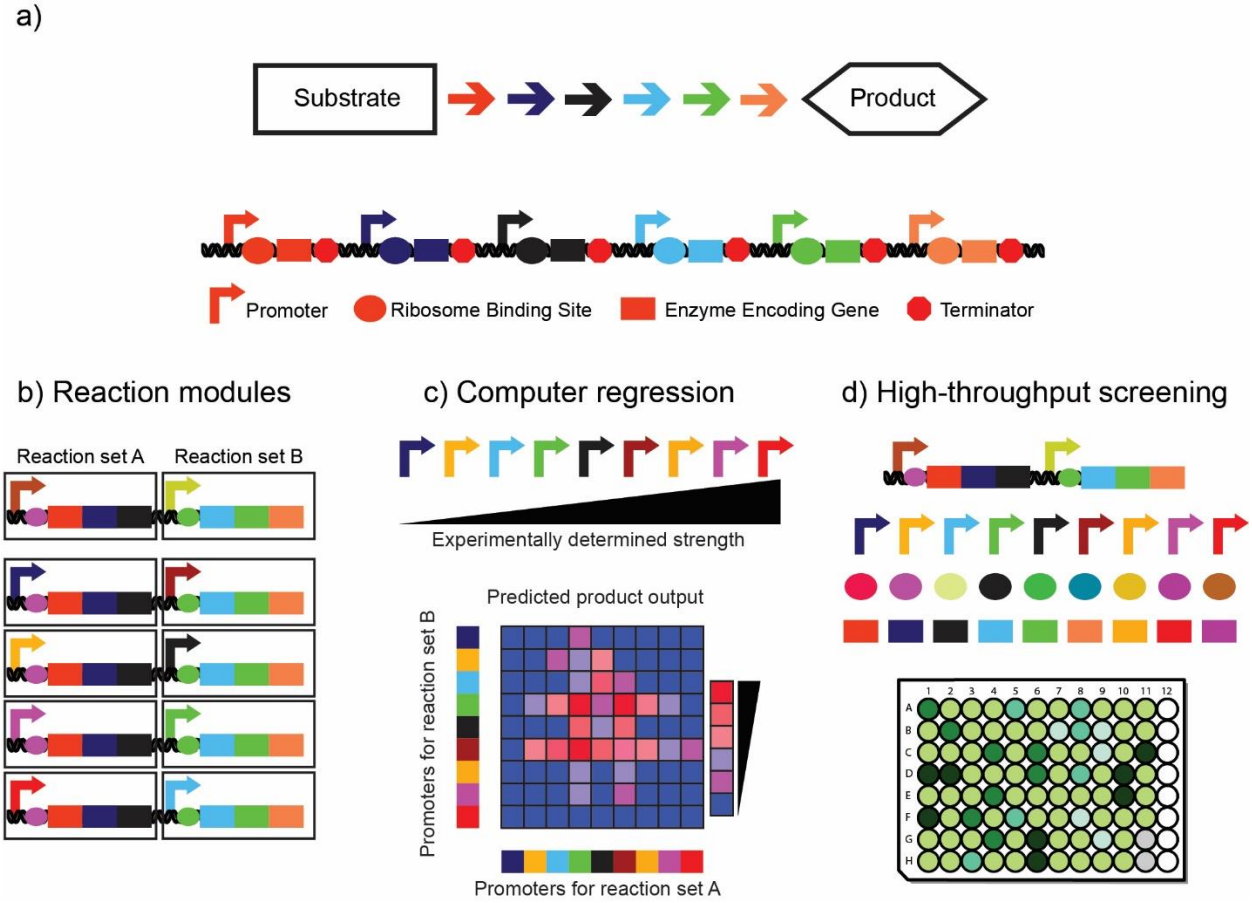


Figure 1.1: There are three main strategies to sample sequence-flux space in metabolic pathways. a) A model pathway. Each arrow represents a different reaction chemistry used to convert an initial substrate into a value-added product. Expression elements like promoters and ribosome binding sites facilitate expression of the enzyme-encoding genes. b) Individual reaction chemistries are grouped into modules. Expression of these modules is varied, reducing the combinatorial search space. c) Given a training set of different sequences with a given output, models can be harnessed to predict optimal pathway expression levels. d) Advanced DNA assembly methods can be used to create unique pathway variants that are then assessed using high-throughput screens or selections.

The focus of this introduction is on new technologies that identify highly productive and robust synthetic metabolic pathways. This introduction will not cover continuous evolution[37], whole genome engineering[38], or computational pathway design[39] – the interested reader can find

excellent reviews on some of these topics elsewhere[40-43]. I begin by describing pathway evaluation of isogenic cultures. I next describe computational approaches to predict high performing genotypes with a limited training set of pathway variants. Next, I consider high-throughput methods to assess metabolic pathways, including population-based screens or selections. New enabling techniques for DNA library construction, sequencing, and evaluation will be described throughout.

EVALUATION OF ISOGENIC CULTURES

One way to evaluate pathway variants is through the use of isogenic cultures. In a typical set-up, a combinatorial library of expression elements or enzyme variants is created, and clonal variants are tested individually. Lu *et al.* optimized a xylose fermentation pathway in *Saccharomyces cerevisiae* by shuffling promoters of various strength in front of each pathway enzyme[44]. Different promoter combinations were made and tested individually for ethanol productivity and enzymatic activity. Solomon *et al.* tested different expression levels of glucokinase (Glk) and galactose permease (GalP) to enable glucose uptake in *Escherichia coli* independent of the phosphotransferase system[45]. Carbon flux was modulated by controlling expression of Glk and GalP under control of synthetic constitutive promoters. Juminaga *et al.* constructed a pathway for L-tyrosine production in *E. coli* MG1655 by modifying plasmid copy numbers, promoter strength, gene codon usage, and placement of genes in operons[46]. The best pathway variant had a volumetric productivity of 55 mg L-tyrosine/L/hr. Ajikumar *et al.* optimized a pathway for overproduction of taxadiene, a key taxol precursor[3]. The authors used a modular approach by separating the pathway into two operons, with one encoding the methylerythritol-phosphate pathway and the other containing genes encoding the downstream

terpenoid-producing enzymes. The promoter strength in front of each operon was systematically varied and taxadiene product measured. Notably, the taxadiene production landscape was highly non-linear in response to operon expression.

Similar isogenic approaches can be used to engineer key rate-determining enzymes or transporters in implanted metabolic pathways. Zhang *et al.* used site-directed mutagenesis of active site residues of the enzymes KivD and LeuA[47]. Fermentations of *E. coli* harboring pathways with different combinations of KivD/LeuA variants were tested by quantification of desired alcohol products. Leonard *et al.* generated combinatorial mutations in the enzymes geranylgeranyl diphosphate synthase and levopimaradiene synthase to tune the selectivity and increase the productivity of levopimaradiene production in *E. coli*[48]. The best strain had a maximum volumetric productivity of 7.3 mg levopimaradiene per L per h. Lee *et al.* improved xylose utilization in *S. cerevisiae* by directed evolution of xylose isomerase[49]. After three rounds of error-prone PCR and screening they isolated a mutant with a 61-fold improvement in aerobic growth rate and an 8-fold improvement in ethanol production and xylose consumption. Screening pathway variants is not only limited to enzymes. Young et al. demonstrated the tunability of yeast sugar transporters through a combination of motif-based design and saturation mutagenesis[50]. This approach was used to identify xylose-specific fungal molecular transporters, which when expressed improved xylose utilization by *S. cerevisiae*.

COMPUTATIONAL PREDICTIONS USING EMPIRICAL TRAINING SETS

Adjusting the right balance of enzyme specific activities within a pathway is crucial as the fitness cost of protein expression[51], catabolism of pathway intermediates, and off-product reactions can all lower specific productivities. While there have been many admirable attempts to

forward engineer biological systems and parts[22, 24, 52-54] and analytical equations describing pathway flux have been formulated[55], tuning metabolic pathways is largely still an empirical exercise. Because of this, computational models have been used to predict high productivity portions of sequence-flux space given sparse flux datasets resulting from testing isogenic cultures. Lee *et al.* used a linear regression model trained on empirical data to relate enzyme expression levels to product titers in a violacein biosynthetic pathway[29]. This simple model could accurately predict promoter combinations resulting in the production of violacein or one of the three alternative products. Another approach to computationally model and improve pathway performance is to correlate targeted proteomics and metabolite data. George *et al.* generated isopentenol pathway variants with differing promoters, operon organization, and codon-usage[30]. They then used HPLC and LC-MS to quantify glucose, organic acids, and pathway intermediates and used mass spectrometry to quantify all proteins in their pathway. Spearman rank correlations were calculated from values of protein area and metabolite concentrations. Based on these relationships, individual variants were reconstructed and tested in time-course experiments to test model predictions. While this method may not capture complex regulatory interactions like feedback inhibition, other methods like ^{13}C metabolic flux analysis studies are more capable to do so[56, 57]. One example is Feng *et al.* where the authors tested different xylose reductase, xylitol dehydrogenase, and xylulose kinase variants in a yeast xylose pathway and used ^{13}C metabolic flux analysis to determine if the different cofactor requirements of the different enzyme variants had any effect on growth and ethanol production[58]. They found that production of ethanol wasn't affected by the cofactor requirements of the xylose pathway however the cofactor-balanced xylose pathway allowed growth under more conditions. Farasat et al. developed a sequence-expression-activity mapping method to find optimal expression conditions with desired activity

for a carotenoid biosynthetic pathway[31]. In a first step, an RBS calculator is used to make a library that spans a large range of protein expression space. Next, a subset of the library is tested for activity and used as a training set for a computational model. A new library is then constructed with targeted expression within a narrow window specified by the model. Zelcbuch *et al.* performed an iterative assembly of three fluorescent reporters, each with an associated RBS, into an operon[59]. This initial search reduced the expression search space for a balanced astaxanthin pathway. In a clever approach, they were able to haplotype the individual non-local RBS sequences included within the operon by sequencing a downstream barcode built using iterative restriction and ligation steps.

ENABLING DNA CONSTRUCTION METHODS

New genetic modification methods like DNA Assembler[60], Golden Gate assembly[61], Gibson cloning[62], sequence and ligase independent cloning (SLIC)[63], site-specific recombination, or versatile genetic assembly system (VEGAS)[64] enable efficient construction of pathway variants with an array of different enzymes, promoters, and RBS sequences. Smanski *et al.* utilized Gibson cloning[62] and Golden Gate assembly[61] to refactor the *Klebsiella oxytoca* nitrogen fixation gene cluster[32] by systematically varying the expression levels of individual genes in the complete 16-gene pathway. Performance of their clusters was assessed by RNA-seq for expression levels and nitrogenase activity assays. The best of the 122 full-length pathways tested resulted in recovery of 57% of the wild-type activity. Layton and Trinh used Gibson cloning to make ester fermentative pathways in *E. coli*[33]. The modular design of their pathway allowed quick replacement of RBS and promoter sequences. Oliver *et al.* improved 2,3-butanediol production in cyanobacteria by using SLIC to swap different RBS sequences in front of each

pathway enzyme[34]. Colloms *et al.* used serine integrase site-specific recombination to rank gene order and RBS sites for more efficient production of violacein and lycopene[35]. Du *et al.* used *S. cerevisiae* native homologous recombination to swap promoters of various expression strength in front of relevant genes[65]. This was used to improve xylose and cellobiose utilization pathways. Kim *et al.* used a similar approach to balance the flux of a xylose utilizing pathway for biofuel production[36]. Importantly, the optimal pathway was strongly dependent on both the host genotype but also the sugar composition of the growth medium. Latimer *et al.* combinatorially tested promoters of the eight gene pathway for xylose utilization in *Saccharomyces cerevisiae*[20]. Library plasmids were made with Golden Gate assembly. Similar to results above, they found that the enrichment of specific yeast promoters in their library after selection was dependent on the number of genes expressed, the culture media conditions, and the cofactor dependence of the enzymes.

ALTERNATIVE HIGH-THROUGHPUT SCREENING METHODS

Many of the above examples utilized medium-throughput plate-based screening or a growth based selection in order to sort variants. There have been recent developments to utilize fluorescence activated cell sorting (FACS) or microfluidic sorting technology in cases without an observable growth phenotype. For example, Wang *et al.* cultured xylose consuming strains in droplets and microfluidic sorting based on the fluorescence of oxidized extracellular metabolites [66]. Michener *et al.* utilized FACS to screen improved variants of caffeine demethylase using a designed RNA biosensor[67]. The RNA biosensor is a combination of ribozyme and aptamer located in the 3' UTR of a fluorescent reporter gene. When the aptamer is bound to a desired ligand, the ribozyme misfolds leading to lower RNA cleavage rates and increasing the

fluorescent output. Tang et al. utilized FACS to screen for *E. coli* clones with enhanced triacetic acid lactone (TAL) production using a an engineered TAL fluorescent reporter[68]. Jha et al. used a FACS screen to identify *E. coli* clones with increased enzymatic production of 3,4 dihydroxy benzoate [69]. In these above examples, the limitation is developing a fluorescent reporter that is coupled to intracellular concentrations of a target metabolite.

POPULATION-BASED MEASUREMENTS

One limitation of high-throughput screening is the inability to haplotype a unique pathway sequence to an output phenotype. Typically, only a few "winners" of the selection are sequenced. This is sub-optimal for two reasons. First, the winner variants depend strongly on the exact selection or screening conditions used, and so a selection must be repeated for each change of fermentation condition or host genotype. Secondly, high-throughput methods do not allow coverage of a complete sequence-flux space for even moderate-length pathways, and losing crucial genotypic information of the pathway makes it impossible to use the powerful computational analyses and prediction tools that have been demonstrated for low-throughput pathways. I envision population-based measurements that can more thoroughly search sequence-flux space and also identify Pareto optimal genotypes that are robust to different processing conditions (**Figure 1.2**). In this document I use the definition of Pareto optimality as the state where it is impossible to make one condition better without making other conditions worse. An example of identifying Pareto optimal genotypes is covered in Chapter 3 of this dissertation.

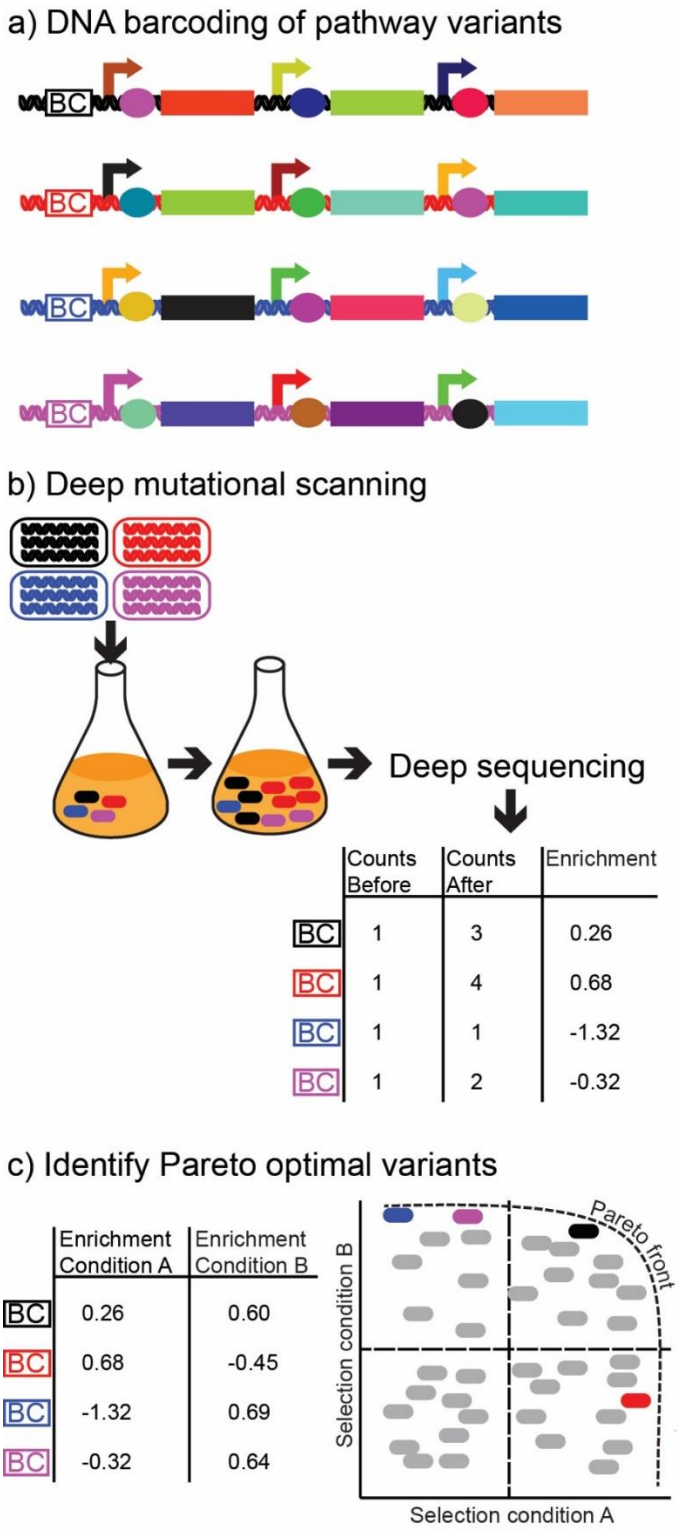


Figure 1.2: Population-based measurements of pathways enable thorough search of flux space. a) DNA barcoding methods allow long DNA constructs to be uniquely identified by short sequences. b) Deep mutational scanning quantifies the enrichment of individual DNA variants after a selection. The enrichment relating the change in frequency of an individual variant can be

Figure 1.2 (cont'd) related to specific productivity. c) Comparing the fitness of individual variants between different selection conditions allows one to find Pareto optimal pathway sequences. This enables the identification of pathways supporting high specific productivity under robust processing conditions.

Recent advances in deep sequencing technology allow the ability to track tens of thousands of pathway variants in a high throughput screen. Most of such methods rely upon "barcoding" individual cells with a short unique identifier DNA sequence (**Figure 1.2a**). A growth selection is performed, and these populations are deep sequenced at the barcode locus. The change in frequency of an individual barcode can be related to the fitness of that unique variant[70, 71]. While in principle such techniques could be used to track individual metabolic pathway variants, most demonstrations have been for studies on evolution. Smith *et al.* developed a barcode sequencing method (Bar-seq), which they validated by performing growth selections of a mixed culture containing yeast deletion strains[72]. The barcode abundance after selection was determined for each deletion strain by deep sequencing the entire population. More recently, Levy *et al.* barcoded 500,000 lineages of *Saccharomyces cerevisiae* and used a growth selection to track time-dependent changes in fitness among the population[73]. Chubiz *et al.* introduced FREQ-Seq, a method to barcode and determine allele frequencies from a mixed population[70]. FREQ-Seq was used to map seven variants of the enzyme Tet(X2), conferring tetracycline resistance, in ten different evolving populations[74].

Frequency analysis of variants within a population can be used to assess if a single variant improves, reduces, or has no effect on function. This approach has been used for evaluation of yeast translation initiation sites[75] and bacterial promoter strengths[21] by coupling these upstream elements to fluorescent reporter proteins. Subsequently, populations are sorted by FACS. In fact, massively parallel sequence-function mapping is now commonplace in

determining the sequence effects on function for proteins[76]. For example, this methodology has been used to improve the affinity of engineered protein binders to Influenza[77]. The question remains how to leverage impressive deep sequencing technology to improve implanted metabolic pathways.

In principle, high-throughput sequence-function mapping can be used to determine metabolic pathways supporting higher or lower flux, provided that it is coupled to a selectable phenotype like growth. I developed a new approach called FluxScan, detailed in Chapter 2, that maps the sequence determinants of flux in living cells (**Figure 1.2b**). First, a selection is designed to allow growth if and only if flux is routed through the implanted pathway. A mutational library is then created and transformed into the strain of interest. After a growth selection is performed for 4-10 generations, the entire population is deep sequenced and compared with the population before selection. The frequency change of each variant can be calculated and converted to a flux value. To demonstrate this method, I determined the effect of flux for over 8,000 single point mutants in a pyrolysis oil catabolic pathway[78]. One designed pathway incorporating fifteen beneficial mutations identified from FluxScan supported a 15-fold improvement in growth rate on levoglucosan, a chief pyrolysis oil constituent.

One significant technical challenge with FluxScan and related deep sequencing approaches is the inability to cover the complete length of metabolic pathways: current long read lengths of the Illumina platform are approximately 300 bp, whereas full operons can exceed 10 kb. One method to escape this limitation is to sequence small contiguous regions of sequence (a gene tile) able to fit on a single read. This "tiling" is then repeated along the length of the entire gene encoding sequence[79]. Other potential solutions to extend deep sequencing include coupling a predefined barcode sequence to a given pathway variant using clever DNA construction

approaches[32, 59] or to utilize the next generation of long-read, highly accurate haplotype sequencing technologies[80, 81].

The ideal outcome of any experiment should be to find Pareto optimal pathway sequences that are robust and transferrable to any process condition. Single isogenic culturing conditions are not suited for this task as each pathway variant would individually have to be tested under each process condition to determine the resulting phenotype. Therefore high-throughput population-based measurements are more capable to resolve the fitness of each sequence variant under each process condition (**Figure 1.2c**) provided that they are performed under diverse conditions. Sequence variants from current high-throughput genomic methods that originate from different process conditions highlight this open problem. Gall *et al.* used the SCALEs method to map the gene expression in *E. coli* that conferred an advantage in the presence of 1-naphthol[82]. They show that genes with enhanced expression depend on the type of culturing method used. Only 25% of clones that were reproducibly enriched in serial transfer cultures were similarly enriched in single batch cultures. Similarly, Warner et al. used TRMR and found differential gene expression depending on the four (valine, D-fucose, methylglyoxal, and saliciin) growth conditions tested[83]. Badarinarayana *et al.* used DNA microarrays to find genes with enhanced expression in *E. coli* when grown in Luria-Bertani or glucose minimal media[84]. They found that under poor media conditions genes that are biosynthetic for missing media nutrients were significantly enriched over cultures in rich media. While these examples highlight gene expression on the genomic scale, the problem also applies to expression and gene variants on the individual pathway level. An example of this is from the aforementioned FluxScan study where the enrichment of individual mutations from the enzyme levoglucosan kinase was strongly dependent on the biophysical properties of the starting enzyme variant from

each selection[78]. Being able to measure the pathway phenotype from different expression elements and individual gene variants under different growth conditions should allow the elucidation of pathway sequences that are optimal over a range of diverse conditions and determine why other sequences fail when these conditions change. These high-throughput population-based measurements can help train computational models by providing empirical data and similarly computational models can help reduce the sequence search space for a more targeted population-based screen.

OUTLOOK

In the near future I believe that robust, high performing pathways can be efficiently identified. New DNA assembly technologies allow for construction of large libraries of pathway variants covering a large range of protein expressions and activities. The number of unique pathway variants that can be made far exceeds that which can be accurately validated using existing technology. In this introduction I have covered current methods to reduce the search space. There is no general method that can assess any metabolic pathway, as there are limitations to each of the main approaches. There are two practical limitations that must be surmounted. First, the relationship between gene expression and pathway flux is highly non-linear. Second, a specific genotype may only support high productivity in a narrow range of process conditions. I suggest that marrying computational modeling with empirical datasets resulting from population-based measurements will allow a more efficient discovery of Pareto optimal gene encoding, expression, or regulatory sequences.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Keasling, J. D. (2008) Synthetic biology for synthetic chemistry, *ACS chemical biology 3*, 64-76.

[2] Markham, K. A., and Alper, H. S. (2014) Synthetic Biology for Specialty Chemicals, *Annual Review of Chemical and Biomolecular Engineering*.

[3] Ajikumar, P. K., Xiao, W.-H., Tyo, K. E. J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T. H., Pfeifer, B., and Stephanopoulos, G. (2010) Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli, *Science 330*, 70-74.

[4] Paddon, C. J., Westfall, P., Pitera, D., Benjamin, K., Fisher, K., McPhee, D., Leavell, M., Tai, A., Main, A., and Eng, D. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin, *Nature 496*, 528-532.

[5] Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B., and Keasling, J. D. (2012) Microbial engineering for the production of advanced biofuels, *Nature 488*, 320-328.

[6] Trinh, C. T., Li, J., Blanch, H. W., and Clark, D. S. (2011) Redesigning Escherichia coli Metabolism for Anaerobic Production of Isobutanol, *Applied and Environmental Microbiology 77*, 4894-4904.

[7] Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., Smith, K. M., Brynildsen, M. P., Chou, K. J. Y., Hanai, T., and Liao, J. C. (2008) Metabolic engineering of Escherichia coli for 1-butanol production, *Metabolic Engineering 10*, 305-311.

[8] Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., and Van Dien, S. (2011) Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol, *Nat Chem Biol 7*, 445-452.

[9] Jung, Y. K., Kim, T. Y., Park, S. J., and Lee, S. Y. (2010) Metabolic engineering of Escherichia coli for the production of polylactic acid and its copolymers, *Biotechnology and Bioengineering 105*, 161-171.

[10] Lindberg, P., Park, S., and Melis, A. (2010) Engineering a platform for photosynthetic isoprene production in cyanobacteria, using Synechocystis as the model organism, *Metabolic Engineering 12*, 70-79.

[11] Borodina, I., Kildegaard, K. R., Jensen, N. B., Blicher, T. H., Maury, J., Sherstyk, S., Schneider, K., Lamosa, P., Herrgård, M. J., Rosenstand, I., Öberg, F., Forster, J., and Nielsen, J. (2015) Establishing a synthetic pathway for high-level production of 3-hydroxypropionic acid in Saccharomyces cerevisiae via β-alanine, *Metabolic Engineering 27*, 57-64.

[12] Thodey, K., Galanie, S., and Smolke, C. D. (2014) A microbial biomanufacturing platform for natural and semisynthetic opioids, *Nat Chem Biol 10*, 837-844.

[13] Bond-Watts, B. B., Bellerose, R. J., and Chang, M. C. Y. (2011) Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways, *Nat Chem Biol 7*, 222-227.

[14] Milo, R., and Last, R. L. (2012) Achieving diversity in the face of constraints: lessons from metabolism, *Science 336*, 1663-1667.

[15] Trenchard, I. J., and Smolke, C. D. (2015) Engineering strategies for the fermentative production of plant alkaloids in yeast, *Metabolic Engineering 30*, 96-104.

[16] Avalos, J. L., Fink, G. R., and Stephanopoulos, G. (2013) Compartmentalization of metabolic pathways in yeast mitochondria improves the production of branched-chain alcohols, *Nature biotechnology 31*, 335-341.

[17] Bastian, S., Liu, X., Meyerowitz, J. T., Snow, C. D., Chen, M. M. Y., and Arnold, F. H. (2011) Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli, *Metabolic Engineering 13*, 345-352.

[18] DeLoache, W. C., Russ, Z. N., Narcross, L., Gonzales, A. M., Martin, V. J. J., and Dueber, J. E. (2015) An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose, *Nat Chem Biol advance online publication*.

[19] Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Jiang, H., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K. W., Fickes, S., Galazzo, J., Gaucher, S. P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L. F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P. S., Keasling, J. D., Reiling, K. K., Renninger, N. S., and Newman, J. D. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin, *Nature 496*, 528-532.

[20] Latimer, L. N., Lee, M. E., Medina-Cleghorn, D., Kohnz, R. A., Nomura, D. K., and Dueber, J. E. (2014) Employing a combinatorial expression approach to characterize xylose utilization in Saccharomyces cerevisiae, *Metabolic Engineering 25*, 20-29.

[21] Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli, *Proceedings of the National Academy of Sciences 110*, 14024-14029.

[22] Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q.-A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and

reliable gene expression via standard transcription and translation initiation elements, *Nat Meth 10*, 354-360.

[23] Lee, M. E., DeLoache, W. C., Cervantes, B., and Dueber, J. E. (2015) A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly, *ACS Synthetic Biology*.

[24] Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression, *Nat Biotech 27*, 946-950.

[25] Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. (2001) On the total number of genes and their length distribution in complete microbial genomes, *Trends in Genetics 17*, 425-428.

[26] Sheppard, M. J., Kunjapur, A. M., Wenck, S. J., and Prather, K. L. J. (2014) Retro-biosynthetic screening of a modular pathway design achieves selective route for microbial synthesis of 4-methyl-pentanol, *Nat Commun 5*.

[27] Tseng, H.-C., and Prather, K. L. J. (2012) Controlled biosynthesis of odd-chain fuels and chemicals via engineered modular metabolic pathways, *Proceedings of the National Academy of Sciences 109*, 17925-17930.

[28] Zhou, K., Qiao, K., Edgar, S., and Stephanopoulos, G. (2015) Distributing a metabolic pathway among a microbial consortium enhances production of natural products, *Nat Biotech 33*, 377-383.

[29] Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J., and Dueber, J. E. (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay, *Nucleic Acids Research 41*, 10668-10678.

[30] George, K. W., Chen, A., Jain, A., Batth, T. S., Baidoo, E. E. K., Wang, G., Adams, P. D., Petzold, C. J., Keasling, J. D., and Lee, T. S. (2014) Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production, *Biotechnology and Bioengineering 111*, 1648-1658.

[31] Farasat, I., Kushwaha, M., Collens, J., Easterbrook, M., Guido, M., and Salis, H. M. (2014) Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria, *Molecular Systems Biology 10*.

[32] Smanski, M. J., Bhatia, S., Zhao, D., Park, Y., B A Woodruff, L., Giannoukos, G., Ciulla, D., Busby, M., Calderon, J., Nicol, R., Gordon, D. B., Densmore, D., and Voigt, C. A. (2014) Functional optimization of gene clusters by combinatorial design and assembly, *Nat Biotech 32*, 1241-1249.

[33] Layton, D. S., and Trinh, C. T. (2014) Engineering modular ester fermentative pathways in Escherichia coli, *Metabolic Engineering 26*, 77-88.

[34] Oliver, J. W. K., Machado, I. M. P., Yoneda, H., and Atsumi, S. (2014) Combinatorial optimization of cyanobacterial 2,3-butanediol production, *Metabolic Engineering 22*, 76-82.

[35] Colloms, S. D., Merrick, C. A., Olorunniji, F. J., Stark, W. M., Smith, M. C. M., Osbourn, A., Keasling, J. D., and Rosser, S. J. (2014) Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination, *Nucleic Acids Research 42*, e23.

[36] Kim, B., Du, J., Eriksen, D. T., and Zhao, H. (2013) Combinatorial Design of a Highly Efficient Xylose-Utilizing Pathway in Saccharomyces cerevisiae for the Production of Cellulosic Biofuels, *Applied and Environmental Microbiology 79*, 931-941.

[37] Esvelt, K. M., Carlson, J. C., and Liu, D. R. (2011) A system for the continuous directed evolution of biomolecules, *Nature 472*, 499-503.

[38] Zeitoun, R. I., Garst, A. D., Degen, G. D., Pines, G., Mansell, T. J., Glebes, T. Y., Boyle, N. R., and Gill, R. T. (2015) Multiplexed tracking of combinatorial genomic mutations in engineered cell populations, *Nature biotechnology*.

[39] Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate, *Biotechnology and bioengineering 106*, 462-473.

[40] Medema, M. H., van Raaphorst, R., Takano, E., and Breitling, R. (2012) Computational tools for the synthetic design of biochemical pathways, *Nat Rev Micro 10*, 191-202.

[41] Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009) Applications of genome-scale metabolic reconstructions, *Molecular Systems Biology 5*.

[42] Woolston, B. M., Edgar, S., and Stephanopoulos, G. (2013) Metabolic Engineering: Past and Future, *Annual Review of Chemical and Biomolecular Engineering 4*, 259-288.

[43] Goldsmith, M., and Tawfik, D. S. (2012) Directed enzyme evolution: beyond the low-hanging fruit, *Current Opinion in Structural Biology 22*, 406-412.

[44] Lu, C., and Jeffries, T. (2007) Shuffling of Promoters for Multiple Genes To Optimize Xylose Fermentation in an Engineered Saccharomyces cerevisiae Strain, *Applied and Environmental Microbiology 73*, 6072-6077.

[45] Solomon, K. V., Moon, T. S., Ma, B., Sanders, T. M., and Prather, K. L. J. (2013) Tuning Primary Metabolism for Heterologous Pathway Productivity, *ACS Synthetic Biology 2*, 126-135.

[46] Juminaga, D., Baidoo, E. E. K., Redding-Johanson, A. M., Batth, T. S., Burd, H., Mukhopadhyay, A., Petzold, C. J., and Keasling, J. D. (2012) Modular Engineering of l-Tyrosine Production in Escherichia coli, *Applied and Environmental Microbiology 78*, 89-98.

[47] Zhang, K., Sawaya, M. R., Eisenberg, D. S., and Liao, J. C. (2008) Expanding metabolism for biosynthesis of nonnatural alcohols, *Proceedings of the National Academy of Sciences*.

[48] Leonard, E., Ajikumar, P. K., Thayer, K., Xiao, W.-H., Mo, J. D., Tidor, B., Stephanopoulos, G., and Prather, K. L. J. (2010) Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control, *Proceedings of the National Academy of Sciences 107*, 13654-13659.

[49] Lee, S.-M., Jellison, T., and Alper, H. S. (2012) Directed Evolution of Xylose Isomerase for Improved Xylose Catabolism and Fermentation in the Yeast Saccharomyces cerevisiae, *Applied and Environmental Microbiology 78*, 5708-5716.

[50] Young, E. M., Tong, A., Bui, H., Spofford, C., and Alper, H. S. (2014) Rewiring yeast sugar transporter preference through modifying a conserved protein motif, *Proceedings of the National Academy of Sciences 111*, 131-136.

[51] Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E., Tomek, K. J., and Whitehead, T. A. (2014) The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate, *PLoS ONE 9*, e109105.

[52] Brewster, R. C., Jones, D. L., and Phillips, R. (2012) Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli, *PLoS Comput Biol 8*, e1002811.

[53] Mutalik, V. K., Guimaraes, J. C., Cambray, G., Mai, Q.-A., Christoffersen, M. J., Martin, L., Yu, A., Lam, C., Rodriguez, C., Bennett, G., Keasling, J. D., Endy, D., and Arkin, A. P. (2013) Quantitative estimation of activity and quality for collections of functional genetic elements, *Nat Meth 10*, 347-353.

[54] Guimaraes, J. C., Rocha, M., Arkin, A. P., and Cambray, G. (2014) D-Tailor: automated analysis and design of DNA sequences, *Bioinformatics 30*, 1087-1094.

[55] Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W., and Milo, R. (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost, *Proceedings of the National Academy of Sciences 110*, 10039-10044.

[56] McAtee, A. G., Jazmin, L. J., and Young, J. D. (2015) Application of isotope labeling experiments and 13C flux analysis to enable rational pathway engineering, *Current Opinion in Biotechnology 36*, 50-56.

[57] Young, J. D. (2014) 13C metabolic flux analysis of recombinant expression hosts, *Current Opinion in Biotechnology 30*, 238-245.

[58] Feng, X., and Zhao, H. (2013) Investigating xylose metabolism in recombinant Saccharomyces cerevisiae via 13C metabolic flux analysis, *Microbial Cell Factories 12*, 114.

[59] Zelcbuch, L., Antonovsky, N., Bar-Even, A., Levin-Karp, A., Barenholz, U., Dayagi, M., Liebermeister, W., Flamholz, A., Noor, E., Amram, S., Brandis, A., Bareia, T., Yofe, I., Jubran, H., and Milo, R. (2013) Spanning high-dimensional expression space using ribosome-binding site combinatorics, *Nucleic Acids Research 41*, e98.

[60] Shao, Z., Zhao, H., and Zhao, H. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways, *Nucleic Acids Research 37*, e16.

[61] Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009) Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes, *PLoS ONE 4*, e5553.

[62] Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases, *Nat Meth 6*, 343-345.

[63] Li, M. Z., and Elledge, S. J. (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC, *Nat Meth 4*, 251-256.

[64] Mitchell, L. A., Chuang, J., Agmon, N., Khunsriraksakul, C., Phillips, N. A., Cai, Y., Truong, D. M., Veerakumar, A., Wang, Y., and Mayorga, M. (2015) Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae, *Nucleic acids research*, gkv466.

[65] Du, J., Yuan, Y., Si, T., Lian, J., and Zhao, H. (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering, *Nucleic Acids Research 40*, e142.

[66] Wang, B. L., Ghaderi, A., Zhou, H., Agresti, J., Weitz, D. A., Fink, G. R., and Stephanopoulos, G. (2014) Microfluidic high-throughput culturing of single cells for selection based on extracellular metabolite production or consumption, *Nat Biotech 32*, 473-478.

[67] Michener, J. K., and Smolke, C. D. (2012) High-throughput enzyme evolution in Saccharomyces cerevisiae using a synthetic RNA switch, *Metabolic Engineering 14*, 306-316.

[68] Tang, S.-Y., Qian, S., Akinterinwa, O., Frei, C. S., Gredell, J. A., and Cirino, P. C. (2013) Screening for Enhanced Triacetic Acid Lactone Production by Recombinant Escherichia coli Expressing a Designed Triacetic Acid Lactone Reporter, *Journal of the American Chemical Society 135*, 10099-10103.

[69] Jha, R. K., Kern, T. L., Fox, D. T., and M. Strauss, C. E. (2014) Engineering an Acinetobacter regulon for biosensing and high-throughput enzyme screening in E. coli via flow cytometry, *Nucleic Acids Research 42*, 8150-8160.

[70] Chubiz, L. M., Lee, M.-C., Delaney, N. F., and Marx, C. J. (2012) FREQ-Seq: A Rapid, Cost-Effective, Sequencing-Based Method to Determine Allele Frequencies Directly from Mixed Populations, *PLoS ONE 7*, e47959.

[71] Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015) Massively parallel single-amino-acid mutagenesis, *Nat Meth 12*, 203-206.

[72] Smith, A. M., Heisler, L. E., Mellor, J., Kaper, F., Thompson, M. J., Chee, M., Roth, F. P., Giaever, G., and Nislow, C. (2009) Quantitative phenotyping via deep barcode sequencing, *Genome Research 19*, 1836-1842.

[73] Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., and Sherlock, G. (2015) Quantitative evolutionary dynamics using high-resolution lineage tracking, *Nature 519*, 181-186.

[74] Walkiewicz, K., Benitez Cardenas, A. S., Sun, C., Bacorn, C., Saxer, G., and Shamoo, Y. (2012) Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance, *Proceedings of the National Academy of Sciences 109*, 21408-21413.

[75] Noderer, W. L., Flockhart, R. J., Bhaduri, A., Diaz de Arce, A. J., Zhang, J., Khavari, P. A., and Wang, C. L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq, *Molecular Systems Biology 10*, n/a-n/a.

[76] Fowler, D. M., and Fields, S. (2014) Deep mutational scanning: a new style of protein science, *Nat Meth 11*, 801-807.

[77] Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., and Baker, D. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing, *Nat Biotech 30*, 543-548.

[78] Klesmith, J. R., Bacik, J.-P., Michalczyk, R., and Whitehead, T. A. (2015) Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli, *ACS Synthetic Biology 4*, 1235-1243.

[79] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N., and Whitehead, T. A. (2015) High-Resolution Sequence-Function Mapping of Full-Length Proteins, *PLoS ONE 10*, e0118193.

[80] Stapleton, J. A., Kim, J., Hamilton, J. P., Wu, M., Irber, L. C., Maddamsetti, R., Briney, B., Newton, L., Burton, D. R., Brown, C. T., Chan, C., Buell, C. R., and Whitehead, T. A. (2016) Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing, *PLoS ONE 11*, e0147229.

[81] Hong, L., Hong, S., Wong, H., Aw, P., Cheng, Y., Wilm, A., de Sessions, P., Lim, S., Nagarajan, N., Hibberd, M., Quake, S., and Burkholder, W. (2014) BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads, *Genome Biology 15*, 517.

[82] Gall, S., Lynch, M. D., Sandoval, N. R., and Gill, R. T. (2008) Parallel mapping of genotypes to phenotypes contributing to overall biological fitness, *Metabolic Engineering 10*, 382-393.

[83] Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. A., and Gill, R. T. (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides, *Nat Biotech 28,* 856-862.

[84] Badarinarayana, V., Estep, P. W., Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G. M. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays, *Nat Biotech 19*, 1060-1065.

CHAPTER 2

Comprehensive sequence-flux mapping of a *Lipomyces starkeyi* levoglucosan kinase in *E. coli*

This chapter is adapted with permission from "Comprehensive sequence-flux mapping of a levoglucosan utilization pathway in E. coli" in *ACS Synthetic Biology* 4:1235-1243 by Justin R. Klesmith, John-Paul Bacik, Ryszard Michalczyk, and Timothy A. Whitehead. Copyright 2015 American Chemical Society. John-Paul Bacik and Ryszard Michalczyk carried out the protein crystallization at Los Alamos National Laboratory.

ABSTRACT

Synthetic metabolic pathways often suffer from low specific productivity, and new methods that quickly assess pathway functionality for many thousands of variants are urgently needed. Here I present an approach that enables the rapid and parallel determination of sequence effects on flux for complete gene-encoding sequences. I show that this method can be used to determine the effects of over 8,000 single point mutants of a pyrolysis oil catabolic pathway implanted in *Escherichia coli*. Experimental sequence-function datasets predicted whether fitness-enhancing mutations to the enzyme levoglucosan kinase resulted from enhanced catalytic efficiency or enzyme stability. A structure of one design incorporating 38 mutations elucidated the structural basis of high fitness mutations. One design incorporating fifteen beneficial mutations supported a 15-fold improvement in growth rate and greater than 24-fold improvement in enzyme activity relative to the starting pathway. This technique can be extended to improve a wide variety of designed pathways.

INTRODUCTION

Production of advanced fuels and chemicals from renewable carbon is a fundamental priority for future societal needs. While metabolic routes have been demonstrated for many important molecules[1], in many cases the specific productivity through these desired pathways is far below that needed for a commercially relevant process. Pathway flux can be increased experimentally using targeted genetic insertions, optimal strain selection[2], promoter or plasmid copy number variation[3], screening of homologous enzymes governing key reactions within the pathway[4], or enzyme engineering[5, 6]. There have also been several published examples of balancing gene expression in multi-gene pathways using libraries of gene-modifying expression cassettes and, in some cases, performing regression analysis for pathway optimization[7-11]. However, for many of these examples, pathway assembly and functional validation have been low throughput. Accordingly, improving activity by high-throughput screening of pathway enzyme variants is a relatively underutilized strategy[6, 12, 13], and progress on this front awaits the development of a high-throughput way to couple genotype to flux.

Recently, deep sequencing methods have been developed that can assess the fitness contribution of thousands of genotypes in a massively parallel fashion[14, 15]. Such methods involve deep sequencing of an entire population before and after a screen or selection[16]. This approach allows, for each variant in the population, quantification of its frequency change, which can be related to an underlying function (e.g. binding affinity, catalytic turnover). While such techniques have been previously applied to many different protein systems[17-19], high-throughput mapping of enzymes has been limited to those involved in antibiotic resistance[20-22]. In principle, high-throughput sequence-function mapping can be used to identify metabolic pathway variants supporting higher or lower flux, provided that it is coupled to a selectable phenotype like growth.

However, until this point deep sequencing-aided interrogation of synthetic pathways has been demonstrated only on regulatory sequences involved in transcription and translation[23, 24].

In this work I describe a new method, FluxScan, which enables the mapping of sequence determinants to pathway flux on a massive scale. I show that FluxScan can be used to identify scores of mutations that improve the specific levoglucosan (LG) consumption rate (which I refer to throughout as "flux") of a designed levoglucosan catabolic pathway in *E. coli*. Under certain pyrolysis conditions, the 1,6-anhydrosugar levoglucosan is the most abundant molecule formed[25, 26]. Efficient LG utilization via microbial fermentation is an essential component of hybrid thermochemical/biochemical approaches to deconstruct woody biomass to advanced biofuels[27, 28]. My model pathway utilizes the enzyme levoglucosan kinase (LGK) to convert LG to the glycolytic intermediate glucose-6-phosphate in an ATP-dependent reaction[29, 30]. LGK has low thermal stability and relatively poor catalytic efficiency limiting LG consumption when expressed in heterologous organisms[29-31]. Beneficial mutations in LGK identified through sequence-function datasets were combined into a single construct, which afforded at least a 24-fold improvement in enzyme activity compared to the starting construct. The growth rate of the final strain was improved by 15-fold over the starting strain in minimal growth media with 4 g/L levoglucosan.

RESULTS

A schematic of FluxScan is shown in **Figure 2.1a.** First, a selection is developed such that microbial growth occurs if and only if flux is routed through a desired pathway. Ideally, changes in pathway flux should result in concomitant changes in growth rates. The specific growth rate of the initial population should be less than that of the maximum possible growth

rate in order to allow elucidation of variants supporting increased or decreased function. Next, pathway variants are created by mutagenesis and transformed into the specific strain used in the experiment. Growth selections are performed for 6-12 generations, and samples of the starting and final populations are deep sequenced. The resulting data allow calculation of the frequency change of each member of the library after selection, which can then be related to a growth rate relative to the starting sequence[16]. The end result is a comprehensive portrait of the effect of thousands of mutants on flux.

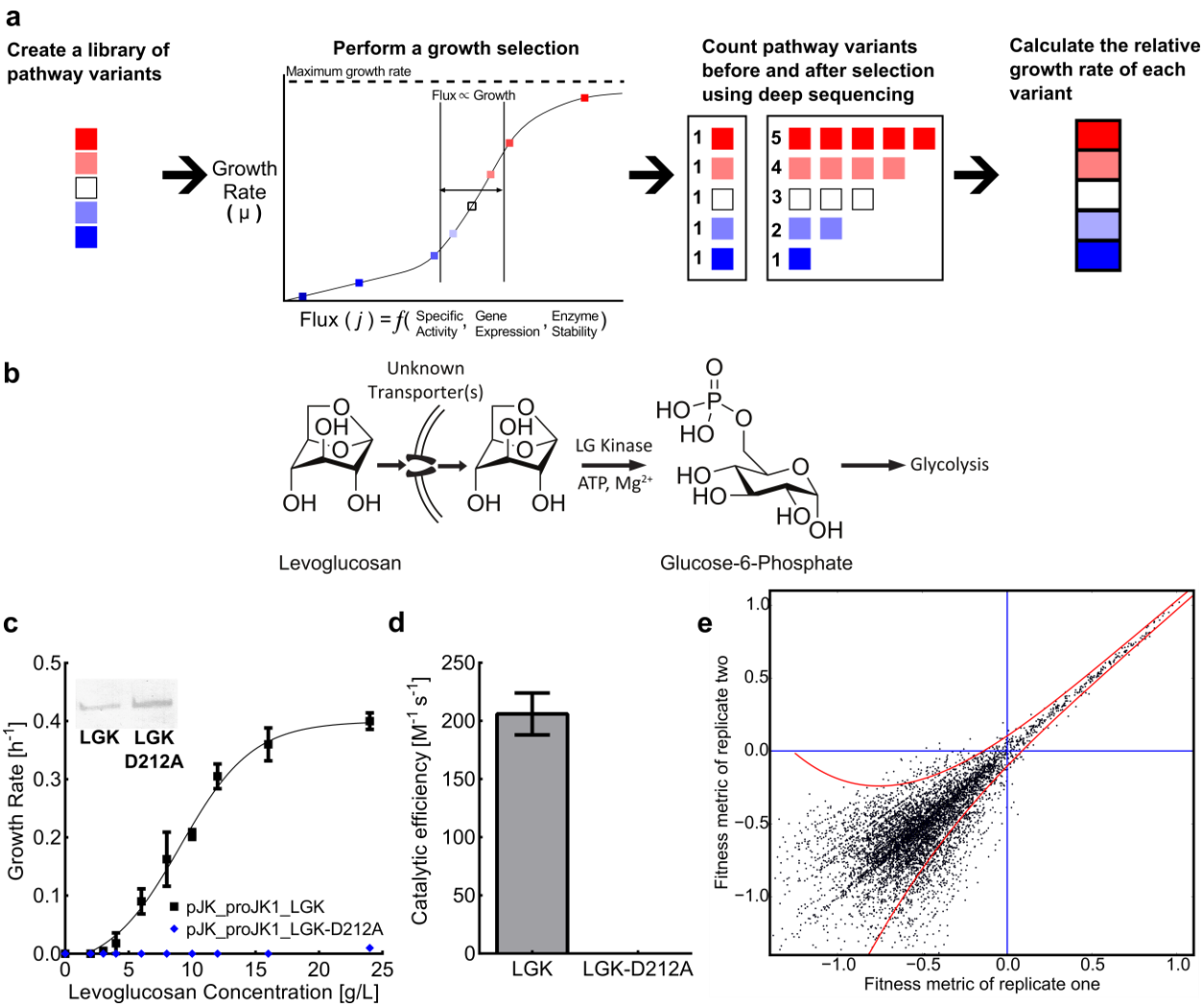


Figure 2.1: FluxScan method overview. a) Schematic of FluxScan. The first step is to create a library of metabolic pathway variants. Next, the entire population is subjected to a growth selection under conditions in which a flux increase through the metabolic pathway results in an

Figure 2.1 (cont'd) increased growth rate. The frequency change of each variant in the population after selection is determined by deep sequencing. These frequencies are used to calculate a variant-specific fitness metric, a proxy for relative growth rate. b) Levoglucosan utilization pathway. After transport across cell membrane, LG kinase converts LG to glucose-6-phosphate in an ATP-dependent reaction. c) Specific growth rates of *E. coli* Tuner pJK_proJK1_LGK (black squares) and pJK_proJK1_LGK-D212A (blue diamonds) as a function of levoglucosan concentration. Error bars represent one standard deviation of experiments repeated at least two times. A sigmoidal fit is plotted to guide the eye. (Inset) A Western blot of supernatant lysates is shown for Tuner expressing plasmids pJK_proJK1_LGK and pJK_proJK1_LGK-D212A. d) Mean catalytic efficiency (k_{cat}/K_m) of LGK and LGK-D212A measured at 30°C. Error bars represent one standard deviation ($n \geq 6$). e) Fitness metric values for biological replicates of 8,056 LGK variants. The theoretical estimation (see **Note A 2.2**) of error between replicates is plotted at two standard deviations (solid red lines).

To demonstrate the utility of the method, I aimed to improve a levoglucosan utilization pathway in *E. coli*. In this pathway, LG is actively transported inside of the cell by an unknown mechanism. Levoglucosan is found in the environment via release from wood burning in forest wild-fires[32]. Next, LGK converts LG into glucose-6-phosphate, the first intermediate in glycolysis (**Figure 2.1b**). Levoglucosan kinase from *Lipomyces starkeyi* is a 439 residue enzyme of the hexokinase family[32]. The closest natural homolog to LGK is the Gram-negative bacteria peptidoglycan recycling enzyme AnmK[33]. Both enzymes share a similar fold and sequence homology in addition to a similar mechanism for substrate hydrolysis and phosphorylation[32].

Under conditions where transport is not limiting, increasing LGK activity results in increasing growth rates. I first established a plasmid-based selection system where a strong synthetic constitutive promoter[34, 35] drives expression of a codon-optimized LGK from *Lipomyces starkeyi* (pJK_proJK1_LGK). When *E. coli* Tuner pJK_proJK1_LGK was grown in M9 salts supplemented with LG as the sole carbon source, increasing concentrations of LG increased the specific growth rate of the strain (**Figure 2.1c**). Western blots showed soluble expression of LGK (**Figure 2.1c**). *E. coli* Tuner expressing the catalytic knockout LGK-D212A[30] was unable to support growth on any concentration of LG tested (**Figure 2.1c, 2.1d**), confirming

that expression of active LGK was responsible for the LG growth phenotype. Finally, I hypothesized that in this system LG consumption rate was determined by LGK activity, not transport rate. To validate this hypothesis, I determined LGK activity using cell lysate assays and compared these with theoretical predictions necessary to support cell growth for two different growth conditions. In both cases, experimental measurements closely matched their respective theoretical predictions, confirming that growth rates are primarily limited by LGK activity rather than transport of LG into the cell (**Note A 2.1, Table A 2.1**).

I then used PFunkel mutagenesis[36] to produce a comprehensive single-site saturation mutagenesis library of LGK, transformed the library into *E. coli* Tuner cells, and grew the population for approximately ten average population doublings at 37ºC using 10 g/L LG as the sole carbon source. The entire LGK variant population was deep sequenced before and after selection, and this information was used to determine a fitness metric (a proxy for relative growth rate; see Materials and Methods) of each LGK variant via an established deep sequencing pipeline[16]. Biological replicates of the selection showed strong reproducibility, with the difference between replicates consistent with counting errors inherent in the deep sequencing quantification pipeline (**Figure 2.1e**, **Note A 2.2**). Using this method, I was able to recover the fitness of 8,056 out of 8,780 (91.8%) possible single nonsynonymous mutations in the protein encoding sequence (**Figure 2.2a, Figure A 2.1, Table A 2.2**). For additional verification, I measured the growth rates of twelve individual variants, and the rank ordering of growth rates matched those measured by the deep sequencing method (**Figure 2.2b**). LG activity measured by lysate assays increased relative to the starting sequence for all five improved variants that were tested (**Figure 2.2c**), supporting the hypothesis that pathway flux can be coupled to growth rates.

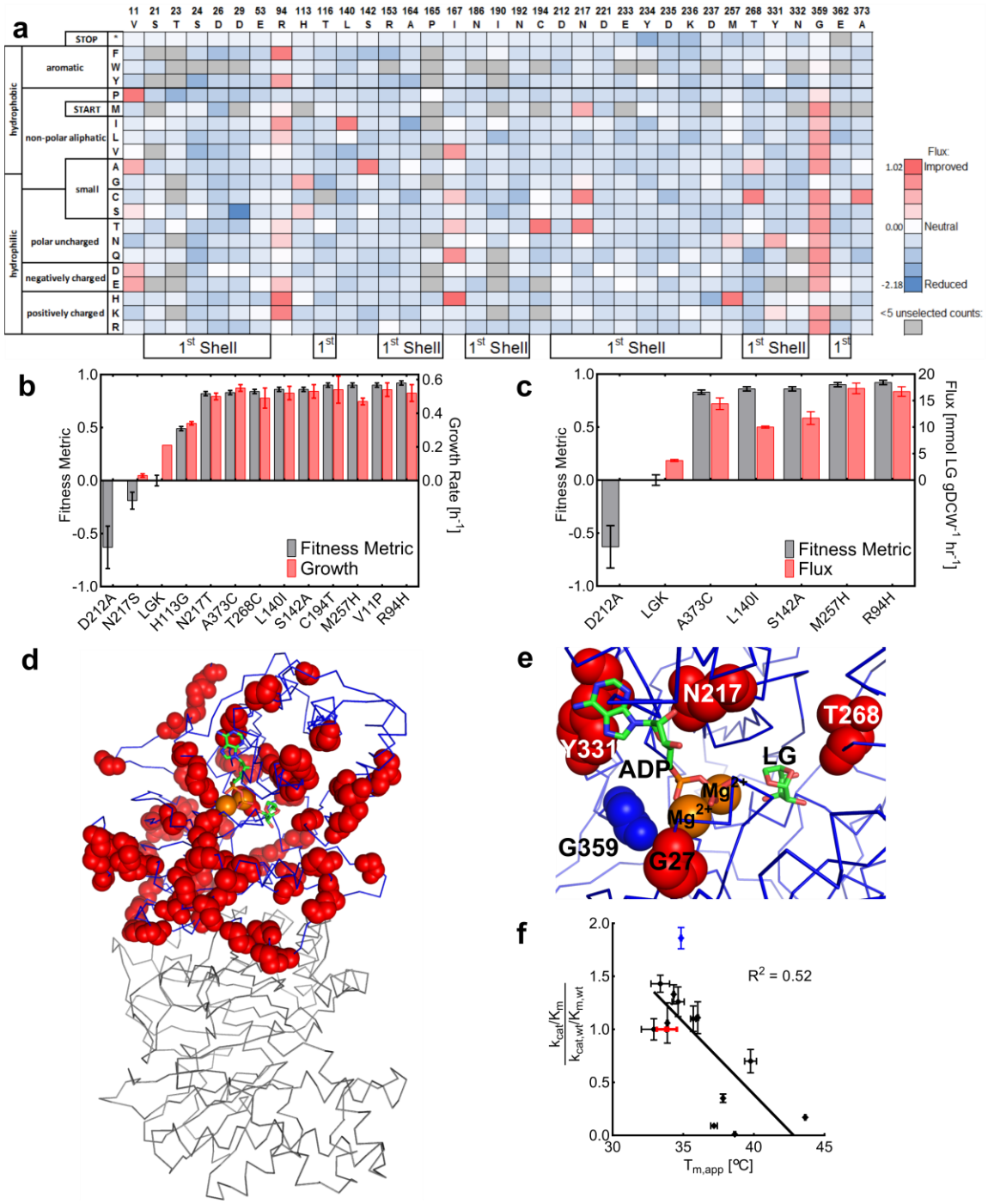


Figure 2.2: Demonstration of FluxScan. a) A fitness metric heatmap for 35/439 LGK residues in an LG utilization pathway. Columns represent residue identity while rows denote mutations. Red shading indicates an increased fitness metric, while reduced fitness is shown in blue. Gray indicates a mutation with less than five counts in the unselected library and was excluded from further analysis. 1st shell indicates residues that contact active site ligands. b) A bar graph of fitness metric (gray, left) and specific growth rate of isogenic cultures (red, right) for selected variants. Error on fitness is two s.d. and the specific growth rate is one s.d. ($n \geq 2$). c) A bar graph of fitness metric (gray, left) and LGK lysate activity of isogenic cultures (red, right) for selected variants. Error on the lysate flux is one s.d. (n=2 cultures, n=3 assay replicates). d) LGK

Figure 2.2 (cont'd) structure (PDB: 4ZLU) showing the locations of each residue with a mutation that improves relative growth rate by at least 20% (red). The ADP and LG ligands are shown as sticks, while magnesiums are shown as orange spheres. The gray line traces the other LGK subunit in the homodimer. e) View of the active site of LGK. Residues shown in red have mutations that result in increased relative growth rate. Residue Gly359 is indicated in blue. f) Relative catalytic efficiency plotted against the apparent melting temperature ($T_{m,app}$) for improved variants. The line is a linear fit of the dataset, with error bars representing one standard deviation ($n \geq 3$). LGK-His_6 is indicated in red and G359R is shown in blue.

In all, 86 (1.0%) point mutants showed improved growth rates greater than 50% relative to the starting sequence, and 215 (2.7%) were able to improve the growth rates greater than 20% (**Table A 2.2**). Analysis of the native crystal structure of LGK[30] showed that these 215 gain-of-function mutations were scattered throughout the structure (**Figure 2.2d,e**): 8 (3.7%) were located in the 1st shell of the active site (G27A, N217C/M/T, Y331K/N, and T268A/C), 34 (15.8%) were found in the 2nd shell, 33 (15.3%) were located at the homodimer interface, 26 (12.0%) were found at the surface of the protein, 93 (43.2%) were at buried positions with less than 25% accessible surface area in the apo LGK structure, and 45 were within 10 residues of the N- or C- terminus.

An individual variant can support increased pathway flux relative to the starting sequence by a number of mechanisms. For example, changes near the 5' untranslated region could change the stability of the mRNA transcript, resulting in a higher concentration of enzyme within individual cells. Alternatively, changes at the protein level could result in a more catalytically efficient enzyme, higher *in vivo* expression yields, or some combination of the two. Since LGK-His_6 has a low apparent melting temperature ($T_{m,app}$) of 33.8 ± 0.7°C (**Table 2.1**), activity may be partially limited by enzyme stability. In such a case, higher thermostability is known to result in a lower degradation rate *in vivo*[37, 38], resulting in higher concentration of active enzyme.

To understand the functional basis for the improved variants, I produced a select number of gain of function variants and tested their *in vitro* catalytic efficiency and thermostability

(**Table A 2.3**). I found that most of these variants were modestly stabilized relative to LGK-His_6, with a median increase in $T_{m,app}$ of 2.1°C (range 0-9.8°C). While no tested variant showed significant enhancement in K_m, four out of twelve tested variants showed statistically significant improvements in the turnover rate (k_{cat}) relative to LGK-His_6. In particular, a variant with the mutation G359R (**Figure 2.2e,f)**, which is adjacent to two residues coordinating an active site Mg^{2+}, had nearly double the turnover rate of LGK-His_6. Interestingly, there was a negative correlation between $T_{m,app}$ and relative catalytic efficiency (**Figure 2.2f**), suggesting competing stability/efficiency trade-offs consistent with previous observations[39-41]. However, LGK G359R showed increased specific activity without corresponding decrease in stability, showing that for some mutations there is not a trade-off (**Figure 2.2f**).

Table 2.1: Number of mutations, apparent T_m and relative catalytic efficiency of LGK designs. The amount of introduced beneficial mutations in total and between certain fitness metric values in each selection for each enzyme design. The T_m apparent measured by the thermal shift assay, error is 1 s.d. And the catalytic efficiency relative to wild-type, error is 1 s.d.

Design	Total number of mutations	Selection	Number of mutations with fitness values				$T_{m,apparent}$ [°C]	Relative Catalytic Efficiency k_{cat}/K_m [M^{-1} s^{-1}]/ $k_{cat,wt}/K_{m,wt}$ [M^{-1} s^{-1}]
		1st	≥0.15	≥0.15	≥0.15	<0.00		
		2nd	≥0.15	-0.15 to 0.15	≤-0.15	≥0.15		
LGK	0		-	-	-	-	33.8 ± 0.7	1.00
LGK.1	3		0	3	0	0	38.9 ± 0.1	1.07 ± 0.08
LGK.2	4		0	3	1	0	42.1 ± 0.1	0.46 ± 0.07
LGK.3	38		5	24	9	0	81.6 ± 0.6	0.00 ± 0.00
LGK.4	35		3	27	5	0	69.2 ± 0.8	0.00 ± 0.00
LGK.5	57[a]		2	44	6	0	65.4 ± 0.1	0.00 ± 0.00
LGK.6	23		0	3	0	20	ND[b]	ND[b]
LGK.7	18		13	5	0	0	47.7 ± 0.0	0.30 ± 0.03
LGK.8	26		8	4	0	14	33.4 ± 0.0	1.06 ± 0.37
LGK.9	15		5	4	0	6	41.0 ± 0.3	2.12 ± 0.13

[a]Five mutations in LGK.5 were not found in the second selection.

[b]LGK.6 was not determined as it was not expressed as a soluble enzyme.

In light of the above results, I reasoned that repeating the selection with a stabilized variant would uncover additional mutations that improve catalytic efficiency. Accordingly, I constructed two variants, LGK.1 and LGK.2, containing a handful of mutations distant from the active site. Both LGK.1 (L140I, S142A, A373C) and LGK.2 (V11P, M257H, T268C, A373C) had an improved $T_{m,app}$ and a catalytic efficiency close to the starting enzyme (**Table 2.1**). In the same genetic background, these variants supported much higher flux and, consequently, were able to support much higher growth rates at low initial concentrations of LG, with LGK.1 performing slightly better than LGK.2. However, LGK.1 was not able to support growth at higher concentrations of LG, possibly due to overflow metabolism or imbalances in the adenylate energy charge[42, 43] (**Table A 2.4).** Consequently, I screened weaker constitutive promoters[16] to reduce basal protein expression to a level where LGK.1 can support half-maximal growth in 4 g/L LG (**Figure 2.3a**). The decreased protein load of LGK.1 necessary to drive pathway flux resulted in a higher basal growth rate in M9 minimal media supplemented with glucose[34]. I then constructed comprehensive single site saturation mutant libraries using this improved variant as the background construct and performed a growth selection for a total of twelve average population doublings. I was able to recover the fitness of 8,312 out of 8,780 (94.7%) possible nonsynonymous mutations in the protein encoding sequence (**Table A 2.2, Figure A 2.2**), with biological replicates showing reproducibility between experimental datasets (**Figure A 2.3**).

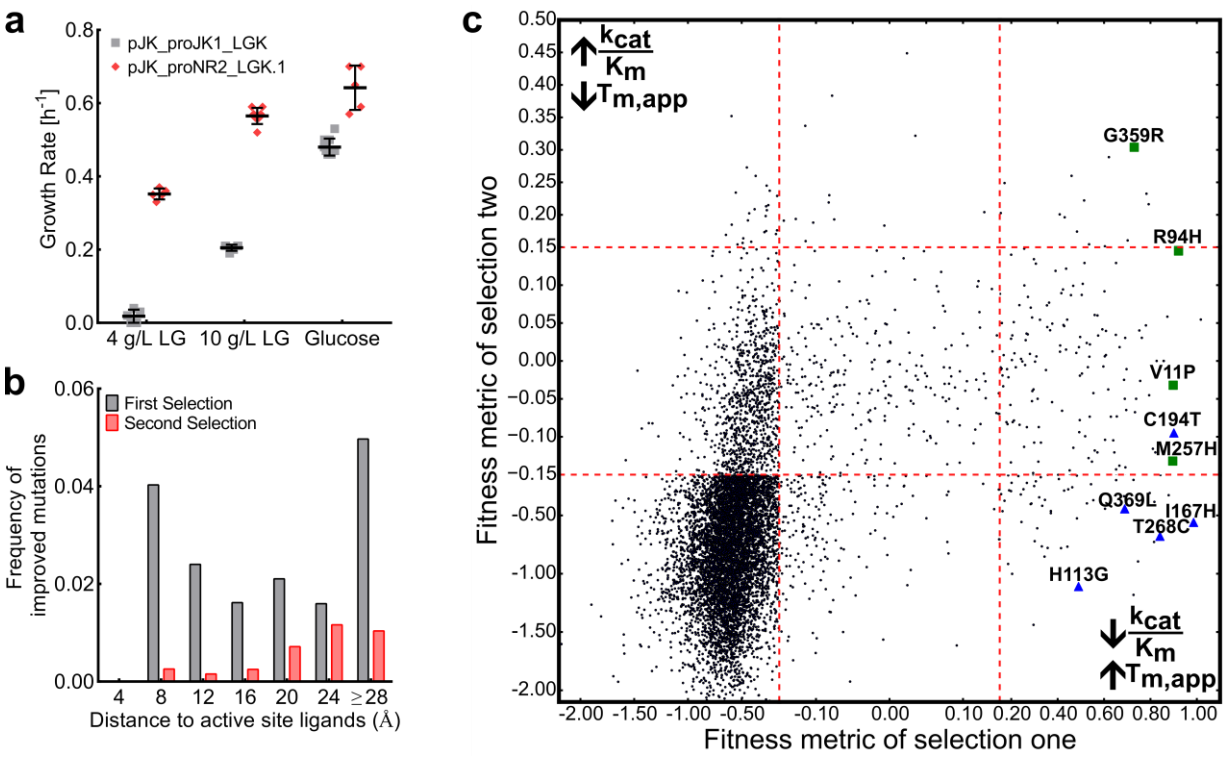


Figure 2.3: Starting with the more stable LGK.1 changes the fitness landscape. a) Comparisons of growth rates in different media for *E. coli* TUNER expressing LGK (gray squares) or the more stable LGK.1 (red diamonds). The growth rate of cultures expressing LGK.1 from the weaker promoter proNR2 was half-maximal at 4 g/L LG. The mean and standard deviation are from at least five samples taken on different days. b) The frequency of mutations that improved relative growth rates >20% is binned for each 4 Å shell as a function of distance from active site ligands in the first selection (gray) and the second selection (red). A higher frequency of beneficial mutations was found in the first selection. c) Comparison of the fitness metric between the first and second selections. The red dashed lines delineate the 20% improvement or reduction in growth rate. Tested mutants with improvements in stability but reduction in catalytic efficiency are indicated with blue triangles, whereas tested mutants with a catalytic efficiency greater than LGK are indicated with green squares.

In stark contrast to the first selection under the wild-type LGK background, only 54 (0.6%) point mutants improved growth rates greater than 20% relative to the starting sequence, and none improved growth rates by more than 50% (**Table A 2.2**). Additionally, mutations with improved fitness were depleted near the active site (**Figure 2.3b)**. In fact, only two variants with mutations near the 1st or 2nd shell (G359L/R) supported improved growth rates greater than 20%, and both were uncovered in the original selection. However, there was a significant decrease in

beneficial mutations at buried positions compared to the first selection (**Table A 2.5**), possibly because such mutations primarily enhance thermostability, which does not result in increased fitness in the more stable LGK.1 background.

I next asked to what extent mutations identified from FluxScan could be combined to improve performance of the levoglucosan pathway. Ideally, high-fitness designs with multiple pooled mutations could be constructed without exhaustively testing all possible combinations. However, some fitness-enhancing mutations result in lower catalytic efficiency, and combining many of these mutations may result in a design with lower flux, and hence fitness. Intersection of the two different selections (**Figure 2.3c**) could provide insight into the biophysical basis of individual mutations. For example, I hypothesize that mutations that are enriched in the first selection but show decreased fitness in the second selection using the more stable LGK.1 should improve thermostability at the expense of catalytic efficiency. By contrast, variants supporting higher fitness in the second selection but lower fitness in the first selection improve catalytic efficiency at the expense of thermostability. Mutations enriched in both selections improve fitness without a compensatory trade-off. In support of this hypothesis, nine variants had *in vitro* properties consistent with predictions from their respective fitness values (**Figure 2.3c**).

I constructed three designs to test the hypothesis that designs incorporating fitness-enhancing mutations from the first selection and decreased fitness in the second selection will result in thermostable, catalytically inactive proteins. LGK.3 and LGK.4 were designed by hand, and LGK.5 was designed by RosettaDesign[44] in fixed backbone mode. The designs incorporated 35-57 mutations from the starting sequence (sequences are listed in **Table A 2.6**), were expressed in *E. coli* and assayed *in vitro* as purified enzymes. Consistent with the hypothesis, all three designs showed large enhancements in $T_{m,app}$, with LGK.3 having a $T_{m,app}$ of $81.6 \pm 0.6^{o}C$

(**Table 2.1**). However, all three designs were catalytically inactive under the conditions tested. A single residue was not responsible for the loss of catalytic activity, as these designs each contain a number of unique mutations. Backcrossing experiments using LGK.3 and LGK resulted in recovery of less than 10% relative activity for a construct incorporating 17/38 possible LGK.3 mutations (**Table A 2.7**). These results suggest that the trade-off of catalytic efficiency for thermostability is gradual and the loss of catalytic activity is not the result of a single mutation.

To understand the structural basis of thermostability of the LGK designs, the crystal structure of LGK.3 (PDB: 4ZXZ) was determined to a resolution of 2.2 Å. The LGK.3 structure coordinates can be superposed to native LGK with bound ADP (PDB: 4YH5[30]) to a RMSD of 0.62Å (**Figure 2.4a**). I found that most mutations improved stability by making very modest changes to core packing. For example, P75L improves hydrophobic packing interactions with residues Leu78, Ile117, and Leu132 at the homodimer interface (**Figure 2.4b**). As another example, C194T (**Figure 2.4c**) forms a hydrogen bond with Asp171 to aid in backbone stabilization while preserving the same van der Waals interactions formed by the cysteine sulfhydryl. This stabilization is energetically important, as LGK C194T has a 6°C higher $T_{m,app}$ than LGK (**Table A 2.3**).

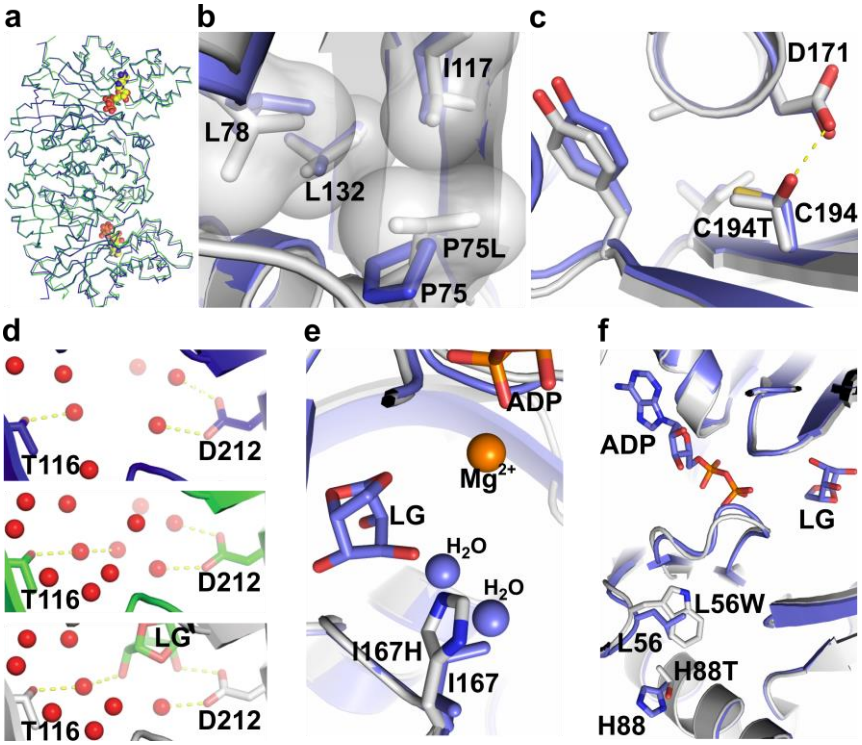


Figure 2.4: The structural basis for the improved stability and inactivity of the LGK.3 design. a) Superposition of wild-type LGK bound to ADP (green lines; PDB: 4YH5) and LGK.3 (blue lines, PDB: 4ZXZ). The RMSD between structures is 0.62Å. b) View of residue 75 in LGK (blue) and LGK.3 (grey). Mutation Pro75Leu in LGK.3 enhances hydrophobic packing with residues L78, I117, and L132 at the homodimer interface. c) View of residue 194 in LGK (blue) and LGK.3 (grey). Mutation C194T forms a new hydrogen bond with Asp171 to aid in main chain stabilization. d) Water bonding arrangements near the levoglucosan binding site. In LGK.3 (top), several of the waters appear to be displaced or are absent when compared to the wild type structure (middle, PDB: 4YH5). When levoglucosan is bound in the wild-type enzyme (bottom) two of the substrate hydroxyls take the place of the conserved water positions (PDB: 4ZLU). e) View of mutation I167H in LGK.3 (gray) and LGK (blue). The two waters near this mutation were not found in the LGK.3 structure providing further support that a disrupted water network could reduce activity. f) View of mutations L56W and H88T in LGK.3 (grey) compared to LGK (blue). The mutations appear to cause a large shift of the loop that coordinates nucleotide binding.

Next, I evaluated the LGK.3 structure to identify structural reasons responsible for LGK.3 inactivity. With the large number of mutations it is difficult to assess from the structure any single point mutant that compromises activity. Native active site residues thought to participate in catalysis are in the same rotameric conformations as in native LGK, including the catalytic aspartate, Asp212. In all of the native LGK structures the water structure is highly

conserved[30]. By contrast, for LGK.3 several of the waters appear to be displaced or are absent, although this may be partially explained by the lower resolution of this structure (**Figure 2.4d**). A I167H mutation appears to be a candidate for diminishing activity since it is very close to the LG binding site and also to Glu362, which binds magnesium (**Figure 2.4e**). Indeed, kinetic analysis of the I167H recombinant point mutant showed a reduced relative catalytic efficiency of 0.17 ± 0.01 (**Table A 2.3**) compared to the wild-type enzyme. Addition of multiple mutations also can cause structural changes that are not existent in point mutants. For example, L56W appears to cause a large main chain shift directly adjacent to the ATP binding site, which may be exacerbated by the nearby mutation of H88T, thereby affecting the enzymes ability to bind ATP (**Figure 2.4f**).

To improve the function of LGK in the catabolic pathway, I next tested four designs harboring 15-26 mutations from the starting LGK.1 sequence (**Table 2.1, Table A 2.5**). Designs were constructed by randomly choosing mutations identified from intersection of the two selections. Potential designs were modeled by Rosetta[44, 45] using the LGK structure as a template. Structures with unresolved steric clashes were discarded. In general, designs had mutations that were distributed throughout the protein but depleted near the active site (**Figure A 2.4**). LGK.6 primarily contained mutations with increased fitness in second selection but decreased fitness in the first selection. Consistent with the hypothesis that these mutations negatively impact thermostability, I was unable to solubly express LGK.6. LGK.7 contained mutations that were beneficial in both selections. I found that this design was improved in stability but showed a reduction in catalytic efficiency, possibly because some unknown combination of mutations showed negative epistasis. The final two designs (LGK.8 and LGK.9) combined mutations from LGK.6 and LGK.7. Design LGK.8 had a $T_{m,app}$ and a relative catalytic efficiency similar to wild-

type LGK, while LGK.9 showed significant improvements in stability and activity at double the catalytic efficiency of LGK (**Table 2.1**).

To test whether LGK.9 improves LG utilization by *E. coli*, I expressed LGK.9 from the same plasmid (pJK_proNR2) used for the LGK.1 selection. *E. coli* Tuner expressing LGK.9 showed at least a 24-fold improvement in flux over the starting sequence and 2.7-fold higher flux over LGK.1 at 4 g/L LG when measured by lysate activity assays (**Figure 2.5a**). The strain expressing LGK.9 showed a 15-fold improvement in growth rate over the starting sequence at 4 g/L LG. However, LGK.9 showed only a 1.3-fold improvement in growth rate over LGK.1 (**Figure 2.5b**), indicating that LG transport into the cell, rather than enzyme activity, may be limiting growth under these conditions. While future engineering efforts should focus on identifying and improving the molecular transporter(s) responsible for LG influx, I conclude that FluxScan is able to markedly improve the performance of enzymes within engineered metabolic pathways.

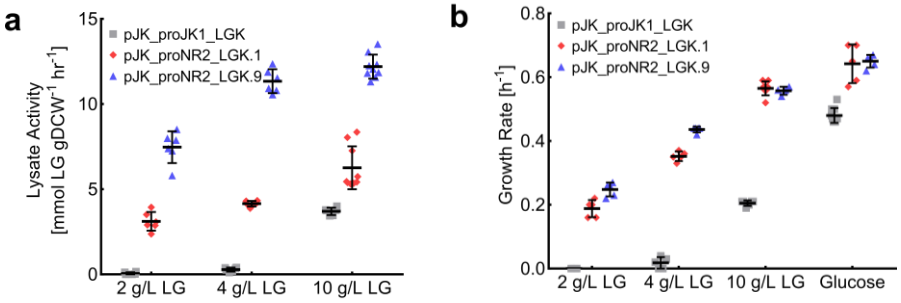


Figure 2.5: Design LGK.9 improves utilization and growth rate using LG as the sole carbon source. a) Enzyme activity assays of lysates of *E. coli* Tuner strains expressing different LGK variants. Strains were grown on M9 salts with the specified LG concentration as the sole carbon source. LGK.9 had greater than a 24-fold improvement of lysate activity compared to the starting construct at 4 g/L LG ($n \geq 2$ cultures, $n = 3$ assays per culture). b) Specific growth rates for *E. coli* Tuner strains expressing the three designs. Cultures were grown aerobically at 37°C in M9 salts with the specified LG concentration as the sole carbon source. pJK_proNR2_LGK.9 supported a 15-fold improvement of the specific growth rate over the starting variant at 4 g/L LG. The bars represent the mean and the error bars SD of cultures grown on different days ($n \geq 4$).

DISCUSSION

In this work I have demonstrated a new method that can be utilized to map the sequence determinants to flux through metabolic pathways in living cells. I have shown that it is possible to determine the flux of thousands of pathway variants in a single experiment and that mutations resulting in flux enhancements can be rationalized from biophysical considerations of the pathway enzyme. Further, I show that gain-of-function mutations can be combined in a rational manner to create enzymes with improved stability and/or catalytic efficiencies. Using mutations identified by FluxScan, I engineered an improved levoglucosan catabolic pathway allowing a 15-fold improvement in growth compared to the starting synthetic pathway. This engineered enzyme should increase the efficiency of utilization of anhydrosugars produced from thermochemical processing of renewable biomass.

FluxScan is advantageous because it allows short selection times (on the order of one day), and it uses liquid cultures as the selection medium. Liquid culture selections facilitate over-sequencing to generate fewer false positives than plate-based selections. This over-sequencing provides resolution of variants with a 10% change in relative growth rate. Whereas current metabolic engineering techniques to enhance activity are focused largely on improving expression elements, FluxScan can be used to directly improve the donated pathway enzymes. Compared to recombineering or continuous evolution selections that result in a handful of winners relevant only under the specific selection conditions (strain type, expression level, media formulations), the fine resolving power of FluxScan allows identification of scores of mutations that are able to improve pathway specific productivity. These datasets can then be coupled to other approaches such as structure-guided rational design to narrow the search space for phenotypic improvements.

In the demonstrated implementation, I identified the relative growth rate for nearly all possible single point mutants in an enzyme. I demonstrated that these single point mutants could be combined to generate better performing designs. I tested five designs that were expected to improve flux. Of these, four showed improved thermostability relative to LGK, while only one had significantly higher catalytic efficiency. Additional work will further our understanding of factors that can contribute to the design of more stable and efficient enzymes using the sequence-flux relationships generated by the method.

While I have demonstrated the power of FluxScan to improve a LG utilization pathway, there are several limitations of the current method. For example, even in the simple one-enzyme system developed in this work there was not a monotonic relationship between growth rate and flux. The LGK.1 design, when placed under the same strong promoter as LGK, was unable to support growth. This effect is attributed to imbalances in adenylate energy charge at the beginning of glycolysis[42, 43]**.** Next, the method is well suited for improving flux when the rate-determining step is known. In the last sets of designs tested, transport limitations become flux-determining as further improvements to the enzyme did not result in corresponding improvements in growth rate using LG as the sole carbon source. The requirement for a growth-based selection currently precludes transferability to fermentative pathways. Finally, a truly transformative approach would be to modulate several rate-determining enzymes and transporters at a single time. However, technical challenges of resolving mutations over kb-length distances by deep sequencing currently limits approaches to, at most, two enzymes.

Future work to improve the utility of FluxScan should focus on extending the method to cover fermentative pathways and to demonstrate modulation of multiple pathway enzymes in a single experiment. With regards to fermentative pathways, one approach to couple flux to growth

is by introducing genetic deletions into a host strain such that an organism grows anaerobically only if flux is routed through the desired fermentative pathway[46, 47]. Additionally, such a method can easily be adapted to couple pathway activity to fluorescence[48], with cells screened using fluorescence activated cell sorting or other approaches. A primary limitation of the methodology is that next generation sequencing read lengths limit interrogation of sequence changes to several hundred contiguous nucleotides. Longer, more complicated pathways await development of sequencing technologies able to map long DNA stretches with unprecedented accuracy[49, 50]. I anticipate such improvements in deep sequencing capability, allowing the use of FluxScan for resolving haplotypes of coupled mutations spread over complete synthetic operons.

MATERIALS AND METHODS

Reagents

All chemicals were purchased from Sigma-Aldrich, unless otherwise noted. All DNA primers were ordered from IDT. Genetic constructs were sequence verified by Genewiz. Selected plasmid constructs have been deposited in the AddGene plasmid repository (www.addgene.org).

Plasmid construction and verification

The plasmid pJK_proJK1_LGK was created by inserting a codon-optimized gene encoding levoglucosan kinase (LGK) (Genscript, Piscataway, NJ) with LEHHHHHH as the C-terminal tag into a pJK-series plasmid using the *NdeI*/*XhoI* restriction sites. The pJK plasmid was created by modifying the promoter and antibiotic resistance gene of pET-29b(+) (Novagen). A variant of the proB promoter sequence[35] was ordered as a gBlock (IDT) and cloned into pET-29b(+) between the *BglI* and *XbaI* restriction sites using standard techniques. On pET-29b(+) the

lacI gene, *lacO* gene, and the T7 promoter were removed between these restriction sites. The antibiotic resistance gene on pET-29b(+) was swapped to TEM-1 BLA (AmpR) from pET-22b(+) (Novagen) using Gibson assembly[51]. The ribosome binding site (sequence AGGAG), pMB1 ori, and the T7 terminator were not modified during the creation of the base plasmid. The plasmid pET29b_LGK-His_{6x} was created by subcloning LGK-His_{6x} into unmodified pET-29b (Novagen) from pJK_proJK1_LGK by *NdeI*/*XhoI* digestion. Individual point mutants were created using Kunkel mutagenesis[52]. The proJK1 promoter sequence is -35:TTTATG and -10:TATAAT, and the proNR2 promoter sequence is -35:CTTACG and -10:TAATAT. LGK designs were constructed from gBlocks (IDT) and cloned into plasmids using Gibson assembly[51]. All protein and nucleic acid sequences of LGK variants tested are listed in **Note A 2.3**.

Comprehensive single-site mutagenesis library preparation

Comprehensive single-site saturation mutagenesis was performed on the LGK protein encoding region of plasmids pJK_proJK1_LGK-WT and pJK_proNR2_LGK.1 using PFunkel mutagenesis[36] essentially as described in Kowalsky et al.[16]. Next, 10 ng of library plasmid DNA was transformed into electrocompetent *E. Coli* Tuner and plated overnight on Nalgene BioAssay plates (245mm X 245mm X 25 mm) with LB Agar and carbenicillin (Sigma-Aldrich). Controls were run to limit double plasmid transformation artifacts[16]. The next day, cells were scraped and used to inoculate a 50 mL TB culture with carbenicillin at an initial OD_{600} of 0.01. After 7 hours of growth at 37°C and 250 rpm this culture was pelleted and washed with 1X M9 salts solution (47.6 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.54 mM NaCl, 18.68 mM NH_4Cl, pH 7.0) three times. These washed cells were used to inoculate a 50 mL culture of M9 minimal media with 4 g/L glucose with carbenicillin at a starting OD_{600} of 0.05. After 16 hours of growth at 37°C at 250

rpm, the cells were washed and resuspended in fresh 1X M9 salts solution at an OD_{600} of 0.45. DMSO was then added to the cell suspension at a final concentration of 7% (v/v), and 1 mL aliquots were flash frozen with liquid nitrogen.

Growth selections

Each frozen cell stock was thawed on ice for 45 minutes and then centrifuged for 5 minutes at 10,000xg. The storage media was aspirated, and the stock was re-centrifuged for 5 minutes at 10,000xg. Pellets were washed three times with 1 ml of 1x M9 salts solution and resuspended to an OD_{600} of 0.03 in M9 minimal media + 4 g/L glucose + carbenicillin. This culture was grown for 17.5 hours at 37°C and 250 rpm. Cultures were then pelleted at 6000xg for 5 minutes. The growth media was aspirated and the cell pellet was washed three times with 1 mL of 1X M9 salts solution. Cells were resuspended in 1X M9 salts solution and used to inoculate at M9 minimal media with carbenicillin and levoglucosan (10 g/L for the initial selection on LGK and 4 g/L for LGK.1) at an initial OD_{600} of 0.02. Unused cell pellets were stored at -80°C for comparison as the unselected population. Cultures were grown for 4.5-5 generations, pelleted and washed with 1X M9 salts solution, and used to inoculate another Hungate tube (ChemGlass) with fresh media for another 4.5 to 5 of generations of growth. At the end of the selection the cells were pelleted at 10,000xg for 5 min and media removed by pipette. Cell pellets were stored at -80°C for sequencing as the selected population.

Deep sequencing analysis

Libraries were prepared for deep sequencing according to Kowalsky et al.[16] using the primers sets listed in **Table A 2.8**. Deep sequencing was performed on an Illumina MiSeq in 150

bp PE and 300 bp PE reads. Enrich[53] was used to process the deep sequencing files to determine the counts and enrichment of each mutation. The fitness metric for a variant i (ζ_I) is defined as the binary logarithm of growth rate of the variant i relative to the growth rate of the starting sequence[16]:

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

where ε_i is the enrichment ratio of the variant, ε_{wt} is the enrichment ratio of the starting sequence, and g_p is the average number of population doublings in the selection. Enrichment and fitness metric error were then calculated using equations in **Note A 2.2**. Custom python scripts used to calculate the fitness metric and statistics are at Github [user: JKlesmith] (www.github.com). The specific command lines used and description of flags are in **Note A 2.4**. The full deep sequencing datasets are provided at figshare (www.figshare.com).

Biochemical characterization

LGK variants cloned in pET-29(+) were expressed in BL21*(DE3) using Studier auto-induction[54] at 37ºC for eight hours then 18ºC overnight. Cell pellets were resuspended in 50 mM HEPES pH 7.6 buffer and were sonicated with a 120 W, 20 kHz FB120 sonicator (Fisher Scientific) with a 1/4" sonicator horn using the settings: 2:30 m total on time, cycled for 30 s on, 30 s off, 37% amplitude. Sonicated cell lysate was applied to a buffer equilibrated Ni-NTA agarose column and subsequently washed and eluted from the column. LGK was desalted using PD-10 columns (GE Healthcare). Purified protein was quantified using the Synergy H1 spectrophotometer by measuring A_{280} over the calculated extinction coefficient of 39,880 M^{-1} cm^{-1} (ExPASy ProtParam).

A coupled glucose-6-phosphate dehydrogenase (G6PD) assay was used to determine the catalytic parameters of the recombinant LGK[55]. The final concentration of each component in a total of 100 µL is 55 mM HEPES pH 7.6, 99 mM NaCl, 1.5 mM NAD+, 2 mM ATP, 20 mM MgCl2, and 0.8 units of recombinant glucose-6-phosphate dehydrogenase from *Leuconostoc mesenteroides* suspended in water (Sigma-Aldrich). Levoglucosan (Carbosynth, UK) was added to each assay well in a 1:2 serial dilution of final concentrations from 550 mM to 17.2 mM. Assay components were sealed and incubated on a 30°C pebble bath for 5 minutes prior addition of 10 µl purified LGK enzyme for a final enzyme concentration of 0.1 µM. Absorbance was monitored at A_{340} by a Synergy H1 spectrophotometer using a kinetic read method every 21 seconds at 30ºC for 20 minutes. Gen5 software was used to correct the pathlength to 1 cm, subtract blank assay wells, and divide by the extinction coefficient of NADH of (6,220 M^{-1} cm^{-1}). The time versus NADH concentration data was exported and the velocity was calculated from the initial slope until a NADH concentration of 0.3 mM or 6 minutes. Prism 6 (GraphPad) was used to non-linearly fit substrate velocity versus levoglucosan concentration using the Michaelis-Menten equation to calculate K_m and V_{max}.

Apparent melting temperatures of protein variants were assessed using a modified SYPRO Orange thermal-shift assay[56, 57]. 45 µL of 5 µM purified enzyme was added to 5 µL of 200x SYPRO Orange (Life Technologies) in 50 mM pH 7.6 sodium phosphate buffer for a total volume of 50 µL. A Bio-Rad CFX96 Real-Time PCR (Bio-Rad) measured fluorescence (ex: 470 nm, em: 570 nm) while the temperature increased from 25°C to 99°C at a gradient of 0.5ºC per 30 s. A Boltzmann sigmoidal fit was used to determine the $T_{m,app}$[56]. All samples were tested in triplicate.

Growth rate and lysate flux measurements of clonal variants

Cells were prepared as above, except that cells were freshly transformed. 0.05 mL of the washed cells were added to 2.95 mL of M9 media with carbenicillin and levoglucosan concentrations ranging from 0 g/L to 24 g/L in 14 mm inner diameter Hungate tubes. Biological replicates were grown aerobically at 37°C at 250 rpm in an I26 Shaker (New Brunswick). Optical density measurements were taken using a Genesys 20 Spectrophotometer (Thermo Fisher Scientific) at approximately 1-hour time intervals until the culture reached an OD_{600} of 0.6. Specific growth rates were calculated by taking the slope of the natural log transformed OD_{600} readings during exponential growth.

LGK activity of cell lysates was assayed using cell lysate preparation procedures adapted from Bienick et al.[34]. Strains were grown aerobically at 37°C at 250 rpm in M9 minimal media with levoglucosan and carbenicillin. Cultures were sampled at an OD_{600} between 0.15 and 0.3 and then were centrifuged at 10,000 xg for 5 min and washed with PBS twice then resuspended in 50 mM HEPES pH 7.6 and 90 mM NaCl. Cells were lysed with a 120 W, 20 kHz FB120 sonicator (Thermo Fisher Scientific) with a 1/8″ sonicator horn using the settings: 39 s total on time, cycled 3 s on, 15 s off at 37% amplitude. The lysate was clarified at 10,000 xg for 5 minutes in a microcentrifuge. After lysing the cell culture, 10 μL of cell lysate was added to the assay to measure the turnover of LG by LGK within the lysate. This rate was then normalized to gDCW using the M9 media OD_{600} to gDCW conversion factor of 0.56 (gDCW/L)-cm/OD_{600}[34].

Crystallization, data collection and structure determination

pET-29b_LGK.3-His_{6x} was transformed in to *E. coli* strain BL21(DE3) GOLD cells (Invitrogen). A 5 ml overnight culture of recombinant *E. coli* BL21(DE3) GOLD cells was added

to 500 mL of Overnight Express Instant LB Medium (Novagen) supplemented with 1% v/v glycerol and 35 μg/mL kanamycin. The culture was then incubated for 20 hours at 30°C with shaking (300 rpm). Cells were pelleted by centrifugation and stored at -80°C. Pellets were thawed in 20 mL of ice-cold lysis buffer (0.5 M NaCl, 20 mM Tris-HCl pH 7.5, 2 mM imidazole) and lysed using sonication. The lysate was clarified by centrifugation and mixed with 2 mL of TALON metal affinity resin (Clontech) with gentle shaking for 30 minutes at room temperature. The talon beads were centrifuged and re-suspended in binding buffer (500 mM NaCl, 20 mM Tris pH 7.5, 0.5 mM TCEP) before being poured into a 20 ml gravity column. The column was washed with 20 mL of binding buffer supplemented with 5 mM imidazole, followed by 20 mL of binding buffer supplemented with 10 mM imidazole. The LGK.3 protein was eluted from the column with 10 mL of binding buffer supplemented with 250 mM imidazole. The protein was further purified by Superdex 200 gel filtration column in crystallization buffer (20 mM Tris pH 7.5, 50 mM NaCl, 0.5 mM TCEP) prior to concentration using an Amicon Ultra-15 concentrator with a 10,000 Da cut-off (Millipore). Chromatographic steps were performed using an AKTA FPLC (GE Healthcare).

LGK.3 crystals were grown at room temperature using the hanging drop vapor-diffusion method by mixing 3 μl volumes of reservoir buffer containing 18% PEG3350 and 500 mM ammonium tartrate and 3 μl LGK.3 (21.5 mg/mL) in crystallization buffer. The resulting crystal was mounted in a glass capillary and a room temperature X-ray diffraction data were collected using a Rigaku X-ray diffractometer. The data were integrated using MOSFLM[58] and scaled and merged using SCALA[59]. Phase estimates for the structure were obtained by molecular replacement using PHASER[60] and the native LGK structure as a search model (PDB identifier: 4YH5). The model was subsequently rebuilt using the PHENIX autobuild routine[61]. Further

iterative model building and refinement of the protein structures were performed using Coot and PHENIX[62]. The stereochemical quality of the final models was assessed using MolProbity[61]. Refinement statistics are presented in **Table A 2.9**.

Computational design using RosettaDesign

RosettaDesign[44] was used on a prepacked[45] PDB in fixed backbone mode to pick the optimal residue from the set of mutations that showed improvement. A resfile was created such that the default was NATAA (default behavior to allow only the natural amino acid) and residues with beneficial mutations were then mutated to the best identity using PIKAA (allowing mutations to specified amino acids) to either the wild-type identity or any of the beneficial mutations found at that position.

APPENDIX

APPENDIX

Note A 2.1: Determination of transport limitations in LG system.

Consider growth of a microorganism using levoglucosan as the sole carbon source [LG]. Extracellular LG is transported inside of the cell at a flux $J_{transport}$, where the enzyme levoglucosan kinase (LGK) acts on intracellular LG at a flux J_{LGK}. The resulting product glucose-6-phosphate (G6P) is fixed into biomass at a flux $J_{Biomass}$:

$$[LG]_{ext} \xrightarrow{Jtransport} [LG]_{int} \xrightarrow{JLGK} [\text{G6P}] \xrightarrow{JBiomass} \text{Biomass} \quad \textbf{(1)}$$

Under the condition of balanced growth:

$$J_{transport} = J_{LGK} = J_{Biomass} \quad \textbf{(2)}$$

Biomass flux can be experimentally determined by the molar yield coefficient of biomass ($Y'_{X|LG}$), the molecular weight of biomass (MW_B), and the exponential growth rate (μ):

$$J_{Biomass} = \frac{\mu}{Y'_{X|LG} MW_B} \quad \textbf{(3)}$$

Assuming *E. coli* aerobic growth and that the yield coefficient for growth on LG is the same as glucose[63] results in the equation:

$$J_{Biomass} = 16.1\ \mu\ [\frac{mmol\ LG}{g\ DCW-hr}] \quad \textbf{(4)}$$

Where growth rate is expressed in units of inverse hours.

With the simplifying assumption that LG concentration is much lower than the Michaelis constant, we can relate LGK flux to Michaelis-Menten kinetics (V_{max}, K_M),:

$$J_{LGK} = \frac{V_{max}}{K_M} [LG]_{int} \quad \textbf{(5)}$$

Which can be rewritten as:

$$J_{LGK} = \left\{\frac{V_{max}}{K_M} [LG]_{ext}\right\} \left\{\frac{[LG]_{int}}{[LG]_{ext}}\right\} \quad \textbf{(6)}$$

Or, alternatively:

$$J_{LGK} = \left\{J_{LGK,measured}\right\} \left\{\frac{[LG]_{int}}{[LG]_{ext}}\right\} \quad \textbf{(7)}$$

Here, $J_{LGK,measured}$ is the flux that can be experimentally determined from lysate assays. We can combine equations **(2)**, **(4)**, and **(7)** to arrive at the desired result:

Note A 2.1 (cont'd)

$$\left\{\frac{[LG]_{int}}{[LG]_{ext}}\right\} = \frac{16.1\mu}{J_{LGK,measured}} \quad \textbf{(8)}$$

Measuring the growth rate of a culture and the specific reactant turnover in the lysate assay can determine the degree of transport limitation in the selection system: ideally, if the system was reaction limited, the ratio of internal to external [LG] should approach unity. Severe transport limitations to growth would occur if the measured LGK flux is much greater than the flux necessary to support aerobic growth.

Note A 2.2: Error approximation for fitness values calculated from digital counting.

An important consideration in deep sequencing of biomolecular libraries is the effect of counting errors on the quantification of individual variant frequencies. In this note we quantify the expected variance of the measured enrichment ratio and fitness metric as functions of depth of coverage of the population library.

Starting from a definition of the enrichment ratio (ε_i) for a given clone i[16]:

$$\varepsilon_i = \log_2\left(\frac{x_{fi}}{x_{oi}}\right) - \log_2\left(\frac{\sum x_{fi}}{\sum x_{oi}}\right) \quad \textbf{(9)}$$

Here, x_{fi} and x_{oi} are the individual sequencing counts of clone i in the selected and unselected populations, respectively.

The variance for any given enrichment ratio ($\sigma_{\varepsilon i}$) can be found by propagation of errors:

$$\sigma_{\varepsilon i}{}^2 = \sigma_{xfi}{}^2 \frac{\delta\varepsilon_i}{\delta x_{fi}}^2 + \sigma_{xoi}{}^2 \frac{\delta\varepsilon_i}{\delta x_{oi}}^2 \quad \textbf{(10)}$$

Because the minimum error associated with counting sequences approximates Poisson noise[16, 17]:

$$\sigma_{xfi}{}^2 = x_{fi} \quad \textbf{(11)}$$

We can write the variance for ε_i as:

$$\sigma_{\varepsilon i}{}^2 = (\log_2 e)^2 [\tfrac{1}{x_{fi}} + \tfrac{1}{x_{oi}}] \quad \textbf{(12)}$$

To estimate the variance for all clones in the population, let us assume that the counts given in the unselected population approximate $\langle x_o \rangle$, the average sequenced depth of coverage. We can then write x_{fi}, the number of sequencing counts for a given clone i as:

$$x_{fi} = 2^{\varepsilon_i} < x_o > \frac{\sum x_{fi}}{\sum x_{fo}} \quad \textbf{(13)}$$

Substituting this into relation (4) yields the desired relationship:

$$\sigma_{\varepsilon i}{}^2 = (\log_2 e)^2 < x_o >^{-1} [2^{-\varepsilon_i} \frac{\sum x_{fi}}{\sum x_{fo}} + 1] \quad \textbf{(14)}$$

We can write this as the standard deviation for clone i:

$$\sigma_{\varepsilon i} = (\log_2 e) < x_o >^{-1/2} [2^{-\varepsilon_i} \frac{\sum x_{fi}}{\sum x_{fo}} + 1]^{1/2} \quad \textbf{(15)}$$

Note A 2.2 (cont'd) Here, the standard deviation is proportional to the inverse square root of the depth of coverage of the unselected library. Because the frequencies of variants in unselected library are log-normally distributed (**Figure A 2.5**), it is more appropriate to calculate the depth of coverage as the median, not the mean value.

Practically speaking, this derivation shows: (i.) the median depth of coverage for the unselected library should be high (in our lab's hands, >50); and (ii.) to minimize errors for variants with lower growth rates, the selected library should be sequenced to a higher depth of coverage compared to the unselected library.

We can apply similar principles to derive the standard deviation for the fitness metric (F_i):

$$\sigma_{fi} = (\log_2 e)^2 < x_o >^{-\frac{1}{2}} \left[\frac{\sum x_{oi}}{\sum x_{fi}} 2^{-(g_p(2^{fi}\left(\frac{\varepsilon_{wt}}{g_p}+1\right)-1))} + 1\right]^{\frac{1}{2}} (2^{fi}(\varepsilon_{wt} + g_p))^{-1} \qquad \textbf{(16)}$$

Here, g_p is the number of average population doublings and ε_{wt} is the enrichment value for the starting sequence.

Note A 2.3: Protein and nucleic acid sequences of tested LGK designs.

> LGK-His_{6x}

MPIATSTGDNVLDFTVLGLNSGTSMDGIDCALCHFYQKTPDAPMEFELLEYGEVPLAQPI
KQRVMRMILEDTTSPSELSEVNVILGEHFADAVRQFAAERNVDLSTIDAIASHGQTIWLL
SMPEEGQVKSALTMAEGAILASRTGITSITDFRISDQAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVCFIPPDVDGRRTDEYYDFDTGPGNVFIDAVVRHFTNGEQEYDKDGA
MGKRGKVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAHDLIRRAEAKGLSPDDIVA
TTTRITAQAIVDHYRRYAPSQEIDEIFMCGGGAYNPNIVEFIQQSYPNTKIMMLDEAGVP
AGAKEAITFAWQGMEALVGRSIPVPTRVETRQHYVLGKVSPGLNYRSVMKKGMAFGG
DAQQLPWVSEMIVKKKGKVITNNWALEHHHHHH

> LGK-His_{6x}

ATGCCGATTGCGACCTCAACGGGTGATAATGTTCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGGCG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATACCACGAGCCCGTC
TGAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGTCCGTCA
GTTCGCGGCCGAACGCAATGTGGATCTGTCAACCATTGACGCAATCGCTTCGCACGG
CCAGACGATTTGGCTGCTGAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGA
CCATGGCAGAAGGCGCTATCCTGGCGAGTCGTACGGGTATTACCTCCATCACGGATT
TCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCG
ATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTA
TCGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCGCACCGATGAATATT
ACGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTTTA
CCAATGGTGAACAGGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTAAAGTC
GATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCG
AAAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCACATGACCTGATTCGTCGC
GCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCACGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGACGA
AATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCA
ATCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCAGGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAAGCCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAGTTAGC
CCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCT
CAGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATCAC
CAACAACTGGGCGCTCGAGCACCACCACCACCACCAC

Note A 2.3 (cont'd)

>LGK.1

MPIATSTGDNVLDFTVLGLNSGTSMDGIDCALCHFYQKTPDAPMEFELLEYGEVPLAQPI
KQRVMRMILEDTTSPSELSEVNVILGEHFADAVRQFAAERNVDLSTIDAIASHGQTIWLL
SMPEEGQVKSALTMAEGAIIAARTGITSITDFRISDQAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVCFIPPDVDGRRTDEYYDFDTGPGNVFIDAVVRHFTNGEQEYDKDGA
MGKRGKVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAHDLIRRAEAKGLSPDDIVA
TTTRITAQAIVDHYRRYAPSQEIDEIFMCGGGAYNPNIVEFIQQSYPNTKIMMLDEAGVP
AGAKEAITFAWQGMECLVGRSIPVPTRVETRQHYVLGKVSPGLNYRSVMKKGMAFGG
DAQQLPWVSEMIVKKKGKVITNNWALEHHHHHH

>LGK.1

ATGCCGATTGCGACCTCAACGGGTGATAATGTTCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGGCG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATACCACGAGCCCGTC
TGAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGTCCGTCA
GTTCGCGGCCGAACGCAATGTGGATCTGTCAACCATTGACGCAATCGCTTCGCACGG
CCAGACGATTTGGCTGCTGAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGA
CCATGGCAGAAGGCGCTATCATAGCGGCTCGTACGGGTATTACCTCCATCACGGATT
TCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCG
ATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTA
TCGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCGCACCGATGAATATT
ACGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTTTA
CCAATGGTGAACAGGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTAAAGTC
GATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCG
AAAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCACATGACCTGATTCGTCGC
GCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCACGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGACGA
AATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCA
ATCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCAGGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAGTTAGC
CCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCT
CAGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATCAC
CAACAACTGGGCGCTCGAGCACCACCACCACCACCAC

>LGK.2

MPIATSTGDNPLDFTVLGLNSGTSMDGIDCALCHFYQKTPDAPMEFELLEYGEVPLAQPI
KQRVMRMILEDTTSPSELSEVNVILGEHFADAVRQFAAERNVDLSTIDAIASHGQTIWLL
SMPEEGQVKSALTMAEGAILASRTGITSITDFRISDQAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVCFIPPDVDGRRTDEYYDFDTGPGNVFIDAVVRHFTNGEQEYDKDGA
MGKRGKVDQELVDDFLKHPYFQLDPPKTCGREVFRDTLAHDLIRRAEAKGLSPDDIVA
TTTRITAQAIVDHYRRYAPSQEIDEIFMCGGGAYNPNIVEFIQQSYPNTKIMMLDEAGVP
AGAKEAITFAWQGMECLVGRSIPVPTRVETRQHYVLGKVSPGLNYRSVMKKGMAFGG
DAQQLPWVSEMIVKKKGKVITNNWALEHHHHHH

Note A 2.3 (cont'd)

>LGK.2

ATGCCGATTGCGACCTCAACGGGTGATAATCCTCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGGCG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATACCACGAGCCCGTC
TGAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGTCCGTCA
GTTCGCGGCCGAACGCAATGTGGATCTGTCAACCATTGACGCAATCGCTTCGCACGG
CCAGACGATTTGGCTGCTGAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGA
CCATGGCAGAAGGCGCTATCCTGGCGAGTCGTACGGGTATTACCTCCATCACGGATT
TCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCG
ATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTA
TCGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCGCACCGATGAATATT
ACGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTTTA
CCAATGGTGAACAGGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTAAAGTC
GATCAGGAACTGGTGGATGACTTTCTGAAACATCCGTATTTCCAACTGGACCCGCCG
AAAACCTGCGGCCGTGAAGTTTTTCGCGATACCCTGGCACATGACCTGATTCGTCGC
GCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCACGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGACGA
AATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCA
ATCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCAGGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAGTTAGC
CCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCT
CAGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATCAC
CAACAACTGGGCGCTCGAGCACCACCACCACCACCAC

>LGK.3

MPIATSTGDNPLDFTVLGLNSGTSMDGIDLALCHFYQKTPDAPMEFELLEYGEVPWAQP
IKQRVMRMIQEDTTSLSELSEVNVILGETFADAVHQFAAERNVDLSTIDAIGSHGQTIWL
NSMPEEGQVKSCLTMGEGAIIAARTGITTITDFRISDIAAGRQGAPLHAFFDALLLHHPTK
LRACQNIGGIANVTFIPPDVDGRLTDEYYDFDTGPGTVMIDAVVRHFTNGEQEYDKDGE
MGKRGKVDQELVDDFLKHPYFQLDPPKTCGREVFRDSLAHDLIRRAEAKGLSPDDIVAT
VTRITAQSIVDAYRRYAPSQEIDEIFLCGGGAYNPNIVEFIQQAYPNTKIMMLDEAGIPAR
AKEAITFAWLGMECLVGRSIPVPSRVETRQGYVLGKISPGLNYRSVMKKGMAFGGDAQ
QLPPVSEMIVKKKGKVITNNWDLEHHHHHH

Note A 2.3 (cont'd)

>LGK.3

ATGCCGATTGCGACCTCAACGGGTGATAATCCGCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACCTGGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGTGGGC
GCAGCCGATTAAACAACGTGTCATGCGCATGATCCAGGAAGATACCACGAGCCTGT
CTGAACTGTCAGAAGTCAACGTGATTCTGGGTGAAACCTTTGCGGATGCCGTCCATC
AGTTCGCGGCCGAACGCAATGTGGATCTGTCAACCATTGACGCAATCGGTTCGCACG
GCCAGACGATTTGGCTGAACAGTATGCCGGAAGAAGGTCAAGTGAAATCCTGCCTG
ACCATGGGGGAAGGCGCTATCATTGCGGCACGTACGGGTATTACCACGATCACGGA
TTTCCGTATTTCCGACATCGCAGCTGGTCGTCAAGGTGCACCGCTGCATGCATTTTTC
GATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGT
ATCGCCAATGTGACCTTTATTCCGCCGGATGTTGACGGCCGTCTGACCGATGAATAT
TACGATTTTGACACGGGTCCGGGCACCGTGATGATCGACGCAGTGGTTCGTCATTTT
ACCAATGGTGAACAGGAATATGATAAAGACGGTGAAATGGGCAAACGCGGTAAAG
TCGATCAGGAACTGGTGGATGACTTTCTGAAACATCCGTATTTCCAACTGGACCCGC
CGAAAACCTGTGGCCGTGAAGTTTTTCGCGATTCTCTGGCACATGACCTGATTCGTC
GCGCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCGTGACCCGTATT
ACGGCACAGAGCATCGTTGATGCATATCGTCGCTACGCGCCGTCACAAGAAATTGA
CGAAATCTTCCTGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCA
GCAAGCTTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTATTCCGGCAC
GCGCAAAAGAAGCGATTACGTTCGCCTGGTTAGGCATGGAATGTCTGGTTGGTCGTA
GCATCCCGGTTCCGAGCCGTGTCGAAACGCGCCAGGGTTATGTGCTGGGCAAAATTA
GCCCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATG
CTCAGCAACTGCCGCCTGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATC
ACCAACAACTGGGACCTCGAGCACCATCACCATCACCAC

>LGK.4

MPIATSTGDNPLDFTVLGLNSGTSMDGIDLALCHFYQKTPDAPMEFELLEYGEVPMAQS
IKQRVMRMIQEETTSLSELSEVNVILGETFADAVHQFAAEKNVDLSSIDAIGSHGVTIWL
NSMPEEGQVKSALTMGEGAIIAARTGITSITDFRISDIAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVTFIPPDVDGRKSDEYYDFDTGPGTVMIDAVVRHFTNGEQEYDKDGE
MGKRGKVDQELVDDFLKHPYFQLDPPKTTGREVFRDSLAHDLIRRAEAKGLSPDDIVAT
VTRITAQAIVDHYRRYAPSQEIDEIFLCGGGAYNPNIVEFIQQAYPNTKIMMLDEAGVPA
DAKEAITFAWQGMECLVGRSIPVPTRVETRQHYVLGKISPGLNYRSVMKKGMAFGGDA
QQLPPVSEMIVKKKGKVITNGGALEHHHHHH

Note A 2.3 (cont'd)

>LGK.4

ATGCCGATTGCGACCTCAACGGGTGATAATCCGCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACCTTGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGATGGC
GCAGTCCATTAAACAACGTGTCATGCGCATGATCCAAGAAGAGACCACGAGCCTTT
CTGAACTGTCAGAAGTCAACGTGATTCTGGGTGAAACCTTTGCGGATGCCGTCCATC
AGTTCGCGGCCGAAAAAAATGTGGATCTGTCAAGTATTGACGCAATCGGTTCGCAC
GGCGTTACGATTTGGCTGAACAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCT
GACCATGGGAGAAGGCGCTATCATCGCGGCTCGTACGGGTATTACCTCCATCACGG
ATTTCCGTATTTCCGACATTGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTT
CGATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGG
TATCGCCAATGTGACCTTTATTCCGCCGGATGTTGACGGCCGTAAATCCGATGAATA
TTACGATTTTGACACGGGTCCGGGCACGGTGATGATCGACGCAGTGGTTCGTCATTT
TACCAATGGTGAACAGGAATATGATAAAGACGGTGAAATGGGCAAACGCGGTAAA
GTCGATCAGGAACTGGTGGATGACTTTCTGAAACATCCGTATTTCCAACTGGACCCG
CCGAAAACCACGGGCCGTGAAGTTTTTCGCGATTCCCTGGCACATGACCTGATTCGT
CGCGCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCGTTACCCGTAT
TACGGCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGA
CGAAATCTTCCTGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCA
GCAAGCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCAG
ATGCAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGT
AGCATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAAT
CAGCCCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCG
ATGCTCAGCAACTGCCGCCCGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTT
ATCACCAACGGTGGCGCGCTCGAGCACCATCACCATCACCAC

>LGK.5

MPIATSTGDNSLDFTVLGLNSGTSMDGIDCALCHFYQENPTAPMEFELLEYGEVPLPKEI
KKRVMRMIQTNRTSPQELAEVNVLLGEHFADAVRIFAKERNVSLSTIDAIASHGQCIWL
QSMPGEGQVKSALTMGEGAIIAARTGITAITDFRISDQAAGRQGAPLQAFFDALLLHHPT
KLRACQNIGGIANVTFIPPCVDGRMTDEYFDFDTGPGMIFIDAVVRHFTNGEQEYDKDG
EMGARGKVDQELVDDFLKHPYFQLDPPKTTGREVFRDSLAYDLIRKAEAKGLSPEDIVA
TTTRITAQAIVDHYKRYAPSQDIDEIFLCGGGANNPNIVEFIQQAYPNTKIMMLDEAGVP
ARAKEAITFAWQGMEALVGRSIPVPTRVETRKPCVLGKISPGKNYRKVMKKGMAFGGD
AQQLPWVSEMIVKKNGKVITNKWDLEHHHHHH

Note A 2.3 (cont'd)

>LGK.5

ATGCCGATTGCGACCTCAACGGGTGATAATTCCCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGGAA
AACCCGACCGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGCC
GAAAGAGATTAAAAAGCGTGTCATGCGCATGATCCAGACCAATCGTACGAGCCCGC
AGGAACTGGCGGAAGTCAACGTGTTGCTGGGTGAACATTTTGCGGATGCCGTCCGTA
TTTTCGCGAAAGAACGCAATGTGTCCCTGTCAACCATTGACGCAATCGCTTCGCACG
GCCAGTGTATTTGGCTGCAAAGTATGCCGGGGGAAGGTCAAGTGAAATCCGCCCTG
ACCATGGGCGAAGGCGCTATCATTGCGGCTCGTACGGGTATTACCGCGATCACGGAT
TTCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGCAGGCATTTTTC
GATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGT
ATCGCCAATGTGACGTTTATTCCGCCGTGCGTTGACGGCCGTATGACCGATGAATAT
TTTGATTTTGACACGGGTCCGGGCATGATCTTCATCGACGCAGTGGTTCGTCATTTTA
CCAATGGTGAACAGGAATATGATAAAGACGGTGAGATGGGCGCCCGCGGTAAAGTC
GATCAGGAACTGGTGGATGACTTTCTGAAACACCCGTATTTCCAACTGGACCCGCCG
AAAACCACGGGCCGTGAAGTTTTTCGCGATTCCCTGGCATATGACCTGATTCGTAAG
GCGGAAGCCAAAGGTCTGAGCCCGGAGGACATCGTGGCCACCACGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATAAACGCTACGCGCCGTCACAAGATATTGACGA
AATCTTCCTTTGCGGCGGTGGCGCCAATAACCCGAATATTGTGGAATTTATCCAGCA
AGCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCACGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAAGCCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCAAACCATGCGTGCTGGGCAAAATTAG
CCCGGGTAAGAATTACCGCAAAGTGATGAAAAAAGGCATGGCATTTGGTGGCGATG
CTCAGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAACGGCAAAGTTATC
ACCAACAAGTGGGACCTCGAGCACCATCACCATCACCAC

>LGK.6

MPIATSTGDNVLDFRVLGLNSGTSMDGIDCALCHFYQKTPDAPMEFELKEYGEVPLQQP
IKQRVMRMILEDTTSPSELSEVNVILGEHFADAVGQFAAECGVDLRTIDAIASHGQTIWL
LSMPEEGQVKSALTMAEGAIIAARTGITSITDFRISDQAAGRQGAPLIAFFDALLLHHPTK
LRACQNIGGIANVCFIPPDVDGRRTDEYYDFDTGPGNVFIDAVVRHFTNGECEYDKDGA
MGKRGVVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAHDLIRRAQAKGLSPDDIVA
TTTRITAQAIVDHYRRFAPSQEIDEIFMCGGGAYNPNIVEFIQQKYPNTKIIMLDECGVPA
GAKEAITFAWQGMECLVGRSIPVPTRVETRQHYVLGKVSPGLNYRSVMKKGMAFGGD
ANQLPWVSAMVVKKEGKVKHNNWKLEHHHHHH

Note A 2.3 (cont'd)

>LGK.6

ATGCCGATTGCGACCTCAACGGGTGATAATGTTCTGGACTTTAGAGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGCAG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATACCACGAGCCCGTC
TGAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGTCGGGCA
GTTCGCGGCCGAATGTGGCGTGGATCTGCGCACCATTGACGCAATCGCTTCGCACGG
CCAGACGATTTGGCTGCTGAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGA
CCATGGCAGAAGGCGCTATCATAGCGGCTCGTACGGGTATTACCTCCATCACGGATT
TCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCG
ATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTA
TCGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCGCACCGATGAATATT
ACGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTTTA
CCAATGGTGAATGCGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTGTGGTC
GATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCG
AAAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCACATGACCTGATTCGTCGC
GCGCAGGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCACGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATCGTCGCTTTGCGCCGTCACAAGAAATTGACGA
AATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCA
AAAATACCCGAACACCAAAATTATCATGCTGGATGAATGCGGTGTCCCGGCAGGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAGTTAGC
CCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCT
AACCAACTGCCGTGGGTTTCTGCAATGGTCGTGAAGAAAGAAGGCAAAGTTAAACA
TAACAACTGGAAGCTCGAGCACCATCACCATCACCAC

>LGK.7

MPIATSEGDNVLDFTVLGLNSGTSMDGIDCALCHFYQATPDAPMEFELLEYGEVPLAQPI
KQRVMRMILEDSTSPSELSEVNVILGEHFADAAHQFAAERNVDLATIDAIASHGQTIWLN
SMPEEGQVKSALTMAEGAIIAARTGITCITDFRISDQAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVCFIPPDVDGRLTDEYYDFDTGPGNVFIDAVVRHYTNGEQEYDKDGA
MGKRGKVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAWDLIRRAEAKGLSPDDIVA
TVTRITAQAIVDHYRRYAPSQEIDEIFMCGGGAYNPNIVEFIQQSYPNTKIMMLDEAGVP
ARAKEAITFAWQGMECLVGRSIPVPTRVETRQPYVLGKVSPGLNYRSVMKKGMAFGG
DAQQLPWVSEMIVKKKGKVITNNWELEHHHHHH

Note A 2.3 (cont'd)

>LGK.7

ATGCCGATTGCGACCTCAGAAGGTGATAATGTTCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGGCA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGGCG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATTCCACGAGCCCGTCT
GAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGCGCATCAG
TTCGCGGCCGAACGCAATGTGGATCTGGCGACCATTGACGCAATCGCTTCGCACGGC
CAGACGATTTGGCTGAATAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGAC
CATGGCAGAAGGCGCTATCATAGCGGCTCGTACGGGTATTACCTGCATCACGGATTT
CCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCGA
TGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTAT
CGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCTGACCGATGAATATTA
CGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTATAC
CAATGGTGAACAGGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTAAAGTCG
ATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCGA
AAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCATGGGACCTGATTCGTCGCG
CGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCGTGACCCGTATTACG
GCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGACGAA
ATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCAA
TCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCACGTGC
AAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGCA
TCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCCGTATGTGCTGGGCAAAGTTAGCC
CGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCTC
AGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATCACC
AACAACTGGGAACTCGAGCACCATCACCATCACCAC

>LGK.8

MPIATSEGDNVLDFTVLGLNSGTSMDGIDCALCHFYQATPDAPMEFELLEYGEVPLAQPI
KQRVMRMILEDSTSPSELSEVNVILGEHFADAAHQFAAERNVDLATIDAIASHGQTIWLN
SMPEEGQVKSALTMAEGAIIAARTGITCITDFRISDQAAGRQGAPLIAFFDALLLHHPTKL
RACQNIGGIANVCFIPPDVDGRLTDEYYDFDTGPGNVFIDAVVRHFTNGECEYDKDGAM
GKRGVVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAHDLIRRAQAKGLSPDDIVATT
TRITAQAIVDHYRRFAPSQEIDEIFMCGGGAYNPNIVEFIQQKYPNTKIIMLDECGVPAGA
KEAITFAWQGMECLVGRSIPVPTRVETRQHYVLGKVSPGLNYRSVMKKGMAFGGDAN
QLPWVSAMVVKKEGKVKHNNWKLEHHHHHH

Note A 2.3 (cont'd)

>LGK.8

ATGCCGATTGCGACCTCAGAAGGTGATAATGTTCTGGACTTTACGGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGGCA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGGCG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATTCCACGAGCCCGTCT
GAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGCGCATCAG
TTCGCGGCCGAACGCAATGTGGATCTGGCGACCATTGACGCAATCGCTTCGCACGGC
CAGACGATTTGGCTGAATAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGAC
CATGGCAGAAGGCGCTATCATAGCGGCTCGTACGGGTATTACCTGCATCACGGATTT
CCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCGA
TGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTAT
CGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCTGACCGATGAATATTA
CGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTTTAC
CAATGGTGAATGCGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTGTGGTCG
ATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCGA
AAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCACATGACCTGATTCGTCGCG
CGCAGGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCACGACCCGTATTACG
GCACAGGCTATCGTTGATCACTATCGTCGCTTTGCGCCGTCACAAGAAATTGACGAA
ATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCAA
AAATACCCGAACACCAAAATTATCATGCTGGATGAATGCGGTGTCCCGGCAGGTGC
AAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGCA
TCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCACTATGTGCTGGGCAAAGTTAGCC
CGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCTA
ACCAACTGCCGTGGGTTTCTGCAATGGTCGTGAAGAAAGAAGGCAAAGTTAAACAT
AACAACTGGAAGCTCGAGCACCATCACCATCACCAC

>LGK.9

MPIATSTGDNVLDFRVLGLNSGTSMDGIDCALCHFYQKTPDAPMEFELLEYGEVPLQQP
IKQRVMRMILEDTTSPSELSEVNVILGEHFADAVGQFAAECGVDLRTIDAIASHGQTIWL
LSMPEEGQVKSALTMAEGAIIAARTGITSITDFRISDQAAGRQGAPLIAFFDALLLHHPTK
LRACQNIGGIANVCFIPPDVDGRRTDEYYDFDTGPGNVFIDAVVRHYTNGEQEYDKDGA
MGKRGKVDQELVDDFLKMPYFQLDPPKTTGREVFRDTLAWDLIRRAEAKGLSPDDIVA
TVTRITAQAIVDHYRRYAPSQEIDEIFMCGGGAYNPNIVEFIQQSYPNTKIMMLDEAGVP
ARAKEAITFAWQGMECLVGRSIPVPTRVETRQPYVLGKVSPGLNYRSVMKKGMAFGG
DAQQLPWVSEMIVKKKGKVITNNWELEHHHHHH

Note A 2.3 (cont'd)

>LGK.9

ATGCCGATTGCGACCTCAACGGGTGATAATGTTCTGGACTTTAGAGTTCTGGGCCTG
AATAGCGGTACGAGTATGGATGGTATTGACTGCGCACTGTGTCATTTCTATCAGAAA
ACCCCGGATGCTCCGATGGAATTTGAACTGCTGGAATACGGCGAAGTTCCGCTGCAG
CAGCCGATTAAACAACGTGTCATGCGCATGATCCTGGAAGATACCACGAGCCCGTC
TGAACTGTCAGAAGTCAACGTGATTCTGGGTGAACATTTTGCGGATGCCGTCGGGCA
GTTCGCGGCCGAATGTGGCGTGGATCTGCGCACCATTGACGCAATCGCTTCGCACGG
CCAGACGATTTGGCTGCTGAGTATGCCGGAAGAAGGTCAAGTGAAATCCGCCCTGA
CCATGGCAGAAGGCGCTATCATAGCGGCTCGTACGGGTATTACCTCCATCACGGATT
TCCGTATTTCCGACCAGGCAGCTGGTCGTCAAGGTGCACCGCTGATCGCATTTTTCG
ATGCTCTGCTGCTGCATCACCCGACCAAACTGCGCGCGTGCCAGAACATTGGCGGTA
TCGCCAATGTGTGTTTTATTCCGCCGGATGTTGACGGCCGTCGCACCGATGAATATT
ACGATTTTGACACGGGTCCGGGCAACGTGTTCATCGACGCAGTGGTTCGTCATTATA
CCAATGGTGAACAGGAATATGATAAAGACGGTGCTATGGGCAAACGCGGTAAAGTC
GATCAGGAACTGGTGGATGACTTTCTGAAAATGCCGTATTTCCAACTGGACCCGCCG
AAAACCACGGGCCGTGAAGTTTTTCGCGATACCCTGGCATGGGACCTGATTCGTCGC
GCGGAAGCCAAAGGTCTGAGCCCGGATGACATCGTGGCCACCGTGACCCGTATTAC
GGCACAGGCTATCGTTGATCACTATCGTCGCTACGCGCCGTCACAAGAAATTGACGA
AATCTTCATGTGCGGCGGTGGCGCCTATAACCCGAATATTGTGGAATTTATCCAGCA
ATCGTACCCGAACACCAAAATTATGATGCTGGATGAAGCAGGTGTCCCGGCACGTG
CAAAAGAAGCGATTACGTTCGCCTGGCAGGGCATGGAATGCCTGGTTGGTCGTAGC
ATCCCGGTTCCGACCCGTGTCGAAACGCGCCAGCCGTATGTGCTGGGCAAAGTTAGC
CCGGGTCTGAATTACCGCTCTGTGATGAAAAAAGGCATGGCATTTGGTGGCGATGCT
CAGCAACTGCCGTGGGTTTCTGAAATGATCGTGAAGAAAAAAGGCAAAGTTATCAC
CAACAACTGGGAACTCGAGCACCATCACCATCACCAC

Note A 2.4: Python scripts used for deep sequencing normalization and statistics.
Script: QuickNormalize.py
Version: 1.9, Build: 20150527
This script normalizes a growth selection or a FACS screen.
Dependencies: Enrich 0.2, Python 2.7 or higher

Inputs: Command line:
python QuickNormalize.py –n [growth or FACS] –s [start residue] –l [tile length] –g [growth: g_p] –d [FACS: std dev] –c [FACS: percent collected] –p [path to enrich output directory] –t [significant unselected counts threshold (default: 5)] –w [path to wild-type AA sequence (default: ./WTSeq)] –o [Output a separate heatmap csv (default: True)]

Flag inputs:
-n String: growth or FACS
-s Integer: Example: 0, this is the start residue of your tile counting from zero
-l Integer: Example: 40, this is the length of the tile in amino acids
-g Float: Example: 10.0, this is the number of doublings of your selection
-d Float: Example: 0.6, this is the standard deviation for a FACS sort
-c Float: Example: 0.05, this is the percent collected for a FACS sort
-p String: ./output/ (Tailing slash is needed), Path to the enrich output directory
-t Integer: 5 (Default: 5), Number of unselected counts to be significant
-w String: ./wt_seq.txt (Default: ./WTSeq), A file of the wild-type amino acid sequence
-o String: True or False (Default: True), Output a CSV heatmap?

Example command line for growth:
python QuickNormalize.py –n growth –s 0 –l 40 –g 10.0 –p ./1/data/output/ > Tile1Normed

Example command line for FACS:
python QuickNormalize.py –n FACS –s 0 –l 76 –d 0.6 –c 0.05 –p ./tile/data/output/ -w ./wt_seq.txt > Tile1Normed

Help command:
python QuickNormalize.py –h

Notes:
It is highly recommended to direct the output to a file using > [file name] such that there is a saved copy of the normalization output. The output includes column and csv heatmap data used by other scripts for further analyses (stats, replicate errors). The output heatmap csv will be named heatmap_startresi_#.csv with # being the number given for the start residue. The wild-type amino acid sequence file is a single line ASCII file with the wild-type amino acid sequence. This file must be stripped of special characters hidden with rich-text editors. GNU nano can be used to edit this file and strip special hidden characters.

Script: QuickStats.py
Version: 1.1, Build: 20150428
Version note: v1.1 and higher is for use with QuickNormalize.py v1.7 and higher.

Note A 2.4 (cont'd) This script calculates the reportable statistics for a deep sequencing run. Dependencies: Enrich 0.2, Python 2.7 or higher

Command line:
python QuickStats.py –f [path to file with normalized output] –p [path to root enrich tile directory]

Flag inputs:
-f String: ./Tile1Normed (File from QuickNormalize output) path to the quicknormalize output
-p String: ./tile/ (Tailing slash is needed), path to the root directory of your tile data for enrich

Example command line:
python QuickStats.py –f ./Tile1Normed –p ./tile/ > Tile1Stats

Help command:
python QuickStats.py –h

Enrich files used:
data/output/counts_sel_example_F_N_include_filtered_B_DNA_qc
data/output/counts_unsel_example_F_N_include_filtered_B_DNA_qc
data/output/counts_unsel_example_F_N_include_filtered_B_DNA_qc.m1
data/output/counts_unsel_example_F_N_include_filtered_B_DNA_qc.m2
data/output/counts_unsel_example_F_N_include_filtered_B_PRO_qc
data/output/counts_unsel_example_F_N_include_filtered_B_PRO_qc.m1
input/example_local_config

Notes:
Normalization of the dataset is required to run this script. Additionally, the script uses the <translate_start> tag from the example_local_config file. Therefore, the enrich patch is needs to be applied. Unlike the normalization and other scripts, this script needs the root directory of the tile (i.e. the directory with the data and input directories). It is highly recommended to direct the output to a file using > [file name] such that there is a saved copy of the stats output. The script outputs all reportable statistics. This file assumes a certain naming scheme for the enrich output (listed above).

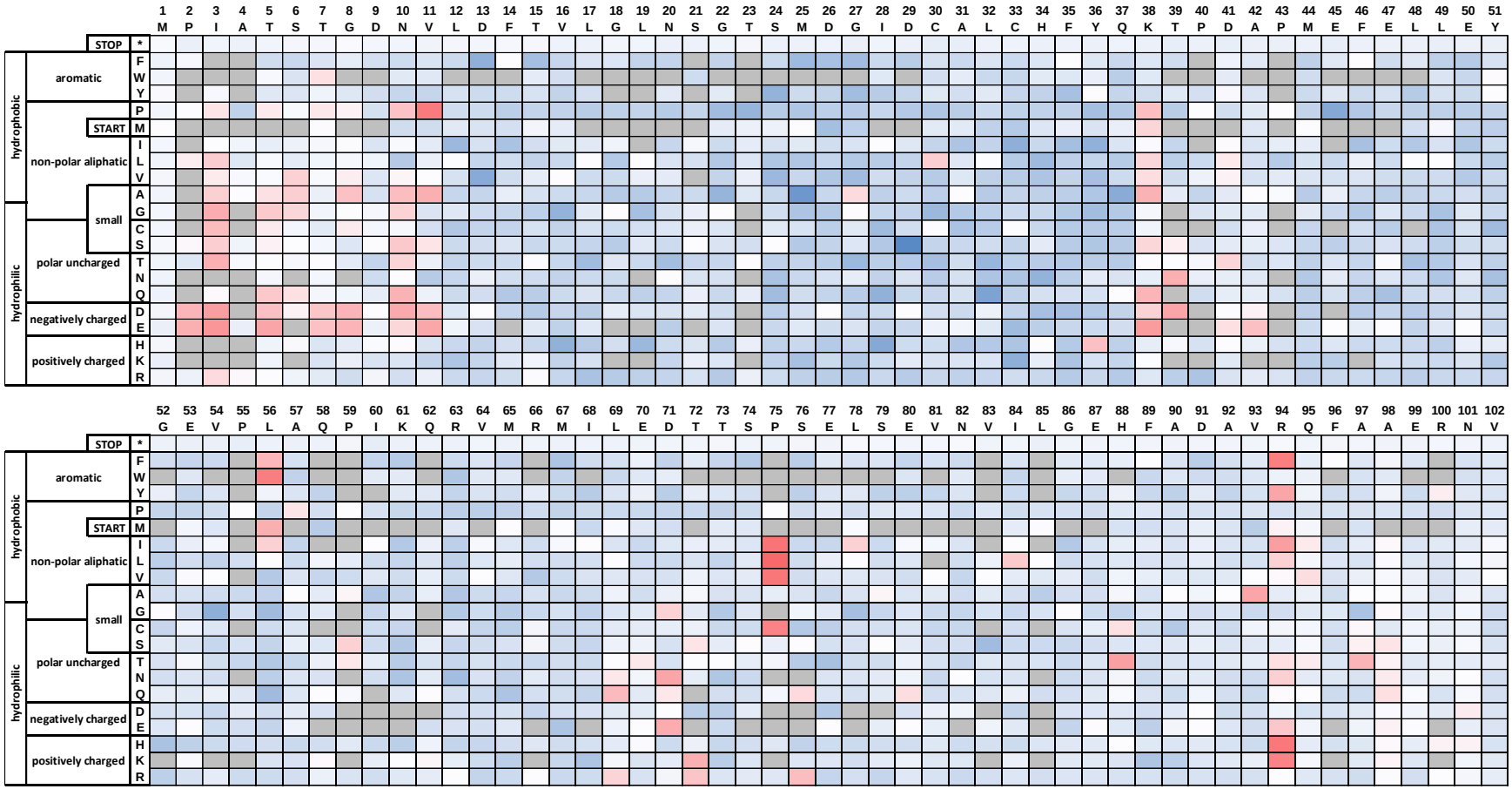


Figure A 2.1: Heatmap of first selection.

Figure A 2.1 (cont'd)

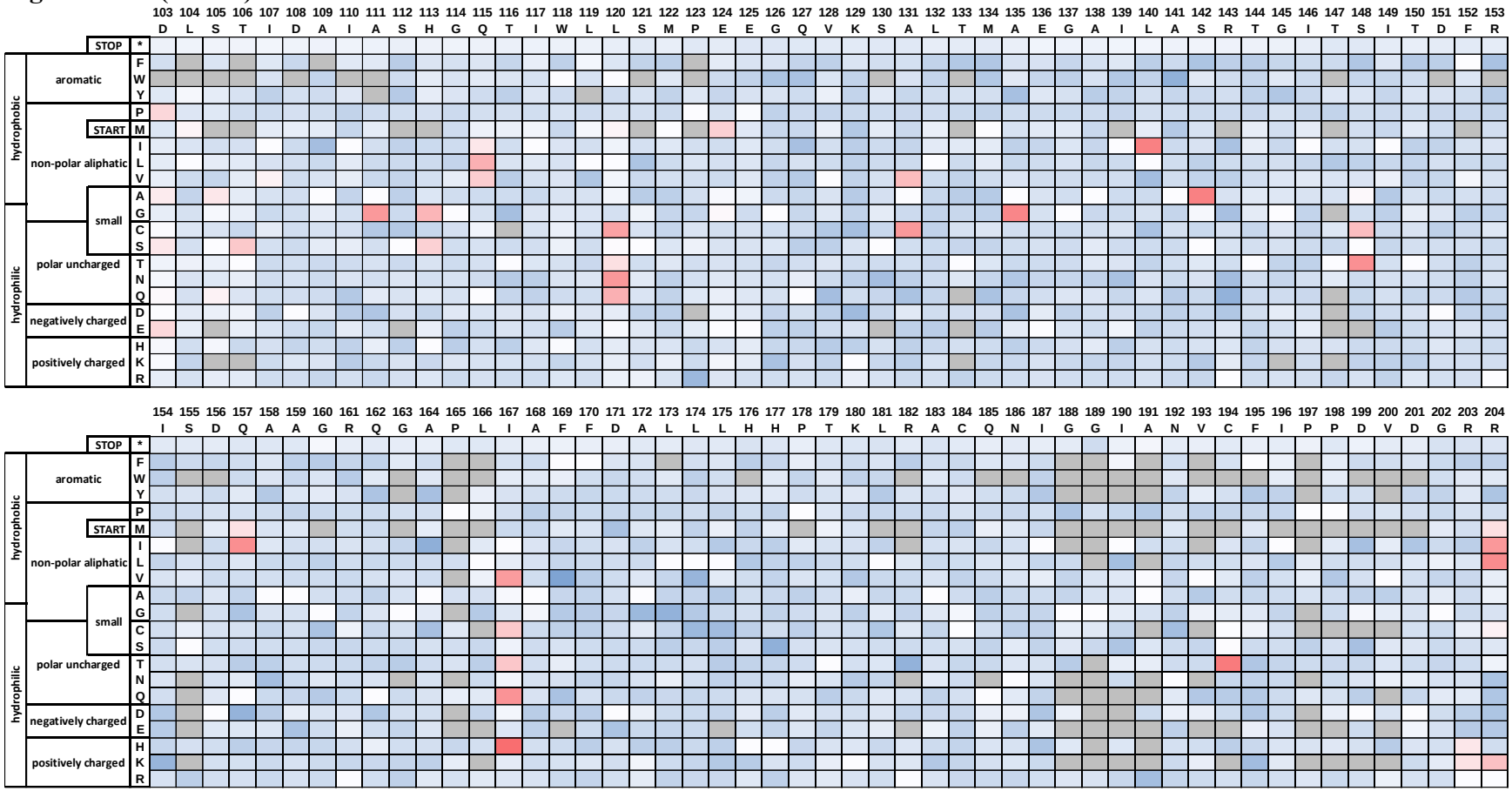


Figure A 2.1 (cont'd)

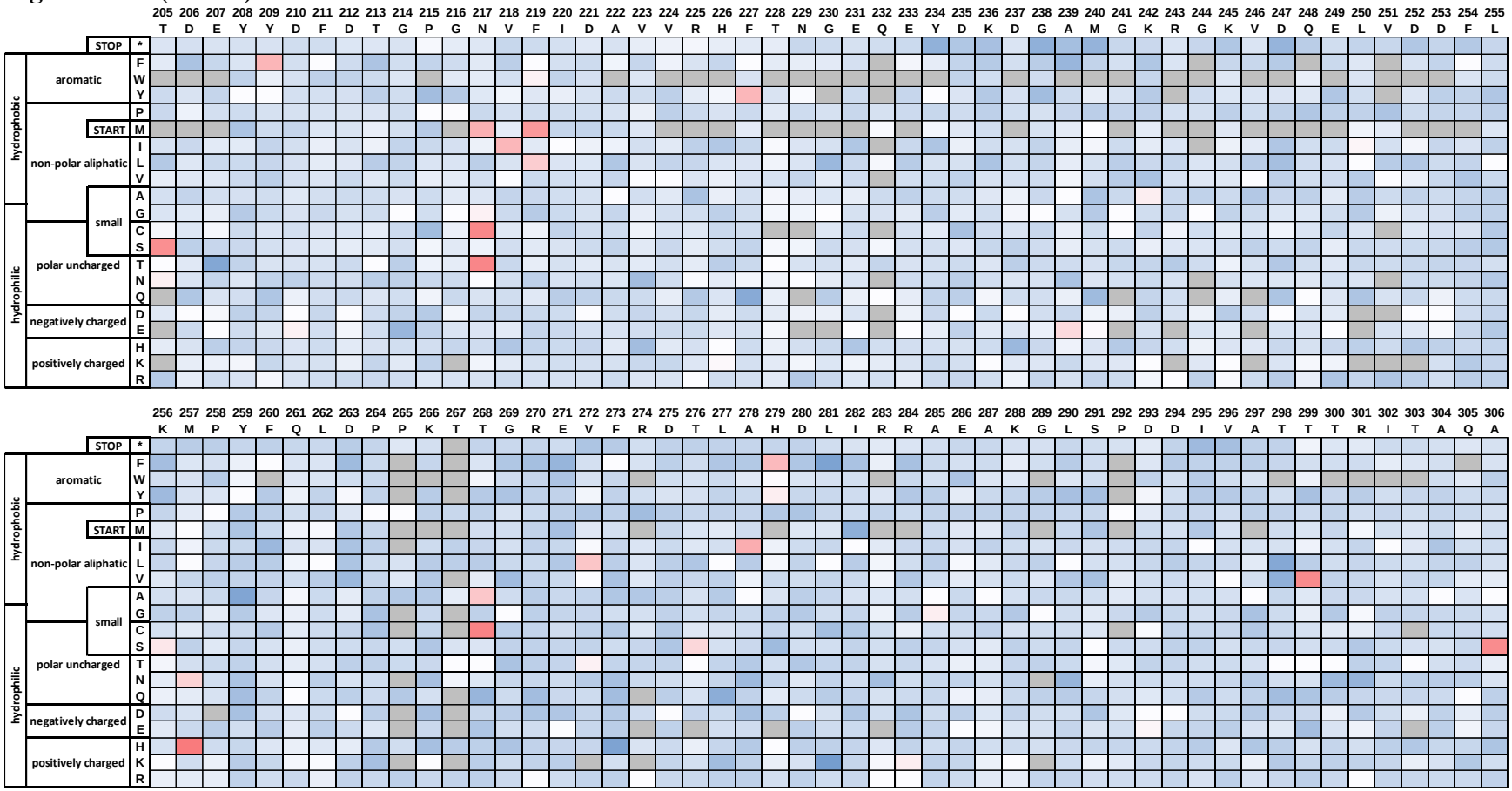


Figure A 2.1 (cont'd)

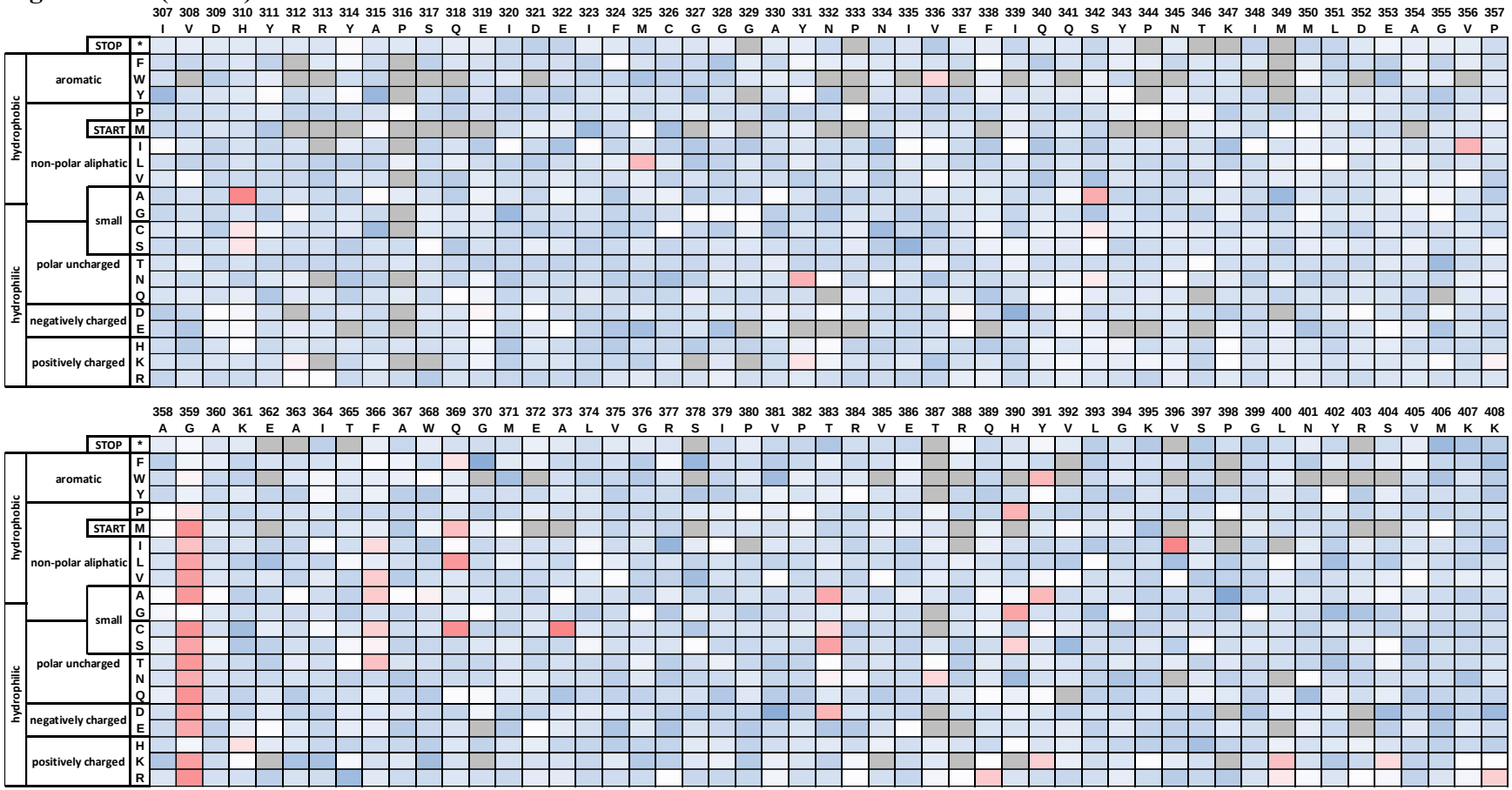
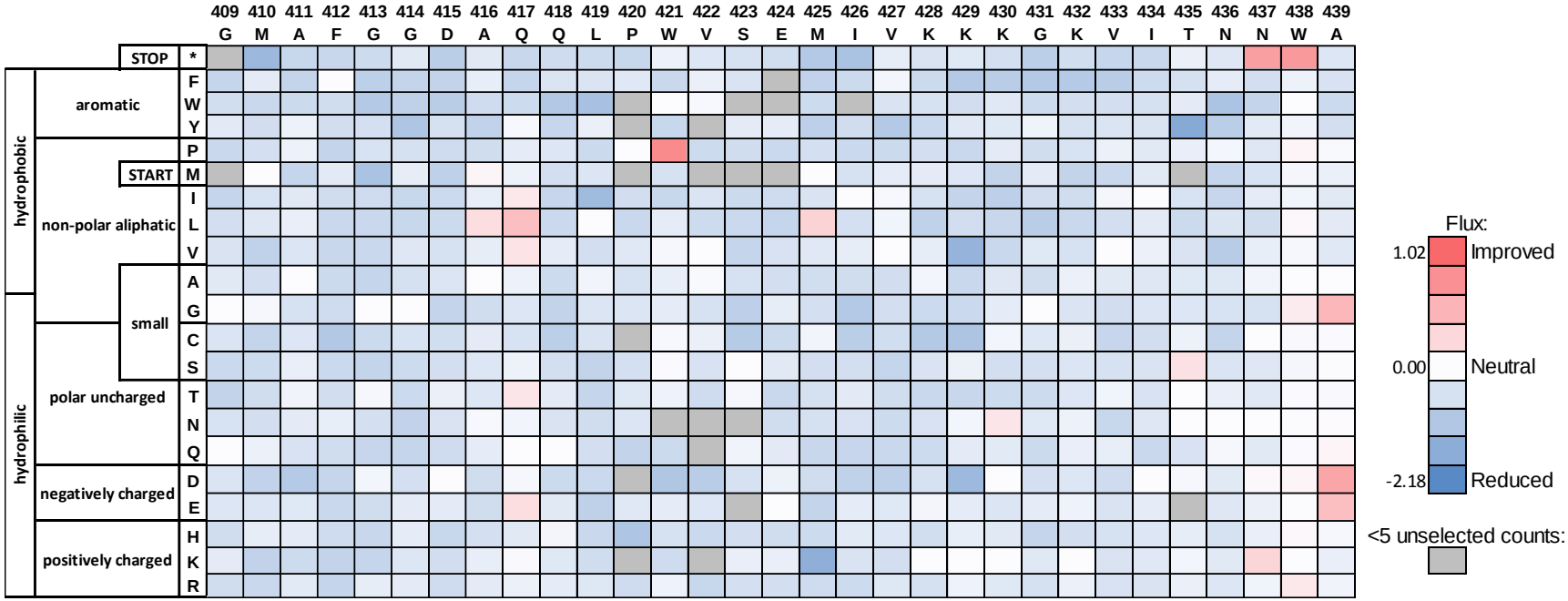


Figure A 2.1 (cont'd)



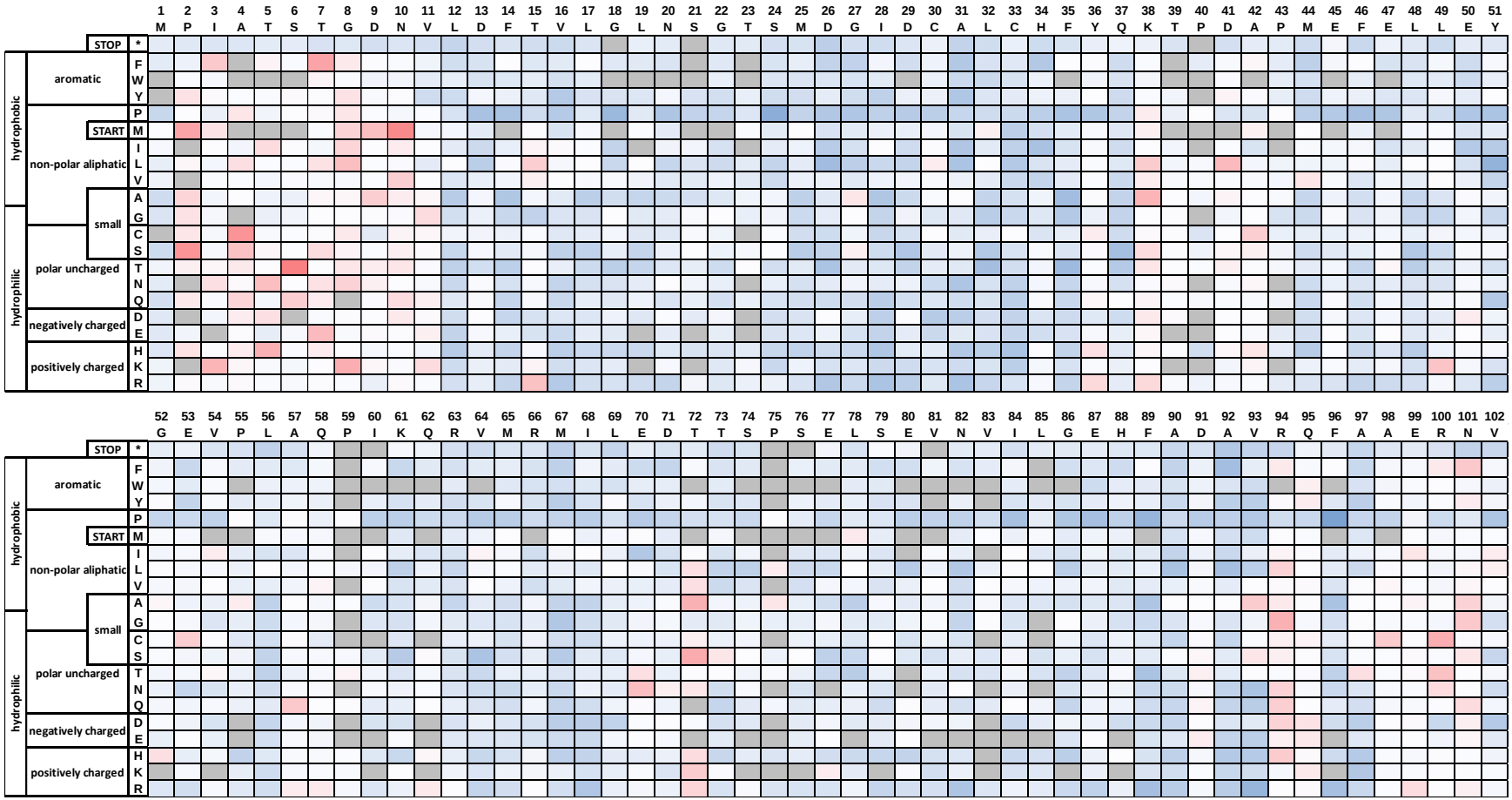


Figure A 2.2: Heatmap of second selection.

Figure A 2.2 (cont'd)

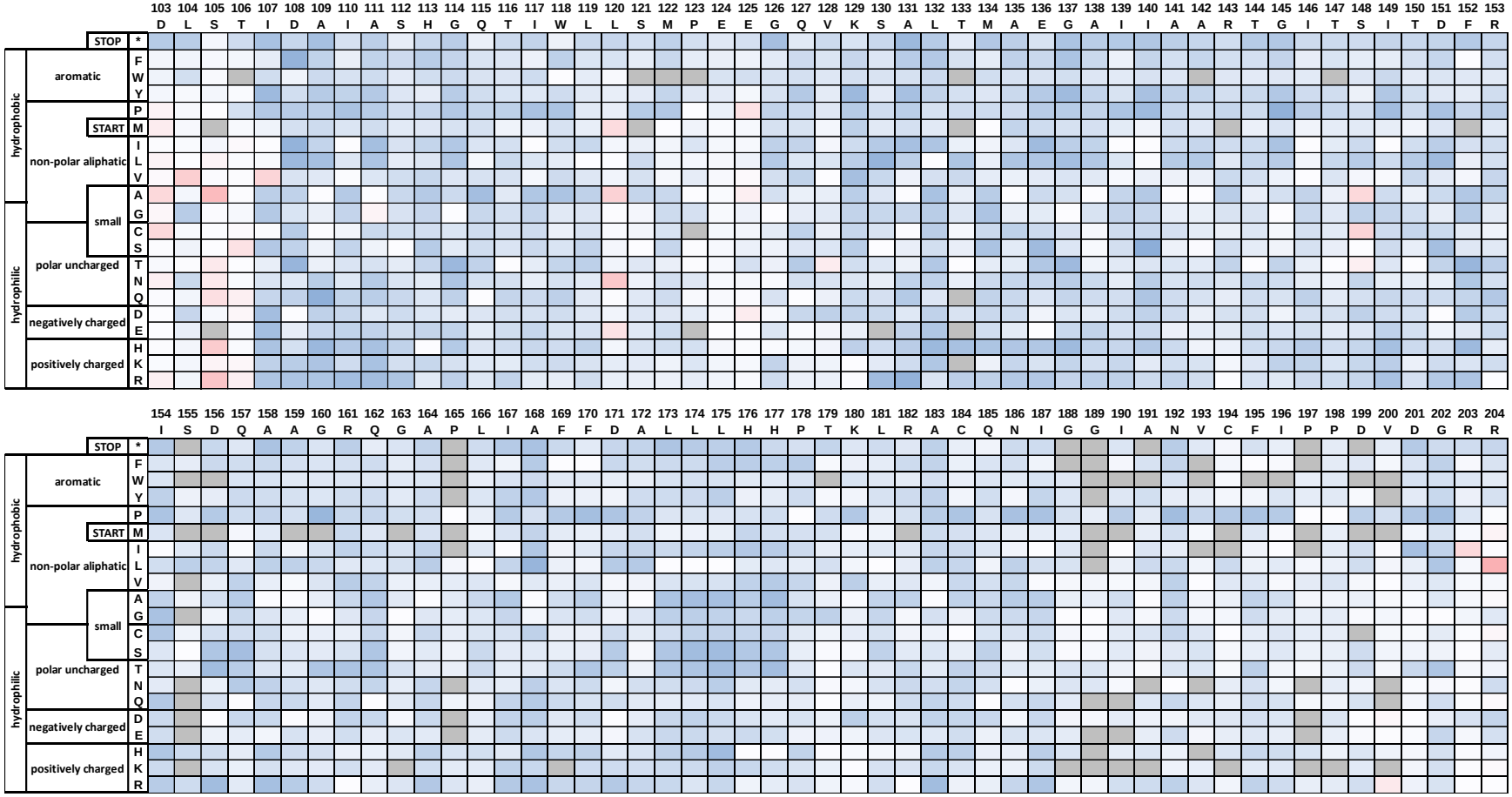


Figure A 2.2 (cont'd)

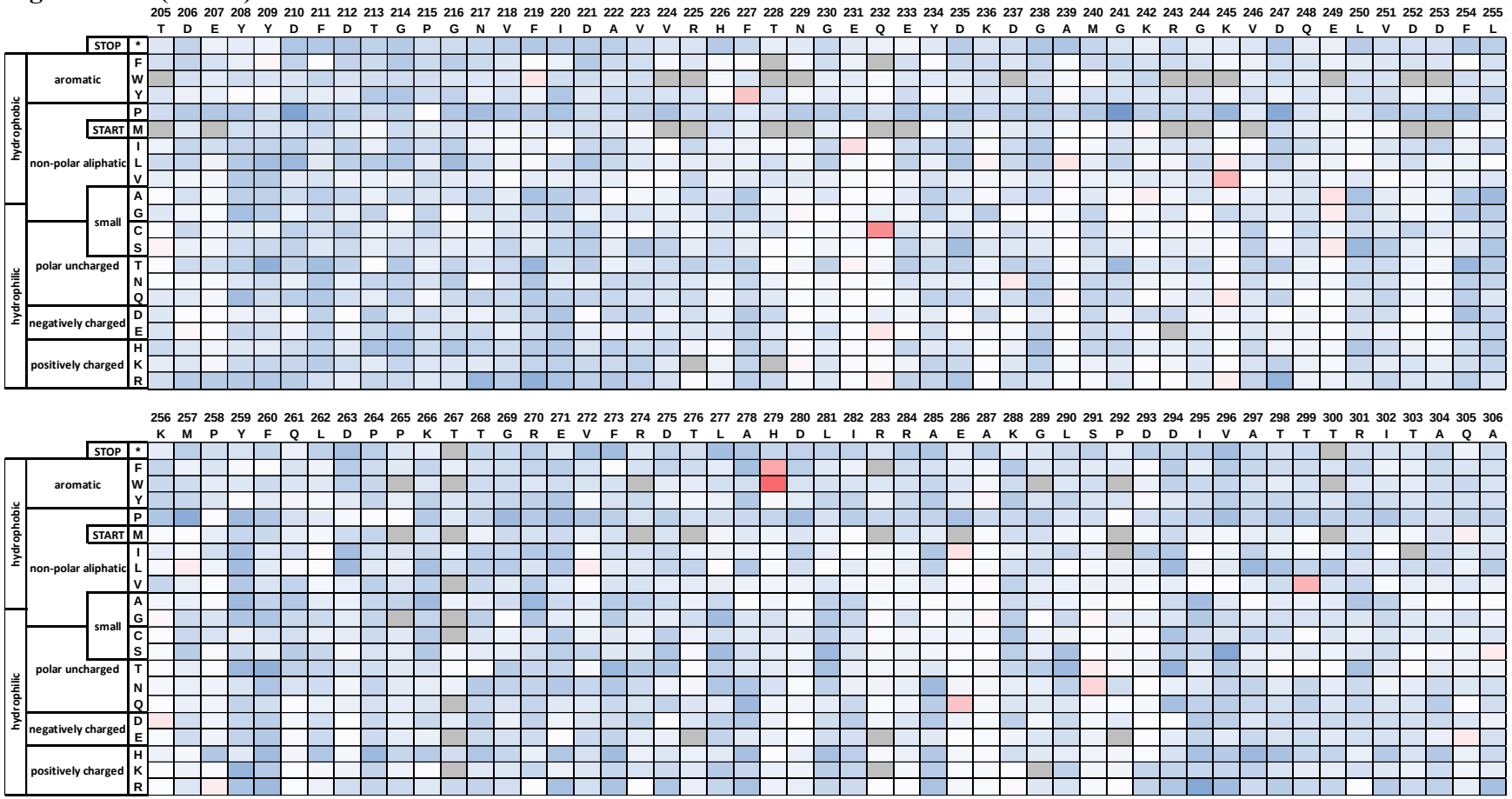


Figure A 2.2 (cont'd)

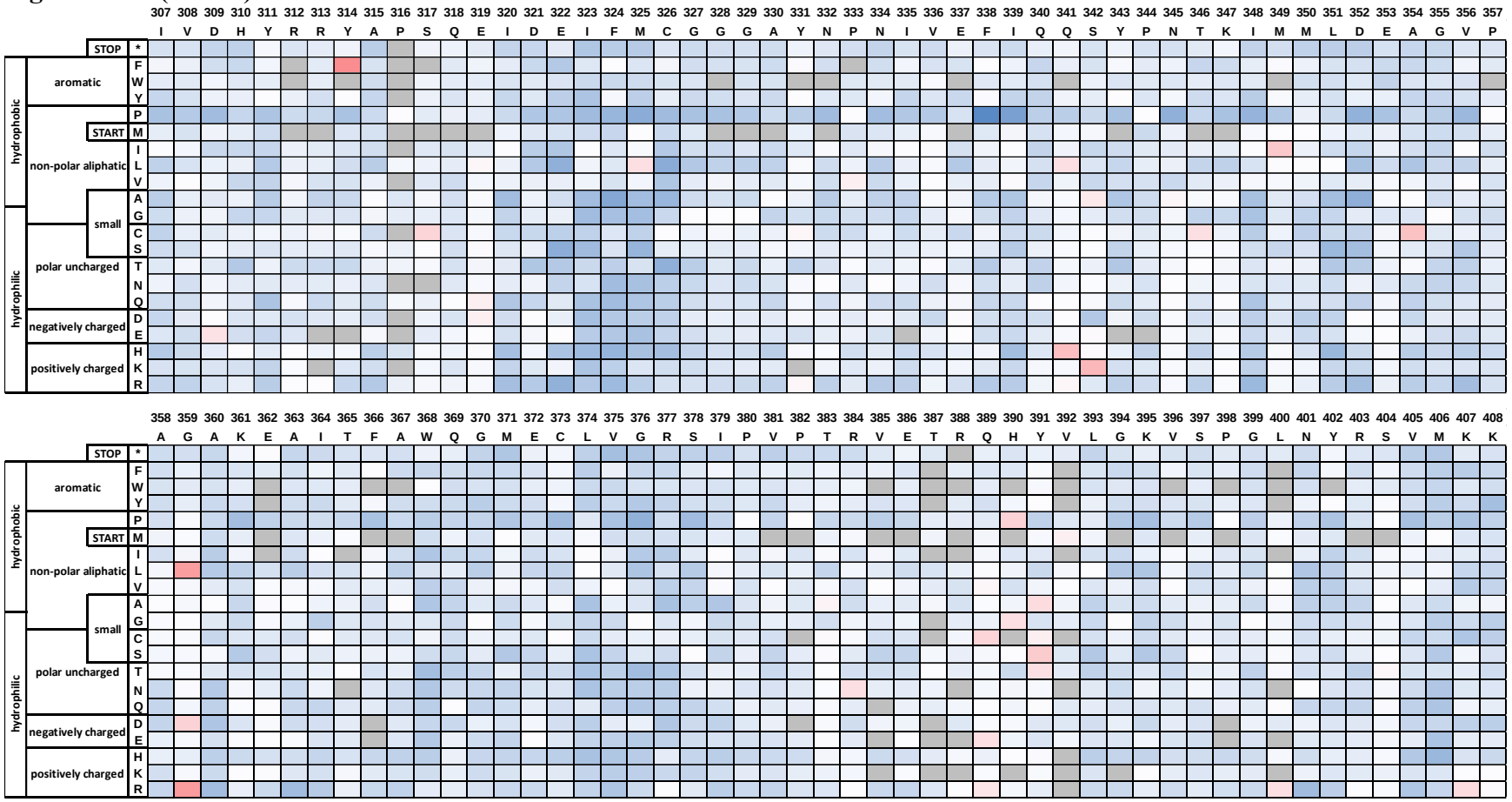
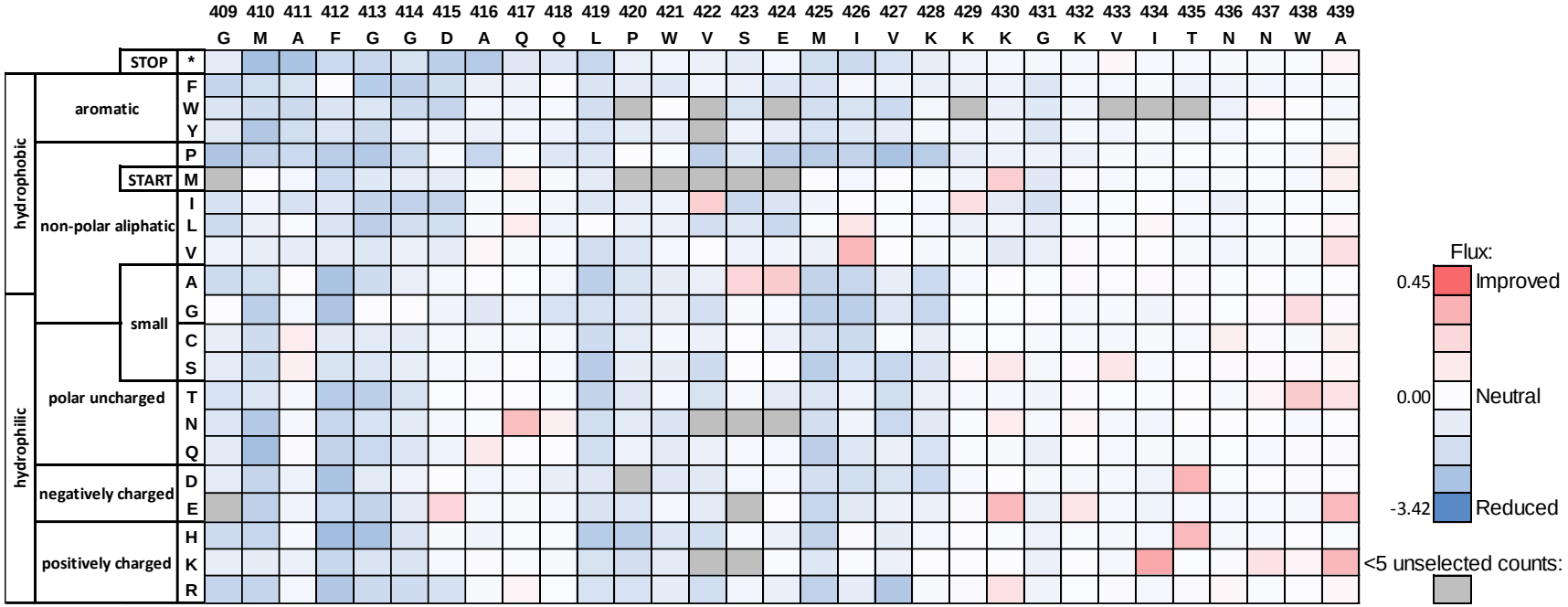


Figure A 2.2 (cont'd)



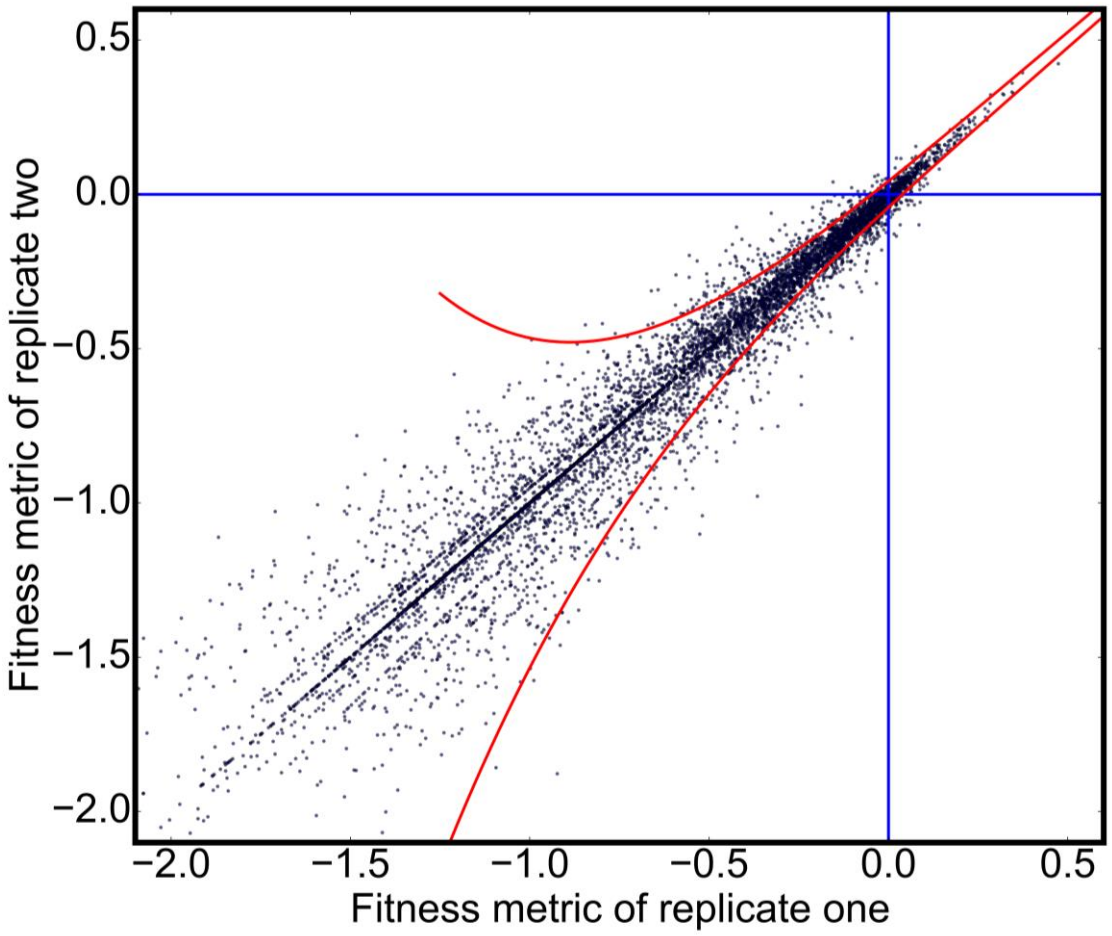


Figure A 2.3: Reproducibility of replicate fitness from second selection. Vertical and horizontal blue lines are used to aid the eye. The red lines demarcate the theoretically predicted error at two standard deviations.

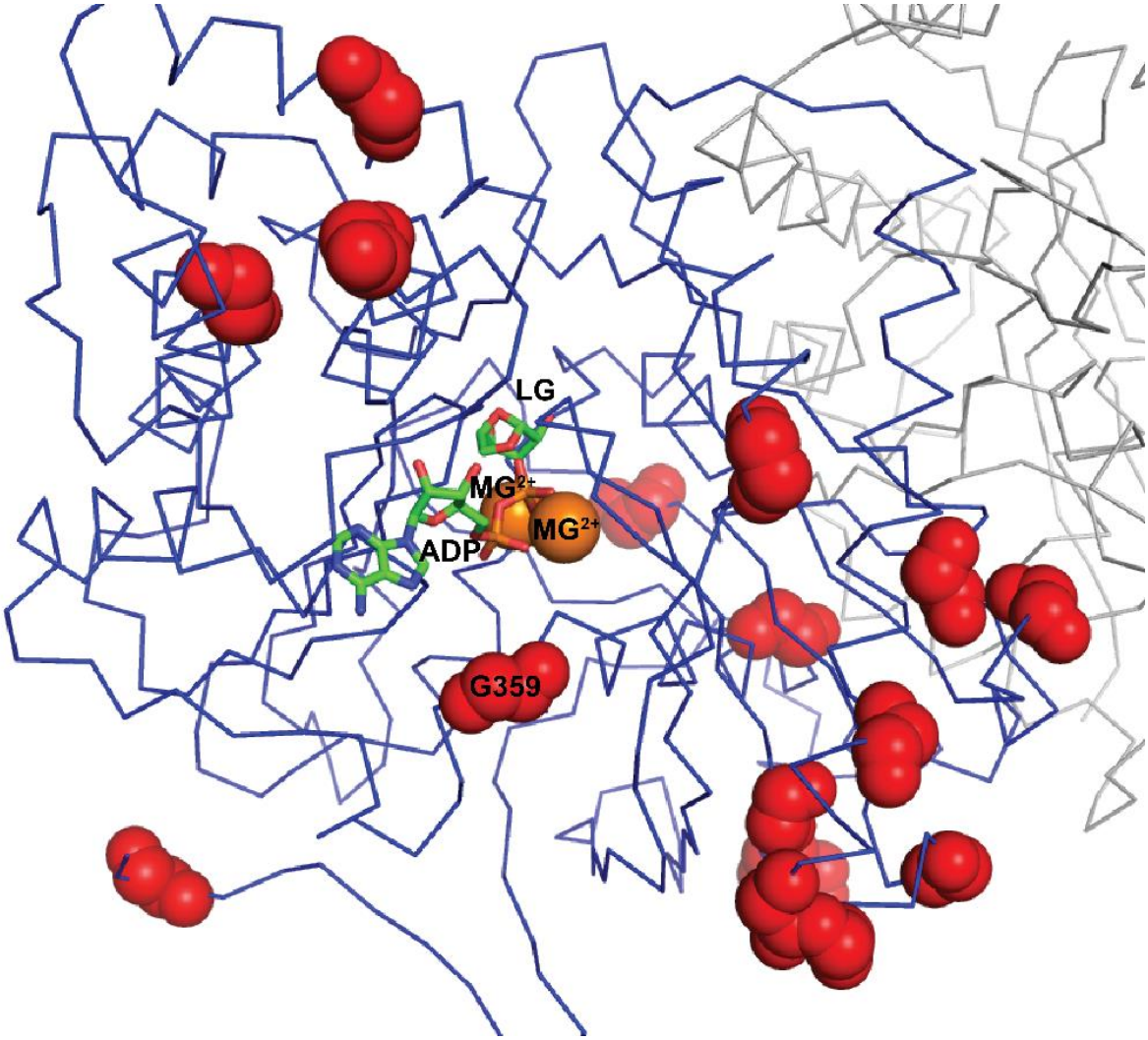


Figure A 2.4: LGK structure (PDB: 4ZLU) showing the locations of each residue mutated in design LGK.9 (red). The ADP and LG ligands are shown as sticks, while magnesiums are shown as orange spheres. The gray line traces the other LGK subunit in the homodimer.

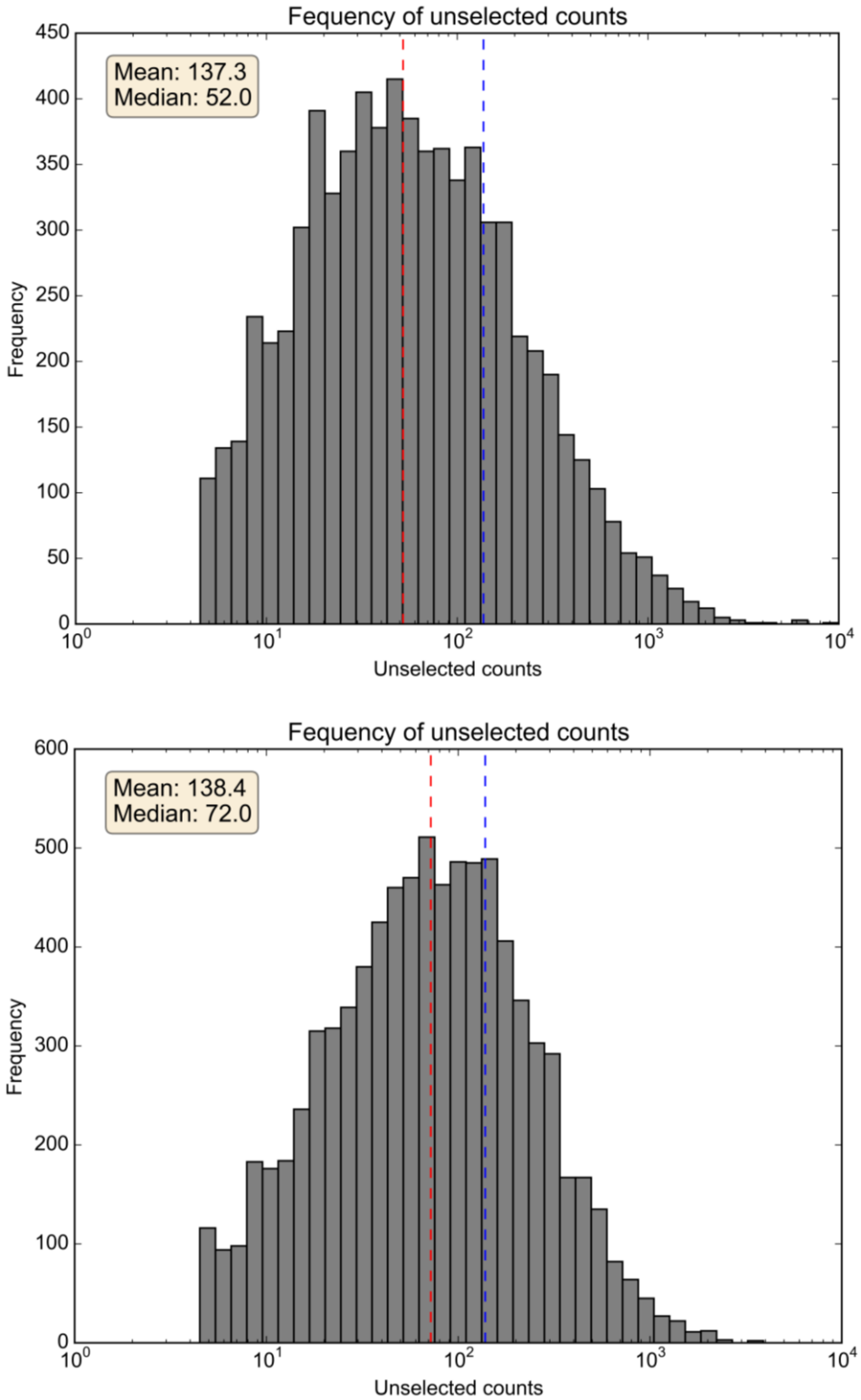


Figure A 2.5: Frequency distribution of mutation counts within the unselected libraries in the 1st selection (top) and 2nd selection (bottom). Dashed vertical lines indicate the median (red) and mean (blue) read coverage of the library.

Table A 2.1: Growth rates, lysate activity, and theoretical flux for cultures expressing pJK_proJK1_LGK on 8 and 10 g/L levoglucosan in M9 minimal media with carbenicillin aerobically at 37°C. Error is reported at one standard deviation (growth rate n≥4, lysate activity assay n=3 from two biological replicates).

LG concentration [g/L]	Growth Rate [h^{-1}]	Measured flux [mmol LG hr^{-1} $gDCW^{-1}$]	Theoretical flux^ [mmol LG hr^{-1} $gDCW^{-1}$]	[LG]i/[LG]e*
8	0.16 ± 0.05	3.04 ± 0.29	2.58	0.85
10	0.21 ± 0.01	3.70 ± 0.21	3.38	0.91

^ Theoretical flux is the calculated glycolytic flux required to support the measured growth rate.
*Calculation of ratio of internal to external LG concentrations. This value should approach unity for reaction-limited reactions.

Table A 2.2: Statistics for read coverage of the combined LGK SSM libraries for the first selection and second selection.

	1st Selection	2nd Selection
Unselected population DNA reads post filter	1,626,972	1,993,785
Selected population DNA reads post filter	3,305,210	4,045,218
Number of single nonsynonymous mutations above a fitness metric of:		
0.00	323	417
0.10	244	99
0.15	215	54
0.30	151	7
0.50	86	0
1.00	1	0
Number of unselected mutations above 5 counts	8,056	8,312
Number of mutations retained in the selected population	7,674	6,039
Percent of possible codon substitutions observed:		
1-base substitution	99.7%	99.9%
2-base substitutions	96.8%	97.2%
3-base substitutions	93.5%	94.6%
All substitutions	95.8%	96.5%
Percent of reads in unselected library with:		
No nonsynonymous mutations	22.8%	34.6%
One nonsynonymous mutation	67.5%	57.7%
Multiple nonsynonymous mutations	9.7%	7.7%
Coverage of single nonsynonymous mutations:	91.8%	94.7%

Table A 2.3: **Summary of thermostability, kinetic parameters, specific growth rates, and lysate flux for selected LGK variants.** Mean and SD for apparent melting temperatures ($T_{m,app}$) of selected LGK variants were measured in 50 mM sodium phosphate buffer pH 7.6 using a SYPRO Orange thermal shift assay (n=3). The kinetic parameters of purified protein variants were measured using a coupled glucose-6-phosphate dehydrogenase assay. Wild-type LGK: k_{cat}: 24 ± 1 s^{-1}, K_m 119 ± 12 mM. Error values are propagated from the standard deviation of three assays from two different experiments. Mean and SD for specific growth rates of isogenic *E. coli* Tuner cultures expressing the specific LGK variant in M9 minimal media with 10 g/L LG and carbenicillin at 37°C at 250 rpm (n≥2). Lysate flux is the *in vitro* activity of culture lysate tested with 10 g/L LG at 30°C. Mean and SD are reported from three assays performed on each culture grown in 10 g/L LG (4 g/L glucose for LGK.1 and LGK.2) in M9 minimal media and carbenicillin (n=2 cultures). ND = not determined.

Variant	$T_{m,app}$ (°C)	k_{cat} (s^{-1})/ $k_{cat,wt}$ (s^{-1})	K_m (mM)/ $K_{m,wt}$ (mM)	k_{cat}/K_m (M^{-1} s^{-1})/ $k_{cat,wt}/K_{m,wt}$ (M^{-1} s^{-1})	Specific Growth Rate (h^{-1})	Lysate Flux (mmol LG $gDCW^{-1}$ hr^{-1})	1st Selection Fitness
LGK	33.8 ± 0.7	1.00	1.00	1.00	0.21 ± 0.00	3.70 ± 0.21	0.00
V11P	33.9 ± 0.1	1.11 ± 0.14	1.08 ± 0.32	1.06 ± 0.19	0.53 ± 0.04	ND	0.90
P75L	35.2 ± 0.2	ND	ND	ND	ND	ND	1.02
R94H	35.7 ± 0.2	1.15 ± 0.15	1.06 ± 0.24	1.10 ± 0.12	0.49 ± 0.07	16.72 ± 0.92	0.92
H113G	38.7 ± 0.1	0.02 ± 0.00	2.13 ± 0.22	0.01 ± 0.00	0.37 ± 0.01	ND	0.49
A135G	36.4 ± 0.8	ND	ND	ND	ND	ND	0.82
L140I	36.0 ± 0.1	0.88 ± 0.10	0.80 ± 0.19	1.11 ± 0.15	0.49 ± 0.04	9.99 ± 0.25	0.86
S142A	34.6 ± 0.4	0.95 ± 0.10	0.77 ± 0.16	1.26 ± 0.14	0.50 ± 0.05	11.70 ± 1.20	0.86
I167H	43.6 ± 0.1	0.22 ± 0.03	1.27 ± 0.24	0.17 ± 0.01	ND	ND	0.99
I167N	35.9 ± 0.1	0.04 ± 0.00	3.86 ± 0.20	0.01 ± 0.00	ND	ND	-0.13
C194T	39.8 ± 0.4	0.95 ± 0.16	1.39 ± 0.42	0.70 ± 0.11	0.53 ± 0.08	ND	0.90
D212A	35.2 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-0.63
N217T	32.9 ± 0.9	0.84 ± 0.06	0.87 ± 0.13	1.00 ± 0.10	0.49 ± 0.02	ND	0.82
N217S	34.4 ± 0.1	0.50 ± 0.06	1.15 ± 0.23	0.44 ± 0.04	0.03 ± 0.01	ND	-0.19
M257H	33.4 ± 0.7	1.31 ± 0.09	0.92 ± 0.10	1.43 ± 0.08	0.45 ± 0.02	17.28 ± 0.95	0.90
T268C	37.8 ± 0.0	0.39 ± 0.04	1.11 ± 0.23	0.35 ± 0.04	0.49 ± 0.06	ND	0.84
A306S	34.9 ± 0.2	ND	ND	ND	ND	ND	0.78
H310A	33.2 ± 1.0	ND	ND	ND	ND	ND	0.80

Table A 2.3 (cont'd)

G359R	34.9 ± 0.1	1.63 ± 0.08	0.88 ± 0.07	1.86 ± 0.10	ND	ND	0.73
Q369L	37.2 ± 0.2	0.08 ± 0.01	0.90 ± 0.13	0.09 ± 0.01	ND	ND	0.69
A373C	34.3 ± 0.1	1.23 ± 0.11	0.93 ± 0.13	1.33 ± 0.09	0.56 ± 0.00	14.43 ± 1.14	0.83
LGK.2	42.1 ± 0.1	0.63 ± 0.05	1.38 ± 0.31	0.46 ± 0.07	-	5.12 ± 2.69	-
LGK.1	38.9 ± 0.1	0.89 ± 0.07	0.84 ± 0.13	1.07 ± 0.08	-	30.91 ± 3.43	-

Table A 2.4: Specific growth rates of *E. coli* Tuner expressing plasmid pJK_proJK1_LGK versus pJK_proJK1_LGK.1.
Isogenic cultures were grown aerobically at 37°C in M9 minimal media with carbenicillin with the specified LG concentration. After a certain point the specific growth rate of the culture expressing LGK.1 was uncoupled to flux as the LG within the growth media was increased. Data represent mean and SD (n≥2). ND = Not determined.

LG Concentration (g/L)	pJK_proJK1_LGK μ (h^{-1})	pJK_proJK1_LGK.1 μ (h^{-1})
2	0.00 ± 0.00	0.38 ± 0.00
4	0.02 ± 0.02	0.55 ± 0.00
10	0.21 ± 0.01	0.04 ± 0.00
20	ND	0.02 ± 0.00
24	0.40 ± 0.01	ND

Table A 2.5: Number of residues with mutations that improve growth rates equal to or greater than 20% relative to the starting sequence as a function of the fraction ASA buried. The analysis excludes the first nine residues as they are not present within the crystal structure. Fraction ASA buried was calculated by calculating the DSSP ACC area[64] and dividing by the individual residue tripeptide ASA[65].

Fraction ASA Buried	Total Number of Residues	Improved in First Selection	Improved in Second Selection
0.00-0.35	37	8	10
0.36-0.75	110	35	19
0.76-1.00	283	50	9

Table A 2.6: List of mutations in each LGK design.

Design	LGK.3	LGK.3.1	LGK.3.2	LGK.3.3	LGK.3.4	LGK.4	LGK.5	LGK.6	LGK.7	LGK.8	LGK.9
WT Residue and Number	Mutation										
T7									E	E	
V11	P	P	P	P		P	S				
T15								R			R
C30	L	L	L	L		L					
K38							E		A	A	
T39							N				
D41							T				
L56	W	W	W			M					
A57							P	Q			Q
P59						S	E				
Q62							K				
L69	Q	Q	Q	Q		Q	Q				
E70							T				
D71						E	N				
T72							R		S	S	
P75	L	L	L	L		L					
S76							Q				
S79							A				
I84							L				
H88	T	T	T	T		T					
V93									A	A	
R94	H	H	H	H		H		G	H	H	G
Q95							I				
A98							K				
R100						K		C			C
N101								G			G

Table A 2.6 (cont'd)

D103							S				
S105								R	A	A	R
T106						S					
A111	G	G	G	G		G					
Q115						V					
T116							C				
L120	N	N	N	N		N	Q		N	N	
E124							G				
A131	C	C	C	C							
A135	G	G	G			G	G				
L140	I	I	I	I		I	I	I	I	I	I
S142	A	A	A	A		A	A	A	A	A	A
S148	T	T	T	T			A		C	C	
Q157	I	I	I	I		I					
I167	H	H	H				Q				
C194	T	T	T			T	T				
D199							C				
R204	L	L	L	L		K	M		L	L	
T205						S					
Y209							F				
N217	T	T	T	T		T	M				
V218							I				
F219	M	M	M	M		M					
F227									Y		Y
Q232								C		C	
A239	E	E	E	E	E	E	E				
K242							A				
K245								V		V	
M257	H	H	H	H	H	H	H				

Table A 2.6 (cont'd)

T268	C		C		C						
T276	S	S	S	S	S	S	S				
H279							Y		W		W
R284							K				
E286								Q		Q	
D293							E				
T299	V	V	V	V	V	V			V		V
A306	S	S	S	S	S						
H310	A	A	A	A	A						
R312							K				
Y314								F		F	
E319							D				
M325	L	L	L	L	L	L	L				
Y331							N				
S342	A	A	A	A	A	A	A	K		K	
M349								I		I	
A354								C		C	
V356	I	I	I	I	I						
G359	R	R	R	R	R	D	R		R		R
Q369	L	L		L	L						
A373	C	C	C	C	C	C		C	C	C	C
T383	S	S	S	S	S						
Q389							K				
H390	G	G	G	G	G		P		P		P
Y391							C				
V396	I	I	I	I	I	I	I				
L400							K				
S404							K				
Q417								N		N	

Table A 2.6 (cont'd)

W421	P	P	P	P	P	P					
E424								A		A	
I426								V		V	
K430							N	E		E	
I434								K		K	
T435								H		H	
N437						G	K				
W438						G					
A439	D	D	D	D	D		D	K	E	K	E

Table A 2.7: Relative activity of LGK backcross designs removing mutations from design LGK.3. Back cross variants of design LGK.3 were tested for activity in cellular lysates after autoinduction. Activity was compared to a cellular lysate of a control culture expressing wild-type LGK. Each construct was expressed in a minimum of 8 cultures each. LGK.3.4 was purified and tested against purified LGK *in vitro*. All assays were tested with a LG concentration of 550 mM.

Variant	Total number of mutations	Relative Activity
LGK.3.1-His_6	37	No Activity
LGK.3.2-His_6	37	No Activity
LGK.3.3-His_6	33	No Activity
LGK.3.4-His_6	18	<10% activity of wild-type

Table A 2.8: PCR primers used to amplify out gene tiles for deep sequencing. The gene was segmented into 11 separate tiles for the 1st selection and 6 separate tiles for the 2nd selection.

Sequence Name	Sequence
TILE1_FWD	GTTCAGAGTTCTACAGTCCGACGATCTTAACTTTAAGAAGGAGATATACAT
TILE1_REV	CCTTGGCACCCGAGAATTCCATCCATCGGAGCATC
TILE2_FWD	GTTCAGAGTTCTACAGTCCGACGATCCTATCAGAAAACCCCG
TILE2_REV	CCTTGGCACCCGAGAATTCCACCAGAATCACGTTGAC
TILE3_FWD	GTTCAGAGTTCTACAGTCCGACGATCCGTCTGAACTGTCAGAA
TILE3_REV	CCTTGGCACCCGAGAATTCCACTTCTTCCGGCATACT
TILE4_FWD	GTTCAGAGTTCTACAGTCCGACGATCCGATTTGGCTGCTG
TILE4_REV	CCTTGGCACCCGAGAATTCCAGGTGCACCTTGACG
TILE5_FWD	GTTCAGAGTTCTACAGTCCGACGATCCCAGGCAGCTGGT
TILE5_REV	CCTTGGCACCCGAGAATTCCACGACGGCCGTC
TILE6_FWD	GTTCAGAGTTCTACAGTCCGACGATCCCGCCGGATGTT
TILE6_REV	CCTTGGCACCCGAGAATTCCACCGCGTTTGCC
TILE7_FWD	GTTCAGAGTTCTACAGTCCGACGATCGATAAAGACGGTGCTATG
TILE7_REV	CCTTGGCACCCGAGAATTCCACGCGACGAATCAG
TILE8_FWD	GTTCAGAGTTCTACAGTCCGACGATCCCCTGGCACATGAC
TILE8_REV	CCTTGGCACCCGAGAATTCCAGCACATGAAGATTTCGTC
TILE9_FWD	GTTCAGAGTTCTACAGTCCGACGATCGCCGTCACAAGAAATT
TILE9_REV	CCTTGGCACCCGAGAATTCCAGAACGTAATCGCTTCTTT
TILE10_FWD	GTTAGAGTTCTACAGTCCGACGATCCGGCAGGTGCA
TILE10_REV	CCTTGGCACCCGAGAATTCCAATCACAGAGCGGTAATT
TILE11_FWD	GTTCAGAGTTCTACAGTCCGACGATCAGCCCGGGTCTG
TILE11_REV	CCTTGGCACCCGAGAATTCCATGGTGGTGCTCGAG

Table A 2.9: Crystallographic data processing and model refinement statistics for LGK.3 (PDB: 4ZXZ).

Data Collection	
Space group	$P2_12_12_1$
Unit cell (Å)	a = 85.18; b = 88.95, c = 139.46 $\alpha = \beta = \gamma = 90.00$
Wavelength (Å)	1.54180
Resolution range (Å)	46.13 – 2.20 (2.32 – 2.20)
Total observations	178822
Total unique observations	49339
I/σ_I	6.9 (1.6)
Completeness (%)	90.9 (79.0)
R_{merge}	0.147 (0.765)
R_{pim}	0.082 (0.506)
Multiplicity	3.6 (2.9)
Refinement Statistics	
Resolution (Å)	46.13-2.20
Reflections (total)	49264
Reflections (test)	2505
Total atoms refined	7034
Solvent	394
R_{work} (R_{free})	0.17 (0.21)
RMSDs Bond lengths (Å) / angles (°)	0.009/1.128
Ramachandran plot Favored/allowed (%)	98.1/1.9
Mean B values ($Å^2$; chain A/B)	30.0/29.8

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., and Lee, S. Y. (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals, *Nature Chemical Biology 8*, 536-546.

[2] Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K. W., Fickes, S., Galazzo, J., Gaucher, S. P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Jiang, H., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L. F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P. S., Keasling, J. D., Reiling, K. K., Renninger, N. S., and Newman, J. D. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin, *Nature 496*, 528-532.

[3] Chang, M. C. Y., Eachus, R. A., Trieu, W., Ro, D. K., and Keasling, J. D. (2007) Engineering Escherichia coli for production of functionalized terpenoids using plant P450s, *Nature Chemical Biology 3*, 274-277.

[4] Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., and Van Dien, S. (2011) Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol, *Nature Chemical Biology 7*, 445-452.

[5] Zhang, K., Li, H., Cho, K. M., and Liao, J. C. (2010) Expanding metabolism for total biosynthesis of the nonnatural amino acid L-homoalanine, *Proceedings of the National Academy of Sciences of the United States of America 107*, 6234-6239.

[6] Zhang, K., Sawaya, M. R., Eisenberg, D. S., and Liao, J. C. (2008) Expanding metabolism for biosynthesis of nonnatural alcohols, *Proceedings of the National Academy of Sciences of the United States of America 105*, 20653-20658.

[7] Zelcbuch, L., Antonovsky, N., Bar-Even, A., Levin-Karp, A., Barenholz, U., Dayagi, M., Liebermeister, W., Flamholz, A., Noor, E., Amram, S., Brandis, A., Bareia, T., Yofe, I., Jubran, H., and Milo, R. (2013) Spanning high-dimensional expression space using ribosome-binding site combinatorics, *Nucleic Acids Research 41*.

[8] Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J., and Dueber, J. E. (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay, *Nucleic Acids Research 41*, 10668-10678.

[9] Latimer, L. N., Lee, M. E., Medina-Cleghorn, D., Kohnz, R. A., Nomura, D. K., and Dueber, J. E. (2014) Employing a combinatorial expression approach to characterize xylose utilization in Saccharomyces cerevisiae, *Metabolic engineering 25*, 20-29.

[10] Du, J., Yuan, Y. B., Si, T., Lian, J. Z., and Zhao, H. M. (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering, *Nucleic Acids Research 40*.

[11] Woo, H. M., Murray, G. W., Batth, T. S., Prasad, N., Adams, P. D., Keasling, J. D., Petzold, C. J., and Lee, T. S. (2013) Application of targeted proteomics and biological parts assembly in E. coli to optimize the biosynthesis of an anti-malarial drug precursor, amorpha-4,11-diene, *Chem Eng Sci 103*, 21-28.

[12] Bond-Watts, B. B., Bellerose, R. J., and Chang, M. C. Y. (2011) Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways, *Nature Chemical Biology 7*, 222-227.

[13] Leonard, E., Ajikumar, P. K., Thayer, K., Xiao, W.-H., Mo, J. D., Tidor, B., Stephanopoulos, G., and Prather, K. L. J. (2010) Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control, *Proceedings of the National Academy of Sciences of the United States of America 107*, 13654-13659.

[14] Araya, C. L., and Fowler, D. M. (2011) Deep mutational scanning: assessing protein function on a massive scale, *Trends in biotechnology 29*, 435-442.

[15] Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. (2010) High-resolution mapping of protein sequence-function relationships, *Nature methods 7*, 741-746.

[16] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N., and Whitehead, T. A. (2015) High-resolution sequence-function mapping of full-length proteins, *PLoS One 10*, e0118193.

[17] Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., and Baker, D. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing, *Nature biotechnology 30*, 543-548.

[18] Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate, *Journal of molecular biology 425*, 1363-1377.

[19] Fowler, D. M., and Fields, S. (2014) Deep mutational scanning: a new style of protein science, *Nature methods 11*, 801-807.

[20] Walkiewicz, K., Benitez Cardenas, A. S., Sun, C., Bacorn, C., Saxer, G., and Shamoo, Y. (2012) Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance, *Proceedings of the National Academy of Sciences of the United States of America 109*, 21408-21413.

[21] Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape, *Molecular Biology and Evolution 31*, 1581-1592.

[22] Stiffler, Michael A., Hekstra, Doeke R., and Ranganathan, R. (2015) Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase, *Cell 160*, 882-892.

[23] Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli, *Proceedings of the National Academy of Sciences of the United States of America 110*, 14024-14029.

[24] Noderer, W. L., Flockhart, R. J., Bhaduri, A., Diaz de Arce, A. J., Zhang, J., Khavari, P. A., and Wang, C. L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq, *Molecular Systems Biology 10*.

[25] Antal, M. J., and Varhegyi, G. (1995) Cellulose Pyrolysis Kinetics - the Current State Knowledge, *Ind Eng Chem Res 34*, 703-717.

[26] Bennett, N. M., Helle, S. S., and Duff, S. J. B. (2009) Extraction and hydrolysis of levoglucosan from pyrolysis oil, *Bioresource Technology 100*, 6059-6063.

[27] Prosen, E. M., Radlein, D., Piskorz, J., Scott, D. S., and Legge, R. L. (1993) Microbial Utilization of Levoglucosan in Wood Pyrolysate as a Carbon and Energy-Source, *Biotechnol Bioeng 42*, 538-541.

[28] Jarboe, L. R., Wen, Z. Y., Choi, D. W., and Brown, R. C. (2011) Hybrid thermochemical processing: fermentation of pyrolysis-derived bio-oil, *Appl Microbiol Biot 91*, 1519-1523.

[29] Dai, J., Yu, Z., He, Y., Zhang, L., Bai, Z., Dong, Z., Du, Y., and Zhang, H. (2009) Cloning of a novel levoglucosan kinase gene from Lipomyces starkeyi and its expression in Escherichia coli, *World Journal of Microbiology and Biotechnology 25*, 1589-1595.

[30] Bacik, J. P., Klesmith, J. R., Whitehead, T. A., Jarboe, L. R., Unkefer, C. J., Mark, B. L., and Michalczyk, R. (2015) Producing glucose-6-phosphate from cellulosic biomass: structural insights into levoglucosan bioconversion, *Journal of Biological Chemistry*.

[31] Layton, D. S., Ajjarapu, A., Choi, D. W., and Jarboe, L. R. (2011) Engineering ethanologenic Escherichia coli for levoglucosan utilization, *Bioresour Technol 102*, 8318-8322.

[32] Bacik, J.-P., Klesmith, J. R., Whitehead, T. A., Jarboe, L. R., Unkefer, C. J., Mark, B. L., and Michalczyk, R. (2015) Producing Glucose 6-Phosphate from Cellulosic Biomass: STRUCTURAL INSIGHTS INTO LEVOGLUCOSAN BIOCONVERSION, *Journal of Biological Chemistry 290*, 26638-26648.

[33] Bacik, J.-P., Whitworth, G. E., Stubbs, K. A., Yadav, A. K., Martin, D. R., Bailey-Elkin, B. A., Vocadlo, D. J., and Mark, B. L. (2011) Molecular Basis of 1,6-Anhydro Bond

Cleavage and Phosphoryl Transfer by Pseudomonas aeruginosa 1,6-Anhydro-N-acetylmuramic Acid Kinase, *Journal of Biological Chemistry 286*, 12283-12291.

[34] Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E., Tomek, K. J., and Whitehead, T. A. (2014) The interrelationship between promoter strength, gene expression, and growth rate, *PLoS One 9*, e109105.

[35] Davis, J. H., Rubin, A. J., and Sauer, R. T. (2011) Design, construction and characterization of a set of insulated bacterial promoters, *Nucleic Acids Research 39*, 1131-1141.

[36] Firnberg, E., and Ostermeier, M. (2012) PFunkel: efficient, expansive, user-defined mutagenesis, *PLoS One 7*, e52031.

[37] Kwon, W. S., Da Silva, N. A., and Kellis, J. T. (1996) Relationship between thermal stability, degradation rate and expression yield of barnase variants in the periplasm of Escherichia coli, *Protein Engineering 9*, 1197-1202.

[38] Parsell, D. A., and Sauer, R. T. (1989) The structural stability of a protein is an important determinant of its proteolytic susceptibility in Escherichia coli, *Journal of Biological Chemistry 264*, 7590-7595.

[39] Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability, *Proceedings of the National Academy of Sciences of the United States of America 103*, 5869-5874.

[40] Romero, P. A., and Arnold, F. H. (2009) Exploring protein fitness landscapes by directed evolution, *Nature reviews. Molecular cell biology 10*, 866-876.

[41] Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability, *Current opinion in structural biology 19*, 596-604.

[42] Solomon, K. V., Moon, T. S., Ma, B., Sanders, T. M., and Prather, K. L. J. (2013) Tuning Primary Metabolism for Heterologous Pathway Productivity, *ACS Synthetic Biology 2*, 126-135.

[43] Kadner, R. J., Murphy, G. P., and Stephens, C. M. (1992) Two mechanisms for growth inhibition by elevated transport of sugar phosphates in Escherichia coli, *Microbiology 138*, 2007-2014.

[44] Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures, *Proceedings of the National Academy of Sciences of the United States of America 97*, 10383-10388.

[45] Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite, *PLoS ONE 6*, e20161.

[46] Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnol Bioeng 84*, 647-657.

[47] Trinh, C. T., Unrean, P., and Srienc, F. (2008) Minimal Escherichia coli Cell for the Most Efficient Production of Ethanol from Hexoses and Pentoses, *Applied and Environmental Microbiology 74*, 3634-3643.

[48] Tang, S.-Y., and Cirino, P. C. (2011) Design and Application of a Mevalonate-Responsive Regulatory Protein, *Angewandte Chemie International Edition 50*, 1084-1086.

[49] Stapleton, J. A., Kim, J., Hamilton, J. P., Wu, M., Irber, L. C., Maddamsetti, R., Briney, B., Newton, L., Burton, D. R., Brown, C. T., Chan, C., Buell, C. R., and Whitehead, T. A. (2016) Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing, *PLoS ONE 11*, e0147229.

[50] Hong, L., Hong, S., Wong, H., Aw, P., Cheng, Y., Wilm, A., de Sessions, P., Lim, S., Nagarajan, N., Hibberd, M., Quake, S., and Burkholder, W. (2014) BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads, *Genome Biology 15*, 517.

[51] Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases, *Nat Meth 6*, 343-345.

[52] Kunkel, T. A. (1985) Rapid and efficient site-specific mutagenesis without phenotypic selection, *Proceedings of the National Academy of Sciences 82*, 488-492.

[53] Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants, *Bioinformatics 27*, 3430-3431.

[54] Studier, F. W. (2005) Protein production by auto-induction in high-density shaking cultures, *Protein Expression and Purification 41*, 207-234.

[55] Hirai, M., Ohtani, E., Tanaka, A., and Fukui, S. (1977) Glucose-phosphorylating enzymes of Candida yeasts and their regulation in vivo, *Biochim Biophys Acta 480*, 357-366.

[56] Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N., and Nordlund, P. (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies, *Analytical Biochemistry 357*, 289-298.

[57] Lavinder, J. J., Hari, S. B., Sullivan, B. J., and Magliery, T. J. (2009) High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering, *Journal of the American Chemical Society 131*, 3794-3795.

[58] Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. In Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography, *Newsletter on Protein Crystallography 26*.

[59] Evans, P. (2006) Scaling and assessment of data quality, *Acta Crystallographica Section D 62*, 72-82.

[60] McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C., and Read, R. J. (2005) Likelihood-enhanced fast translation functions, *Acta Crystallogr D Biol Crystallogr 61*, 458-464.

[61] Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution, *Acta Crystallographica Section D 66*, 213-221.

[62] Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics, *Acta Crystallogr D Biol Crystallogr 60*, 2126-2132.

[63] Clark, D. S., and Blanch, H. W. (1997) *Biochemical engineering*, CRC Press.

[64] Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577-2637.

[65] Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins, *Journal of molecular biology 105*, 1-12.

CHAPTER 3

Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning

John-Paul Bacik and Ryszard Michalczyk carried out the protein crystallization in this work at Los Alamos National Laboratory.

ABSTRACT

Proteins are marginally stable, and an understanding of the sequence determinants for improved protein solubility is highly desired. For enzymes, it is well known that many mutations that increase protein solubility decrease catalytic activity. These competing effects frustrate efforts to design and engineer stable, active enzymes without laborious high-throughput activity screens. To address the trade-off between enzyme solubility and activity, I performed deep mutational scanning using three different, complementary solubility screens/selections for two full-length enzymes. I assayed a TEM-1 beta lactamase variant and levoglucosan kinase using yeast surface display screening, GFP fusion screening, and a twin-arginine translocation pathway selection. I then compared these scans with previously determined experimental fitness landscapes. There was significant correlation between solubility and fitness datasets, although the correlation coefficients were very low. 5-10% of all single missense mutations improve solubility, matching theoretical predictions of global protein stability. For a given solubility-enhancing mutation, the probability that it would retain wild-type fitness was correlated with evolutionary conservation and distance to active site, and anti-correlated with contact number. Hybrid classification models were developed that could predict solubility-enhancing mutations that maintain wild-type fitness with an accuracy of 90%. The downside of using such classification models is the removal of rare, Pareto optimal mutations that improve both fitness and solubility. In order to reveal the biophysical basis of enhanced protein solubility and function, the crystallographic structure of one such LGK mutant was determined. Beyond fundamental insights into trade-offs between stability and activity, these results have potential biotechnological applications.

INTRODUCTION

Solubility is a fundamental biophysical property of proteins. In this work I define solubility as a function of the protein aggregation propensity[1, 2] and thermodynamic stability. Understanding the distribution of solubility-modulating mutations can sharpen biophysical models undergirding evolutionary theories combining molecular evolution and population genetics. There are also a number of biotechnological applications: improving protein solubility can enhance the total turnover number for biocatalysts, increase expression yield of enzymes needed in biomanufacturing, or bolster formulation stability of therapeutic proteins. From this applied perspective, general approaches are desired to identify mutations that improve the solubility of a protein whilst maintaining function.

Computational approaches have been used to evaluate and design thermodynamic stability[3-5] and aggregation propensity[6-10]. There also exist several high-throughput experimental screens that can be used to increase soluble protein expression and protein solubility[11-17]. A major challenge for the above approaches is that solubility-enhancing mutations often reduce the specific activity of enzymes[18-22]. Additionally, because the stabilizing effect of most beneficial mutations is modest[10, 23, 24], many mutations from the starting sequence are typically needed to increase solubility over wild type. Together, these facts necessitate running a secondary screen for activity for positive hits from the solubility screen[25], increasing time and effort.

Comprehensive evaluation of the trade-off between enzyme activity and solubility could identify classifiers used to predict whether a given solubility-enhancing mutation is deleterious for enzyme activity. For example, earlier directed evolution experiments have shown that solubility-enhancing mutations are enriched at or near active site residues – most such mutations are deleterious for activity[26-28]. Additionally, a powerful enzyme engineering approach is to

choose only those mutations that have been oversampled in the evolutionary history of the protein family[29-31]. This 'back-to-consensus' strategy rests on the supposition that mutations to the consensus sequence of the protein family maintain enzyme function and improve stability. Other classifiers beyond the above two may exist.

In this work I used deep mutational scanning[32-39] to assess the sequence determinants of solubility for two different full-length enzymes. Enzymes with known fitness landscapes using experimentally derived functional selections[18, 36] were studied, allowing a direct comparison between protein fitness and solubility. Three existing complementary *in vivo* high-throughput solubility screens/selections were assessed for the ability to identify mutations that confer solubility. For one enzyme the fraction of solubility-enhancing mutations was between 4-5%, which is in line with theoretical predictions[40]. Several limitations in commonly used *in vivo* screens were identified, which should reduce false positives and false negatives in future high-throughput datasets. Classification methods were developed to select mutations that improve solubility without impacting fitness with an accuracy of 90%. Notably, these classifiers do not require a high-resolution protein structure or homology model. A structure of a Pareto optimal enzyme variant was solved to show the biophysical basis of enhanced solubility and function. Together, these results provide experimental illustration of the trade-off between enzyme solubility and function, and provide a means by which active, stable mutants can be uncovered without a high throughput activity screen.

RESULTS

Deep Mutational Scanning for Solubility

Deep mutational scanning was performed on two full-length enzymes: a 263-residue TEM-1 beta-lactamase variant and a 439-residue levoglucosan kinase (LGK) using different *in vivo* high-throughput solubility screens/selections (**Figure 3.1**). For TEM-1 BLA, catalytic activity was abolished by mutating the active site residue Ser70 to Ala because one selection involves growth on beta-lactam antibiotics. The destabilizing mutation D179G[27, 41] was introduced because TEM-1 is stable at the selection and screening temperatures of 30-37ºC. I refer to this resulting construct (TEM-1 S70A, D179G) as TEM-1.1 in the remainder of the work. Comprehensive single-site saturation mutagenesis libraries were constructed in all genetic backgrounds using Nicking Mutagenesis[42], and libraries were harvested, prepared, and deep sequenced in a standardized pipeline[43] (**Figure B 3.1**).

I tested three previously developed solubility screens (**Figure 3.1**). In yeast surface display (YSD)[44], proteins are fused in-frame with a C-terminal epitope tag and an N-terminal Aga2p domain that localizes the fusion protein to the outer cell surface. Binding with a fluorescently conjugated anti-epitope antibody allows discrimination of variants that express on the cell surface from ones that cannot. I used fluorescence activated cell sorting (FACS) to collect a reference population of all yeast and the top 5% of displaying population determined by fluorescence intensity. For the second FACS-based screen I fused LGK to fluorescent protein variant mGFPmut3[45], induced fusion protein expression by IPTG in *E. coli*, and sorted the library by fluorescence intensity. Sort statistics are given in **Table B 3.1**.

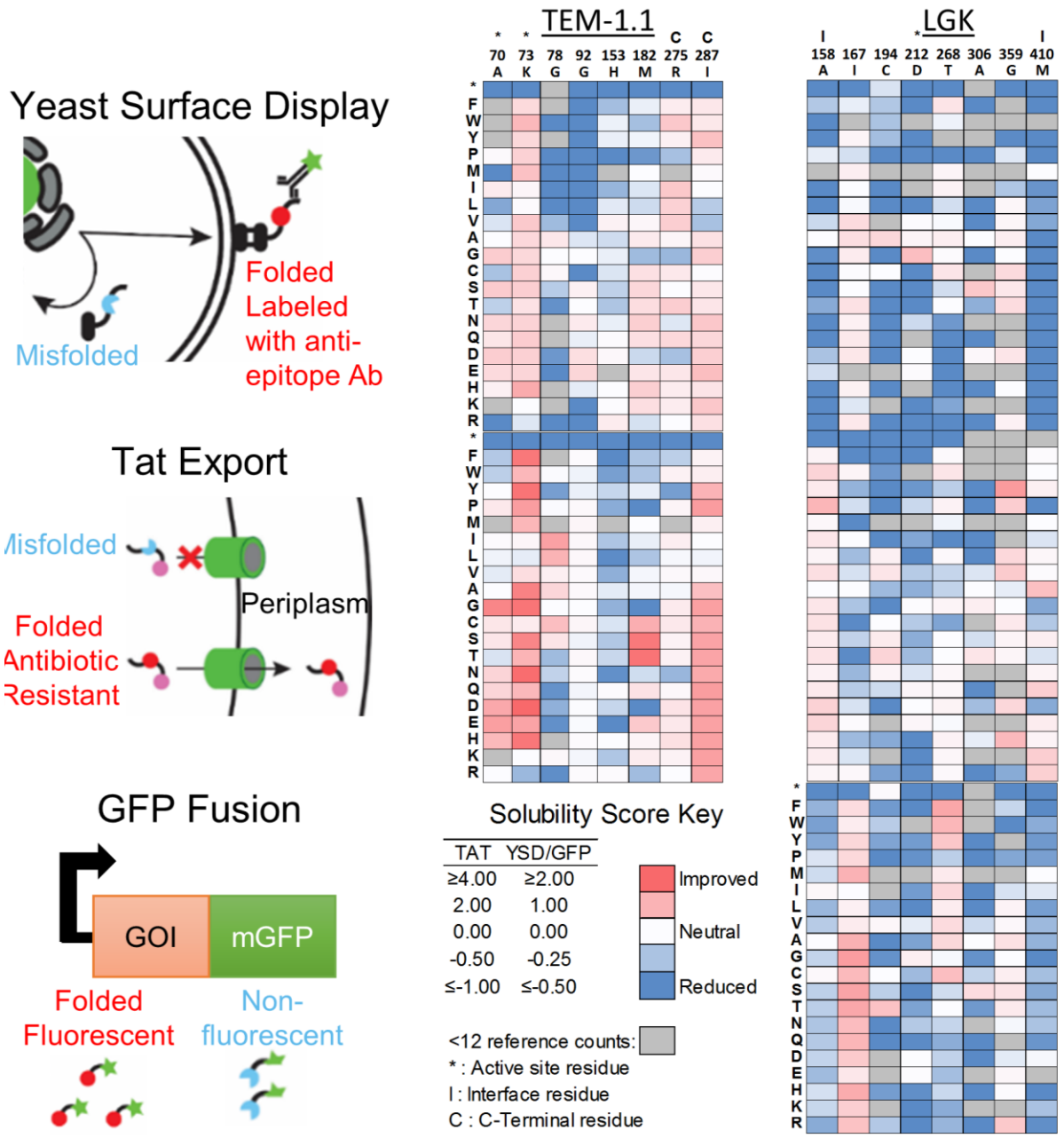


Figure 3.1: **Overview of solubility deep mutational scans for TEM-1.1 and LGK.** The first column shows different screens used in the present work. In yeast surface display (YSD) the protein is exported to the surface and labeled by a florescent antibody that is specific for a c-terminal epitope tag. The top 5% of cells by fluorescence intensity are collected by FACS. For TAT export, a protein is fused to a C-terminal beta-lactamase that requires periplasm localization for activity. Variants are selected on plates containing high antibiotic concentrations. In the last screen, a GFP variant is fused to the protein of interest, expression induced in bacteria, and the top 5% of cells by fluorescence intensity are screened by FACS. The second and third columns show heatmaps of solubility scores for selected residues of TEM-1.1 and LGK. Active site

Figure 3.1 (cont'd) residues are indicated by (*), interface residues by (I), and residues proximal to the C-terminus by (C).

I also performed a twin-arginine translocation (Tat) export selection[46]. A protein of interest is fused in-frame with a N-terminal ssTorA Tat periplasmic export signal peptide and a C-terminal, TEM-1 BLA with a deleted Sec export signal sequence[46]. *E. coli* expressing this fusion protein will survive in the presence of beta-lactam antibiotics if the fusion protein is present in the periplasm, as TEM-1 BLA activity is dependent on formation of a disulfide bond. I prepared a selection plasmid using a codon-swapped C-terminal Δ4-25 TEM-1 designed to minimize recombination during experiments performed on TEM-1.1.

There was on average 93.2% (4,918/5,260 (TEM-1.1) and 7,952/8,560 (LGK)) coverage of single non-synonymous mutations identified across all libraries. 5,466 (67.2%) single amino acid mutations were present in all three LGK screens, while 3,690 (73.8%) single amino acid substitutions were shared in both TEM-1 screens. Enrichment ratios calculated from deep sequencing were converted to a solubility score centered about a wild-type score of zero. The per-position scores are visualized in heatmaps shown in **Figure 3.1** and **Figures B 3.2-3.6**. Detailed statistics for each deep mutational scan are in **Tables B 3.2-3.4**.

Validation of solubility datasets

I evaluated the ability of the screens to select for higher solubility variants and deplete low solubility mutants in several ways. First, all five solubility datasets showed statistically significant reductions in nonsense compared with missense mutations ($p<0.0001$; **Figure 3.2a, Figure B 3.7**). Second, I reasoned that there would be correlation between these solubility scores and existing deep mutational scanning fitness datasets for LGK[18] and TEM-1[36], as enzyme fitness is subject to the biophysical constraints of folding. The correlation between solubility and

fitness datasets were statistically significant ($p < 10^{-17}$ in all cases) but the overall correlation coefficients were very low (**Figures B 3.8-3.9**). Third, the fraction of tolerable residues at each position (see Methods) was negatively correlated with contact number (average number of neighboring residues – a measure of packing density)[47] for LGK and a subset of the TEM-1.1 datasets (positions 61-215 using the Ambler sequence convention - see below for justification) (**Figure 3.2b, Figure B 3.10**).

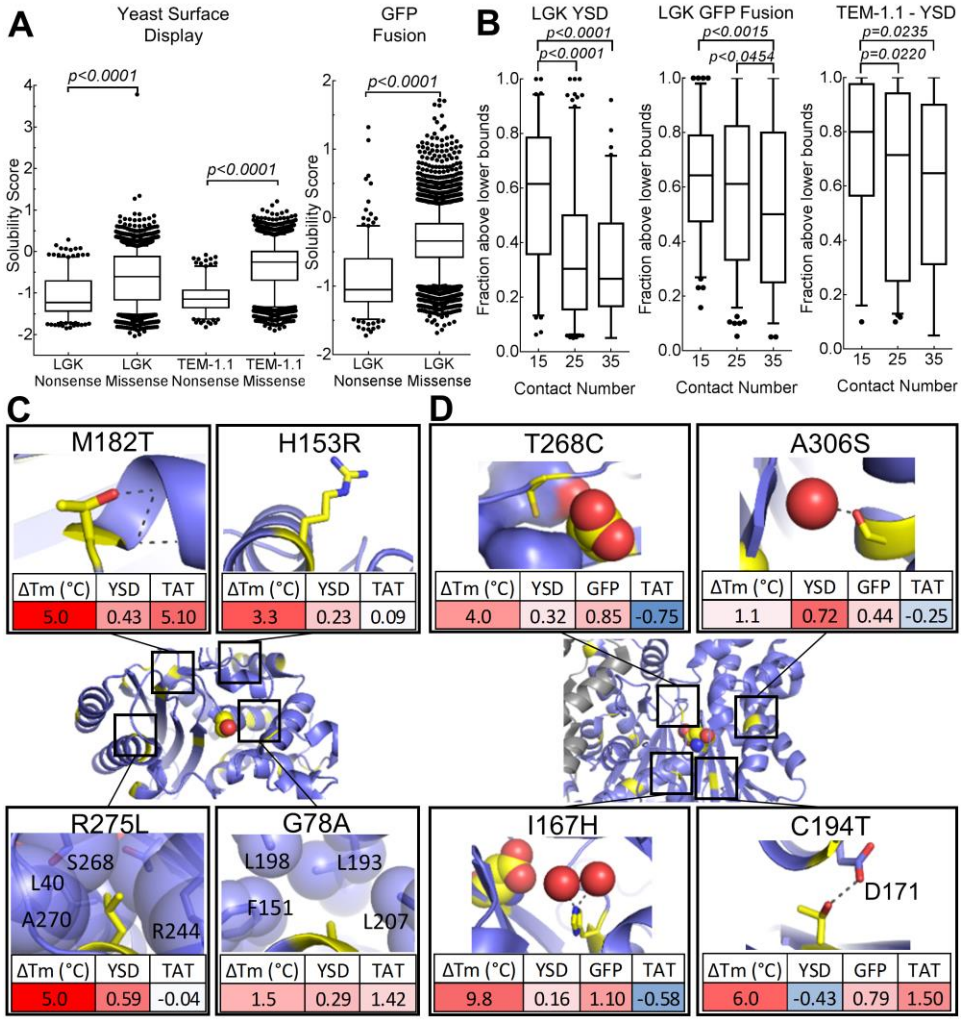


Figure 3.2: Validation of solubility datasets. a) Nonsense vs. missense solubility scores for YSD (LGK, TEM-1.1) and GFP fusion (LGK). b) The fraction of beneficial mutations above the lower bounds versus contact number for LGK and TEM-1.1 (residues 61 to 215). c,d) Known stabilizing mutations (yellow) mapped onto (c) TEM-1.1 (PDB ID: 1M40) and (d) LGK (PDB: 4ZLU). Insets show the structural basis of the stabilizing mutations, shown as yellow sticks, along with the corresponding solubility scores identified by deep sequencing.

I also evaluated the ability of the solubility deep mutational scans to identify known stabilizing mutations in TEM-1 (**Figure 3.2c, Table B 3.5**) and LGK (**Figure 3.2d, Table B 3.6**) that are located at the surface, core, and, for LGK, at the homodimer interface. I identified a mutation as solubility enhancing if its solubility score was above 0.15 – for screens utilizing FACS this value corresponds to a mean fluorescence intensity of 10% above the wild-type sequence. For TEM-1, 15/19 mutations with an *in vitro* characterized change in melting temperature (ΔT_m) $\geq$ 1^{o}C were recorded as solubility enhancing in the YSD dataset (Fisher's exact test p-value <0.0001). For the LGK datasets, the GFP fusion screens identified 9/12 of these mutations (p-value <0.0001) whereas YSD identified 6/11 (p-value <0.0001). 11/12 of the false negatives from the YSD and GFP fusion screens were just below the significance cut-off used. The notable exception was LGK C194T, which had a very low YSD solubility score (**Fig. 3.2d**). Since there is an ASN at position 192 that is surface exposed but in the catalytic active site, THR194 introduces a potential N-linked glycosylation site. I speculate that this aberrant glycosylation at the active site results in misfolded protein that would be retained in the ER. I conclude that YSD and GFP fusion solubility screens are able to identify gain of thermodynamic stability variants.

By contrast, in my hands the TAT selection solubility datasets identify 3/10 (p-value 0.42) TEM-1 and 1/12 (p-value 0.32) LGK known stabilizing mutations, although it is noted that the very stabilizing mutations TEM-1 M182T (ΔT_m=5^{o}C) and LGK C194T (ΔT_m=6^{o}C) were strongly enriched. The inability of the TAT screen to enrich known stability-enhancing mutations may reflect the fact that the selection criteria for TAT export are not dominated by protein thermal stability.

Distribution of Solubility Scores

What is the distribution of solubility scores for the two enzymes? Here, I restricted my analysis to GFP and YSD screens because of the number of false negatives observed for the TAT pathway selection. The distributions for LGK and TEM-1.1 are multi-modal with the mean value below the wild-type solubility score of zero (**Figure 3.3a**). For the LGK datasets, 4.5% (YSD) and 6.0% (GFP) of possible single missense mutations were above a fitness metric of 0.15. However, 14.5% of mutations were identified in the TEM-1.1 dataset using the same criteria. The numbers reported above may overestimate the percentage of solubility-enhancing mutations if the number of false positives outweighs the number of false negatives, and vice versa.

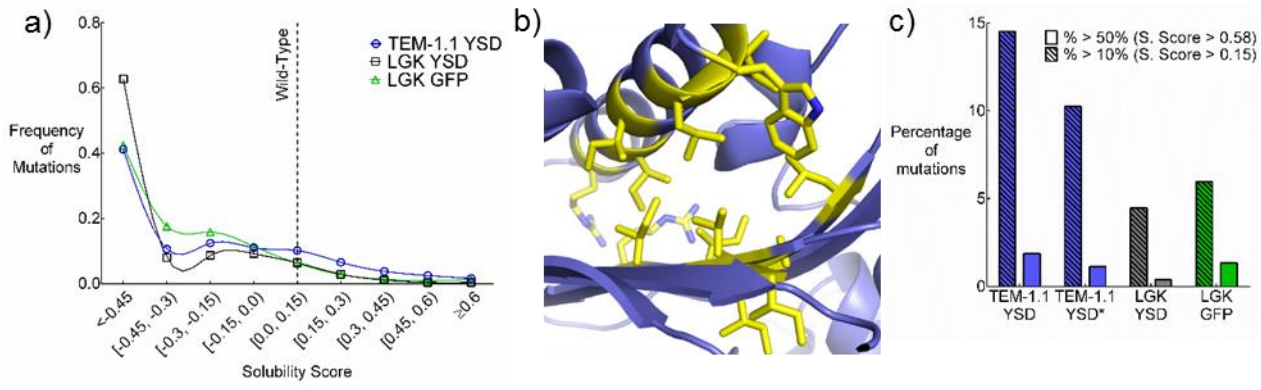


Figure 3.3: Distribution of solubility-enhancing mutations. a) Frequency of mutations for TEM-1.1 YSD (blue), LGK YSD (black), and LGK GFP (green) found at each fitness metric level. Each dataset is fit with a cubic spline to help guide the eye. b) Positions with more than 10 beneficial mutations in TEM-1.1 YSD dataset are shown as yellow sticks. These false positives are predicted to disrupt the C-terminal helix, presumably to promote accessibility of the c-myc epitope tag. c) The percentage of mutations with solubility scores above a 10% (hatched fill) and 50% increase (solid fill) in function for TEM-1.1 YSD (blue), LGK YSD (gray), and LGK GFP (green). TEM-1.1 YSD* covers residues 61 to 215 to remove the section with false positives indicated in panel B.

To identify potential false positives/negatives in the LGK solubility datasets, I compared results from GFP and YSD screens. While only 20% (113/574) of solubility-enhancing mutations were shared between the GFP and YSD screens, most mutations beneficial for one selection were slightly below the cutoff in the other selection (**Figure B 3.11**). However, there

were outliers. 22 mutations (4%) were positive in YSD and strongly negative in GFP fusion screen, with 36 (6%) for the reverse case. No statistically significant single metric was found for these outliers except that the mutations negative in YSD and positive in GFP fusion screen were more buried than all beneficial mutations (average 95% SASA burial; Fisher's exact test p-value=4.6×10^{-15}). These outliers were, on average, strongly deleterious in fitness according to a previous deep mutational scanning dataset[18]. Based on this consideration, I speculate that these outliers are false positives resulting from the GFP fusion screen.

Given that the percentage of solubility-enhancing mutations observed in TEM-1.1 dataset exceeds theoretical estimates for purely thermodynamically stabilizing mutations[40], I diagnosed potential problems with the screens. In the YSD screen, most positions that allow ten or more of these substitutions map to the C-terminus (**Figure 3.3b**). Because the N- and C-termini are so close to one another, and I did not include a linker region between the C-terminus and the myc epitope tag, I speculate that mutations enriched from this screen destabilized the helix positioning to avoid steric clashes between the fluorescently conjugated anti-myc antibody and either the N-terminal Aga2p or TEM-1.1. Thus, the YSD screen awards mutations that enhance antibody binding at the expense of core destabilization. Restricting analysis to portions of the protein not affected by this set of false positives (Ambler positions 61-215) results in 10.3% of fitness enhancing mutations, still higher than the theoretical estimates but closer to the solubility score distributions found in the LGK dataset (**Figure 3.3c**).

Restricting mutational search space to variants represented in the evolutionary history of the enzyme family is a proven stabilization strategy in protein engineering ("back-to-consensus") [29-31]. I calculated the proportion of solubility and fitness-maintaining hits that could be uncovered by back-to-consensus using previously published near-comprehensive experimental fitness

landscapes for TEM-1[36] and a thermally stabilized LGK[18]. In both cases these experimental fitness landscapes were determined for enzymes stable in their genetic background. Thus, most neutral or beneficial mutations maintain similar catalytic efficiencies to wild type. I first classified the experimental fitness values into neutral (≥80% of wild-type), slightly deleterious (>50% and <80%) and deleterious (<50%) bins. It is important to note that the "neutral" bin also includes those mutations that improve fitness. The tolerance for a mutation at a given position as determined by evolutionary history was evaluated using a position-specific scoring matrix (PSSM). Whereas 32% of all TEM-1 mutations are neutral, 69% of mutations with a PSSM score of ≥ 3 were neutral (**Table B 3.7**). For LGK, 28% of all mutations were neutral but 57% of conserved (PSSM ≥ 3) mutations were neutral (**Table B 3.7**). Using a less restrictive cut-off (PSSM ≥ 0) does not appreciably change the findings. While these results suggest that including evolutionary history increases the probability of finding a mutation that improves stability and does not reduce the activity of the enzyme, I also note that the probability of an evolutionary sampled deleterious or moderately deleterious mutation is 31% (TEM-1) to 43% (LGK) (**Table B 3.7**). Thus, choosing mutations solely through evolutionary conservation is insufficient to engineer stable, active enzymes without a secondary activity screen.

Classification methods improve chances of finding soluble, active enzyme variants

Many solubility-enhancing mutations decrease enzyme specific activity. For example, it is well known that catalytically active residues are poorly optimized for solubility[5, 18, 26, 27]. Additionally, false positives like those observed in the TEM-1.1 datasets are often deleterious for fitness. Thus, additional metrics are needed to identify mutations that impart solubility and do not decrease activity.

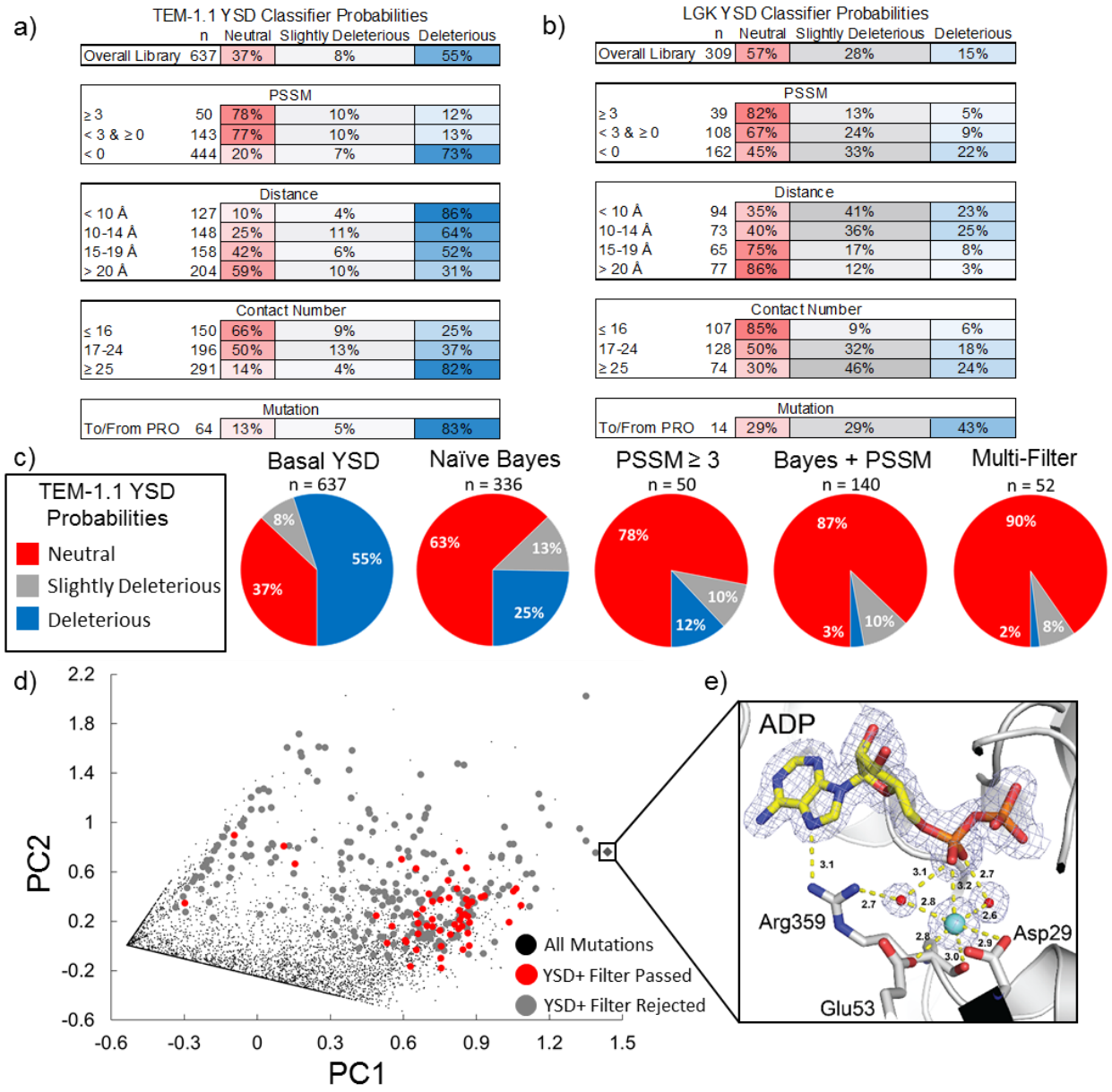


Figure 3.4: Classification methods improve probabilities of selecting mutations conferring solubility and activity but remove rare, globally optimal mutations. a,b) Classifier probabilities for YSD deep mutational scan for a) TEM-1.1 and b) LGK. n is the total number of mutations found in a given bin, and PSSM is the site-specific preferences found in the evolutionary history of the enzyme. c) Classification methods improve probabilities of selecting neutral mutations. d) Principal component analysis of the four experimental LGK deep mutational scanning datasets. PC1 correlates with enzyme fitness, while PC2 correlates with enzyme solubility. Beneficial mutations from YSD screen are shown as circles colored by whether they pass (Red) or fail (Gray) the multiple filter classification method. Pareto optimal mutation G359R (boxed) fails the filtering due to its close distance to the active site, low evolutionary conservation, and high contact number. e) Crystal structure of LGK G359R. G359R makes direct and water-mediated hydrogen bonds with ADP near the active site. A chloride ion

Figure 4 (con'd) also appears to be coordinated in this region, possibly also contributing to the stability of the enzyme. Carbon atoms are shown in gray or yellow for the protein and ligand carbon atoms, respectively. Nitrogen, oxygen and phosphorous atoms are shown in blue, red, and orange. Waters and the chloride are shown as red and cyan spheres, respectively. The $2mF_o$-DF_c electron density map is contoured to 1σ. Bond distances are in angstroms. For clarity, the magnesiums in the active site have been omitted from the figure.

For this analysis, I used previously published near-comprehensive experimental fitness landscapes for TEM-1[36] and a thermally stabilized LGK[18] using the same classification bins (neutral, slightly deleterious, deleterious) as above. For the datasets developed in this work, the probability of finding a deleterious mutation among the list of solubility-enriched variants ranges between 15% (LGK-YSD) and 55% (TEM-1.1-YSD) (**Figure 3.4a-b, Table B 3.8**). To improve my chances of finding solubility-enhancing mutations of neutral fitness, I assessed mutations according to size, chemical type, contact number of the original residue, distance to active site as determined by Cα distance to nearest active site ligand, and evolutionary conservation as quantified by PSSM.

I first addressed whether PSSM improves the identification of active mutations. Solubility-enhancing mutations with a PSSM ≥ 3 are more likely to maintain fitness and are less likely to be deleterious (**Figure 3.4a-b**). For example, only 12% of solubility-enhancing mutations with a PSSM ≥ 3 observed from the TEM-1.1 YSD datasets are deleterious with regards to fitness. In contrast, non-conserved solubility-enhancing mutations are likely to be deleterious for fitness, with probabilities ranging from 22% for LGK-YSD to 73% for TEM-1.1 YSD.

Other classifiers yielded similar results. For example, distance to active site was correlated with increased probability of finding a neutral mutation, whereas contact number was anti-correlated (**Figure 3.4a-b**). However, mutations sorted by size and chemical type did not

show improvements in classifications across all selections (**Table B 3.9**), with the exception of mutations to or from proline, which are generally disfavored.

Next, I tested different classification methods to improve the chances of finding solubility-enhancing mutations conferring neutral fitness. I looked at filtering, naïve Bayes classification, and a hybrid method combining filtering on PSSM (≥ 3) with Bayes analysis on the remaining classifiers (**Figure 3.4c, Table B 3.10**). One filter was strictly on PSSM, whereas multiple filtering included PSSM (≥ 0), distance to active site (≥ 15Å), contact number (≤ 16), and no mutations involving a proline. The multiple filter classification performed best in all three datasets: for the YSD datasets the probability of finding a neutral mutation is 90% with only a 2% (TEM-1.1) or 3% (LGK) chance of uncovering a deleterious mutation. The hybrid Bayes method was next, with a 77-87% chance of finding a neutral mutation and a 3-4% chance of a deleterious mutation for the YSD datasets. However, increased accuracy using multiple filtering is at the expense of number of mutations identified, as the hybrid Bayes method identified approximately 3-fold more mutations than the multiple filtering method (mean 130 for hybrid Bayes versus 43 for filtering).

The trade-off with strict filtering is the removal of rare, globally beneficial mutations. This balance can best be visualized by taking a principal component analysis of the four experimentally determined LGK fitness datasets. PC1 maps mostly to enzyme fitness as determined by growth competition, with PC2 correlating more strongly with enzyme solubility (**Figure 3.4d**). In this case, the multiple filtering method removes the Pareto optimal mutation[48] G359R that is enriched in all four datasets (**Figure 3.4d**). *In vitro*, LGK G359R shows an increased ΔT_m of 1.1ºC and improves the k_{cat} over wild-type by approximately 60%[18]. G359R was removed from consideration because it is not evolutionarily conserved, it is close to the ATP

binding cleft, and Gly359 is in a relatively packed portion of the protein. To evaluate the structural basis for this globally beneficial mutation, the structure of this mutant in the presence of ADP and magnesium (**Figure 3.4e**) was solved. The additional hydrogen bonding interactions from the Arg359 side-chain to the nucleotide in the active site may lead to stronger binding of ATP during the catalytic reaction and possibly have polarizing effects that enhance phosphate transfer. From this regard it is also interesting to note that a strong electron density peak near Arg359 and ADP, modeled as a chloride ion, may also affect electrostatic interactions of the required reactants and promote catalysis.

DISCUSSION

Deep mutational scanning has previously been used to evaluate enzyme function on a massive scale[18, 36, 37, 49]. In the present work deep mutational scanning was combined with existing solubility screens for two different full-length enzymes. The addition of mutational solubility data produces a more complete picture of the fitness landscape for an enzyme. For example, the LGK datasets show that 4-5% of mutations are beneficial for solubility, which is consistent with theoretical predictions (**Figure 3.3c**). However, for the TEM-1.1 YSD dataset the fraction of solubility-enhancing mutations was 10.3% after removing known false positives, still higher than theoretical predictions but closer to the LGK dataset. Further experiments on different enzymes are needed to evaluate the level of concordance with current theories. In the present work I find at least 40% of given solubility-enhancing mutation result in enzymes with impaired fitness. While this trade-off has generally been modeled as solubility-reducing residues in the active site of the enzyme[26], a significant number of mutations occur at distances up to 15Å away from any active site residues for all datasets.

There are limitations to consider when applying each screen/selection to a new enzyme. Deep mutational scanning using the GFP fusion in the bacterial cytoplasm is much faster than YSD, but cytoplasmic expression will not work for proteins requiring disulfide bonds (like TEM-1) or additional posttranslational modifications to fold. In my hands the TAT export selection was not able to identify known thermally stabilizing mutations compared with the other screens. These results could be explained by the fact that export through the TAT pathway does not select for thermally stabilizing mutations, as previously shown (25). On the other hand, YSD screening can be used for most homo-oligomeric enzymes, including ones that contain disulfides or glycosyl groups.

Applying simple classification methods to the deep mutational scanning screens increases the probability of selecting soluble, active mutants. Evolutionary conservation by itself is a strong predictor of solubility-enhancing mutations. However, other classification methods are still needed, as the probability of finding a stable, active mutation solely by PSSM alone is 52-66% using the back-to-consensus approach commonly used in protein engineering (**Table B 3.7)**. Filtering on multiple classifiers gives a probability of selecting a deleterious mutation of 2-3% (**Figure 3.4c, Table B 3.11**); this is a low enough error rate for 10-20 mutations to be combined additively and still maintain activity. However, strict filters remove rare, Pareto optimal mutations like G359R that improve both stability and catalytic efficiency.

From a protein engineering and design perspective, the next step is to test the generality of the approach to find active solubility-enhancing mutations. There are a number of considerations that make this deep mutational scanning approach attractive. First, there exist multiple screens shown to identify gain of stability mutations. Second, even already stable enzymes can be scanned using the recursive destabilizing approach developed by Bradbury and

colleagues[41]. In fact, I used this approach in the present work by destabilizing the robust TEM-1 using the known D197G mutation. Finally, the classifiers used were designed so that high-resolution structures are not necessary. PSSM and type of mutation do not require structural information, while contact number and distance to active site can be approximated even with crude homology models. The remaining step is to automate a design process to combine many solubility-enhancing mutations simultaneously. To that end, excellent results have already been demonstrated on the LGK system[18] and in the presence of a solved structure using the PROSS method[10]. I anticipate, since the mutations are already known, that all-atom design using Rosetta or similar software packages will prove successful even using crude homology models. I anticipate the use of this suggested approach to improve the solubility of difficult proteins involved in biomanufacturing and metabolic engineering.

MATERIALS AND METHODS

Reagents

All DNA primers were ordered from IDT and genetic constructs were sequence verified by Genewiz. All chemicals and plates were purchased from Sigma-Aldrich.

Plasmid construction

The pSALECT and pETConNK plasmid backbones were used as previously described[42]. In short, a ΔS4-A25 truncation using the Ambler consensus numbering system[50] of TEM-1.1 BLA (**Notes B 3.1 and 3.2**) was cloned in-between the *NdeI* and *XhoI* sites of the two backbones to create the pSALECT-TEM1.1/csTEM1 and pETConNK- TEM1.1 plasmids. The codon

optimized DNA sequence for LGK[18] was cloned in-between the *NdeI* and *XhoI* sites of the two backbones to create the pSALECT-LGK/csTEM1 and pETConNK-LGK plasmids.

The plasmids pET29b-TEM1.1/mGFPmut3 and pET29b-LGK/mGFPmut3 were created by cloning genes between *NdeI* and *XhoI* sites. Overhang PCR was used to add a 5' XhoI site and a 3' His_{6x}, 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 already present in the plasmid. The mGFPmut3 construct was cloned between the *XhoI* and *BbvCI* sites using standard techniques to make the fusion construct LGK-Leu-Glu-mGFPmut3-His_{6x}. Plasmids and full maps are freely available on AddGene (www.addgene.org).

Library construction

Comprehensive single-site mutagenesis was performed on the gene encoding sequences for LGK and TEM-1 within the three plasmid backbones using Nicking Mutagenesis[42]. Mutagenic primers encoding degenerate bases (NNN) were used for residues G8 to T435 for LGK and H26 to W290 for TEM-1.1. Plasmids were transformed into *E. coli* XL1-Blue and plasmids were extracted using a Qiagen miniprep kit the following day.

For generation of YSD libraries, chemically competent EBY100 yeast was transformed with 5 µg of pETConNK based library plasmid DNA and grown in 50 mL SDCAAps (Synthetic complete media supplemented with amino acids, 2% (w/v) dextrose, and 10,000 u/mL penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA))[44] for 24 hours at 30ºC. The cells were passaged into fresh 50 mL SDCAAps media and grown for another 24 hours. Yeast were stored in yeast storage buffer (20 mM HEPES 150 mM NaCl pH 7.5, 20% (w/v) glycerol) [51] at -80ºC in 1 mL aliquots at an OD_{600}=1.0 (1 yeast OD_{600} = $2x10^7$ cells/mL).

10 ng of pSALECT based library plasmid DNA was transformed into electrocompetent *E. coli* MC4100 (Coli Genetic Stock Center, New Haven, CT) and grown on a Nunc Bioassay Plate (245 mm X 245 mm X 25 mm) at 30ºC overnight. Transformation controls were performed to limit double plasmid transformation [43]. Cells were scraped and used to inoculate a 100 mL LB culture with 34 μg/mL chloramphenicol at an initial OD_{600} of 0.05 and grown at 30ºC and 250 rpm. When the cultures reached mid-log (OD_{600} 0.40 to 0.60) DMSO was added at a final concentration of 7% (v/v), and 1 mL aliquots were flash frozen in liquid nitrogen. The same approach was taken for pET29b based GFP fusion DNA libraries, except that 50 ng of pET29b library plasmid DNA was transformed into electrocompetent *E. coli* BL21*(DE3).

Screening procedures

Yeast display library cell stocks were thawed at room temperature and were used to inoculate a 1 mL SDCAAps at 30ºC at an initial OD_{600} of 1.0 then grown for 6 to 8 hours. These cells were used to start a 1.1 mL culture in SGCAAps media (Synthetic complete media supplemented with amino acids, 2% (w/v) galactose, and 10,000 u/mL penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA)) at an initial OD_{600} of 1.0 and grown at 30ºC for 18 hours. The next day cells were spun down at top speed for 30 sec and the media pipette removed. Cold PBSF (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, and 2 mM KH_2PO_4 with 1 g/L BSA) was added to the pellets to an OD_{600} of 2. The cells were washed with chilled PBSF, suspended to a concentration of $2x10^6$ cells per mL PBSF, and labeled with 1 μL of anti-c-myc-FITC (Miltenyi Biotec, San Diego, CA) per $2x10^5$ cells. Cell sorting and collection was done on a BD Influx cell sorter (Becton Dickinson, Franklin Lakes, NJ). FSC/SSC (gate 1), FSC/FITC (gate 2), and FITC (gate 3) gates were set. Two populations were collected: an unselected population with gate 1,

and a top 5% display population containing all three gates. Complete sorting statistics can be found in **Table B 3.1**. Following sorting the cells were grown in 10 mL of SDCAAps at 30ºC for 24 hours and were stored at -80ºC in yeast storage buffer at a concentration of $4x10^7$ cells per mL. DNA was extracted from the yeast and prepared for sequencing using previously published protocols[43].

GFP fusion library cell stocks were thawed on ice for 45 minutes prior to washing with fresh sterile TB media. For each library, 3 mL cells were inoculated in Hungate tubes at an initial OD_{600} of 0.003. The cultures were grown at 37ºC and 250 rpm aerobically until an OD_{600} of 0.05-0.06, at which time IPTG was added to a final concentration of 250 µM. The cells were collected when the cultures reached an OD_{600} between 0.28-0.30. Each culture was pelleted at 10,000xg and washed with cold sterile PBS twice. Cell sorting and collection were done on a BD Influx with two sorting gates (FSC/SSC (gate 1), and FL-1 with a 530/40 filter [488 nm] (gate 2)). Two populations were collected: a reference population with gate 1, and the top 5% of cells by fluorescent intensity set by gate 2. Complete sorting statistics can be found in **Table B 3.1**. The collected cells were added to 10 mL of fresh TB media with kanamycin and grown aerobically at 25ºC until mid-exponential phase. The cultures were then washed with PBS, pelleted at top speed, and the DNA was extracted using a Qiagen miniprep kit.

TAT export library cell stocks were thawed on ice for 45 minutes prior to washing with fresh LB media. The washed cells were used to start a 5 mL culture containing LB with 34 µg/mL chloramphenicol inoculated at an initial OD_{600} of 0.05. The cells were grown aerobically at 30ºC and 250 rpm until a culture OD_{600} of 0.8. The unselected library was prepared by pelleting 1 mL culture at 17,000xg for 2 min and storing the pellet at -20ºC. Libraries were plated on 100 (TEM-1.1) or 200 µg/mL (LGK) carbenicillin plates.

The LGK libraries were plated at 0.1 OD_{600}/mL on two 100 mm diameter Petri plates, while the TEM-1 libraries were plated at 3.2 OD_{600}/mL on Nunc Bioassay Plates (245 mm X 245 mm X 25 mm). The number of cells plated was sufficient to support a 200-fold coverage of the theoretical DNA library in viable cells. Plates were cultured at 30ºC in a humidified incubator for 12 hours. The following day the plates were scraped with 1x PBS, pelleted, and a Qiagen miniprep kit was used to extract DNA from saved cell pellets.

Deep sequencing and data analysis

Libraries were prepared for deep sequencing using a previously developed two step PCR method[43] with PCR primers listed in **Table B 3.11.** The pooled library was extracted and cleaned with a Qiagen gel cleanup kit. Deep sequencing was performed using an Illumina MiSeq in 300 bp paired-end mode. Sequencing data were processed using Enrich[52] to quantify the amount and enrichment of each mutation. Deep sequencing statistics are listed in **Tables B 3.2-3.4**. For the yeast display and GFP fusion experiments the solubility score for a variant i (ζ_i) is defined as:

$$\zeta_i = log_2(\frac{F_i}{F_{wt}}) \quad (1)$$

Where F_i is the mean fluorescence of variant i, and F_{wt} is the mean fluorescence of the wild type sequence. This solubility score is calculated using experimental observables in the deep sequencing experimental pipeline according to:

$$\zeta_i = \log_2(e)\sqrt{2}\,\sigma' \left(erf^{-1}\left(1-\phi 2^{(\varepsilon_{wt}+1)}\right) - erf^{-1}\left(1-\phi 2^{(\varepsilon_i+1)}\right)\right) \quad (2)$$

where ε_i is the enrichment ratio of the variant, ε_{wt} is the enrichment ratio of the starting sequence, σ is the standard deviation of the population, and ϕ is the percent of cell collected of the gating population[43]. For the TAT experiments the enrichments of variant i was normalized to the starting sequence by the equation:

$$\zeta_i = \varepsilon_i - \varepsilon_{wt} \quad (3)$$

where ε_i is the enrichment ratio of the variant and ε_{wt} is the enrichment ratio of the starting sequence. Python scripts to calculate solubility scores are publically available at Github [user: JKlesmith] (www.github.com). Processed deep sequencing datasets are deposited at figshare (www.figshare.com).

PSSM Analysis

The PSSM analysis was performed closely following Goldenzweig et al.[10]. In short, a blastp search[53] of the nonredundant database for LGK and TEM-1 was performed with an e-value cut-off of 10^{-4} and filtered to the top 20,000 results. Synthetic or engineered constructs were excluded from the hits. Hits were also excluded if they covered less than 85% of the query sequence or if their sequence identity was less than 34% for LGK or 40% for TEM-1. Cd-hit[54] was used at 98% clustering threshold and default parameters. MUSCLE[55] was then used to produce a multiple sequence alignment of the top 700 clusters. DSSP[56] was then used to identify residues that are a part of loops and a part of secondary structure elements. Insert sequences in loop regions were removed such that the alignment has no gaps in the wild-type sequence. An alignment of sequences without any frameshifts was then independently extracted from each structured and non-structured region. PSI-BLAST[57] was then used on each region with the wild-type sequence as the query sequence.

LGK G359R expression and purification

Recombinant *E. coli* BL21(DE3) GOLD cells harboring plasmid pET29b_LGK-G359R were grown to an OD_{600} of ~ 0.5 at 37 °C, with shaking, in 500-ml volumes of LB medium

supplemented with 35 μg/ml kanamycin. Expression of LGK was induced with 1 mM IPTG for 3 h at 30 °C, with shaking. Cells were pelleted by centrifugation and stored at -80 °C. Pellets were thawed in 20 ml of ice-cold lysis buffer (0.5 M NaCl, 20 mM Tris-HCl pH 7.5, 0.1 mM PMSF, 2 mM imidazole) and lysed using a sonicator (Ameco). The lysate was clarified by centrifugation and mixed with 2 ml of TALON metal affinity resin (Clontech) with gentle shaking for 30 min. at room temperature. The TALON beads were centrifuged and re-suspended in binding buffer (500 mM NaCl, 20 mM Tris pH 7.5, 0.5 mM TCEP) before being poured into a 20 ml gravity column. The column was washed with 20 ml of binding buffer supplemented with 5 mM imidazole (Sigma), followed by 20 ml of binding buffer supplemented with 10 mM imidazole. The LGK protein was eluted from the column with 10 ml of binding buffer supplemented with 250 mM imidazole. The protein was further purified by gel filtration (HiPrep 26/60 Sephacryl S-200 HR) in 20 mM Tris pH 7.5, 50 mM NaCl, 0.5 mM TCEP prior to concentration using an Amicon Ultra-15 concentrator with a 10,000 Da cut-off (Millipore). Chromatographic steps were performed using an AKTA FPLC (GE Healthcare).

LGK crystallization, data collection and structure determination

LGK crystals were grown at room temperature using the hanging drop vapor-diffusion method by mixing equal volumes of reservoir buffer (22% polyethylene glycol (PEG) 3350, 0.2 M KSO_4, 100 mM Tris pH 6.8) and LGK (23 mg/ml) in crystallization buffer (50 mM NaCl, 2 mM ADP, 4 mM $MgCl_2$, 0.5 mM TCEP, 20 mM Tris pH 7.5). Crystals were cryoprotected by dragging them through a drop containing cryoprotectant solution, reservoir buffer supplemented with 9% sucrose (w/v), 2% glucose (w/v), 8% glycerol (v/v), 8% ethylene glycol (v/v)), prior to being flash-cooled in liquid nitrogen. Data were collected at the Stanford Synchrotron Radiation

Lightsource beamline BL7-1, integrated using MOSLFM[58] and scaled and merged using SCALA.

Structure was determined using rigid body refinement using (PDB identifier: 5BSB) as the starting model followed by iterative model building and refinement performed using Coot and PHENIX[59]. The stereochemical quality of the final models was assessed using MolProbity[60]. Refinement statistics are presented in **Table B 3.12**. All structural figures were prepared using PyMOL[61].

APPENDIX

APPENDIX

Note B 3.1: The amino acid sequence for TEM-1.1. Mutations S70A and D179G are underlined in red highlight.

HPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMATFKVLLCGAVLSRV
DAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIG
GPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERGTTMPAAMATTLRKLLTGELLTL
ASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVI
YTTGSQATMDERNRQIAEIGASLIKHW

Note B 3.2: The DNA sequence for TEM-1.1.

CACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGT
GGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGA
AGAACGTTTTCCAATGATGGCCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCC
CGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGA
CTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAA
GAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTC
TGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGAT
CATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGA
CGAGCGTGGCACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAA
CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGG
ATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTG
ATAAATCTGGAGCAGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCA
GATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATG
GATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGG

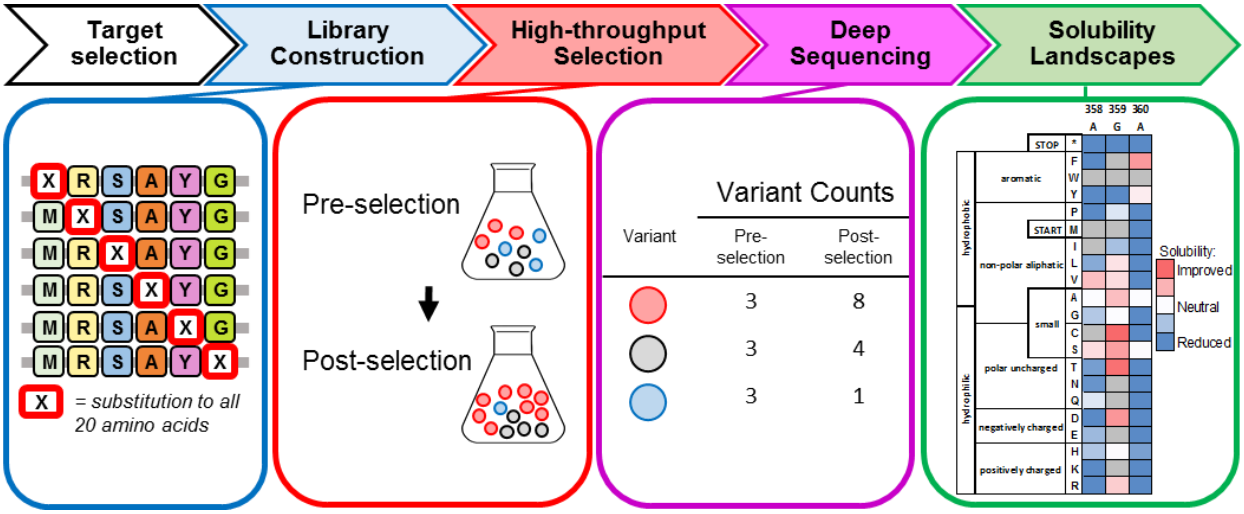


Figure B 3.1: Deep sequencing pipeline. A target protein is first mutagenized such that a DNA library encodes all possible amino acids. Next, a high-throughput selection or screen is performed to enrich beneficial mutants and deplete deleterious mutations. Deep sequencing is used to count each mutation to allow the frequency of that mutation in the population to be calculated. Finally, the frequencies are normalized to a solubility score for each mutation.

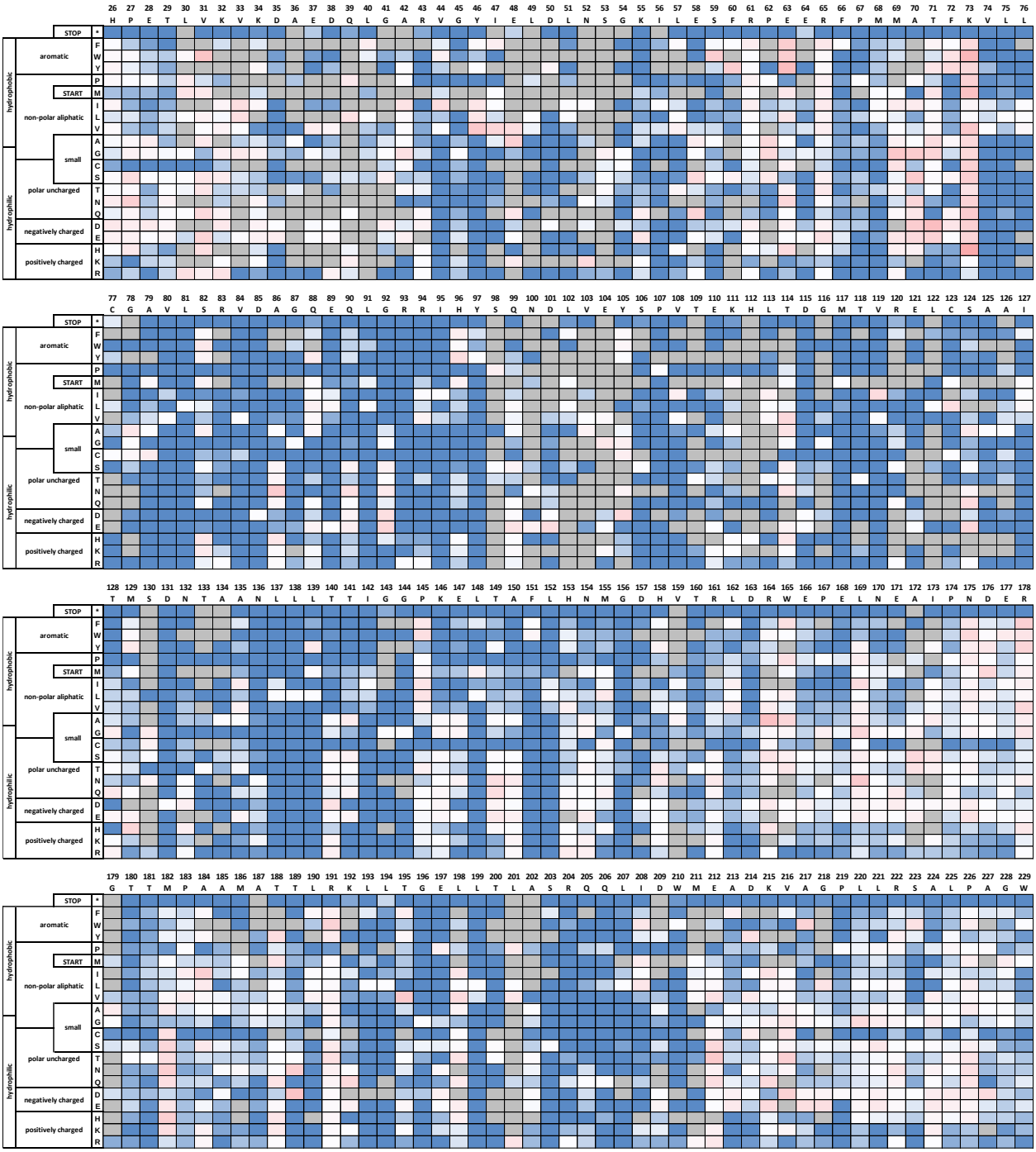
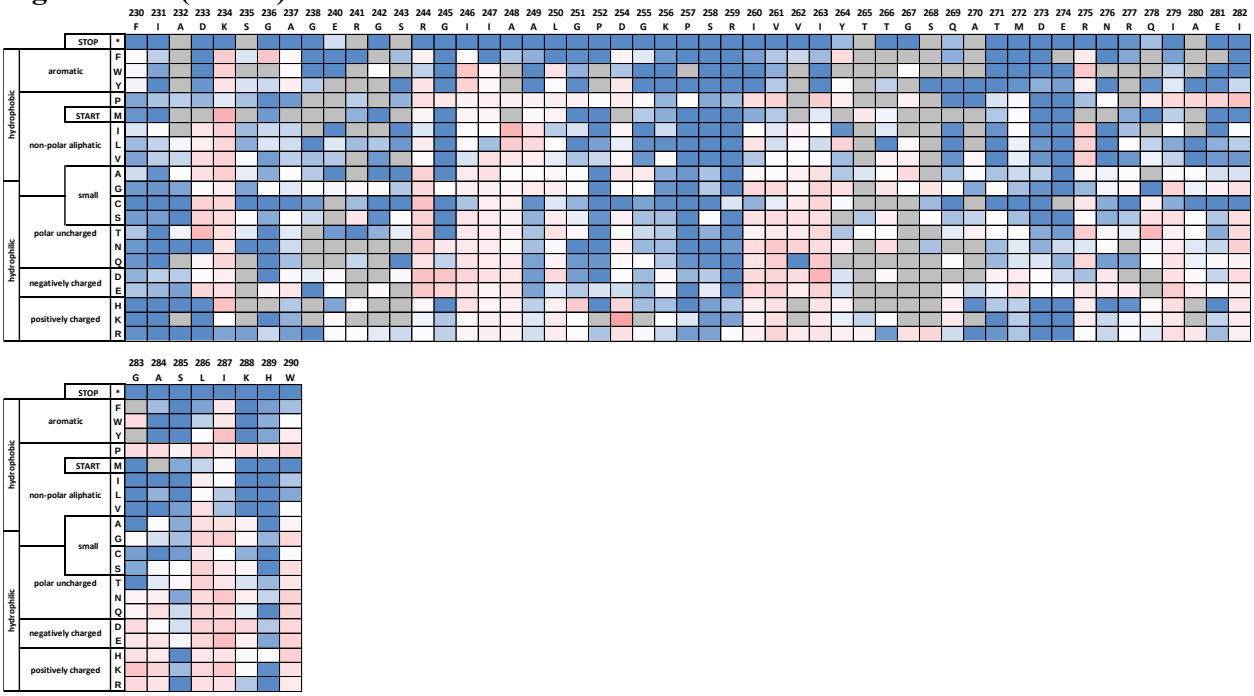


Figure B 3.2: Heatmap of solubility score of TEM-1.1 variants screened by yeast display.

Figure B 3.2 (cont'd)



Solubility Score Key

YSD	
≥2.00	Improved
1.00	
0.00	Neutral
-0.25	
≤-0.50	Reduced

<12 reference counts:

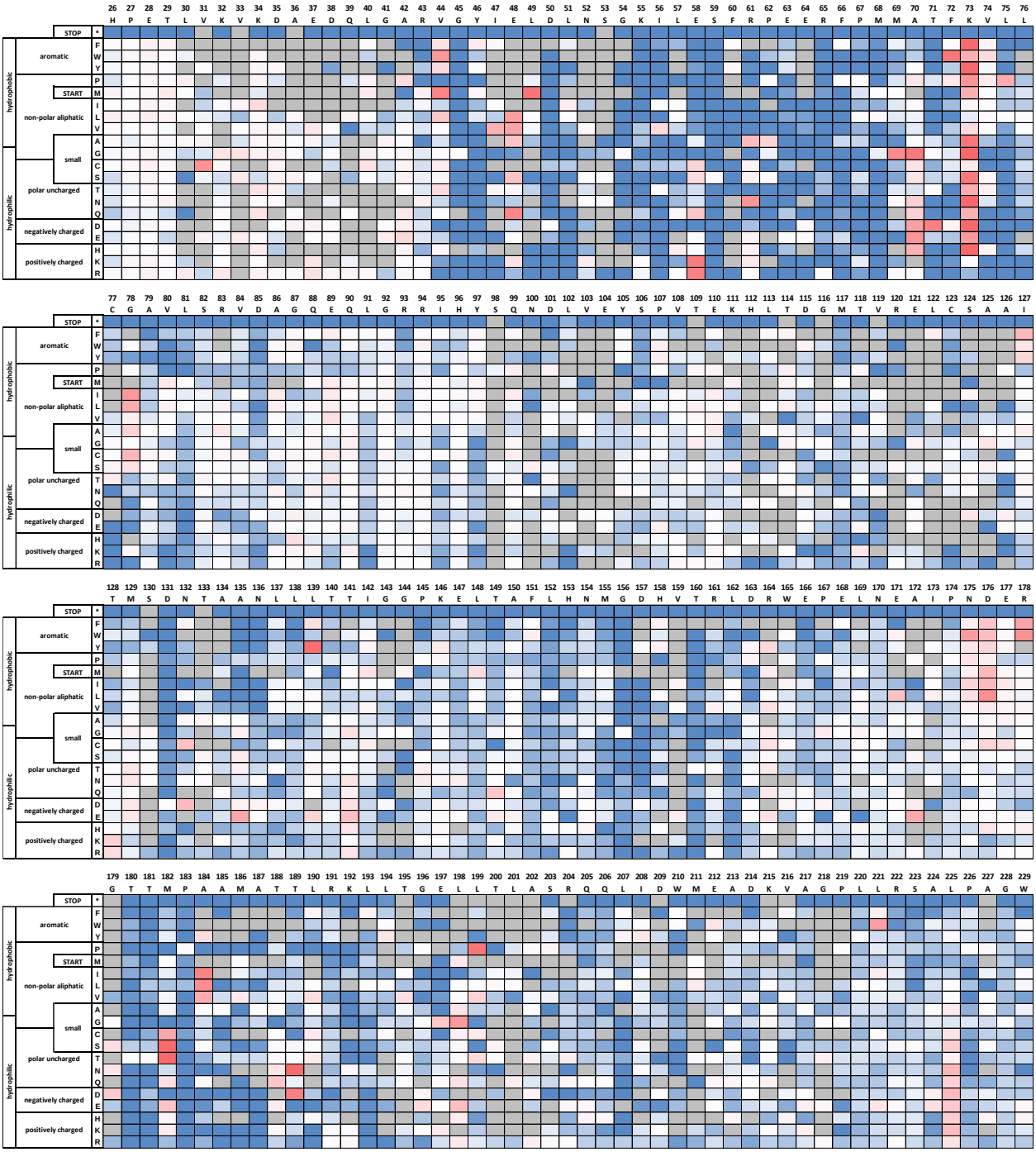
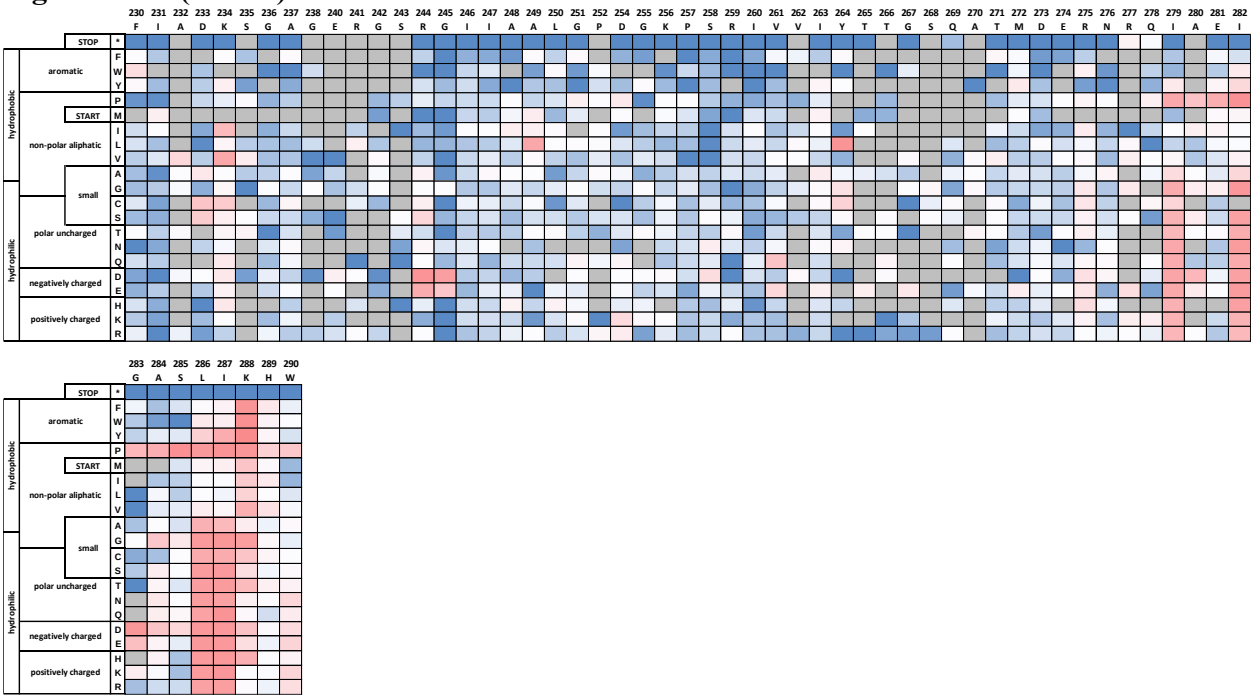


Figure B 3.3: Heatmap of solubility score of TEM-1.1 variants screened by TAT export.

Figure B 3.3 (cont'd)



Solubility Score Key

TAT	
≥4.00	Improved
2.00	
0.00	Neutral
-0.50	
≤-1.00	Reduced

<12 reference counts:

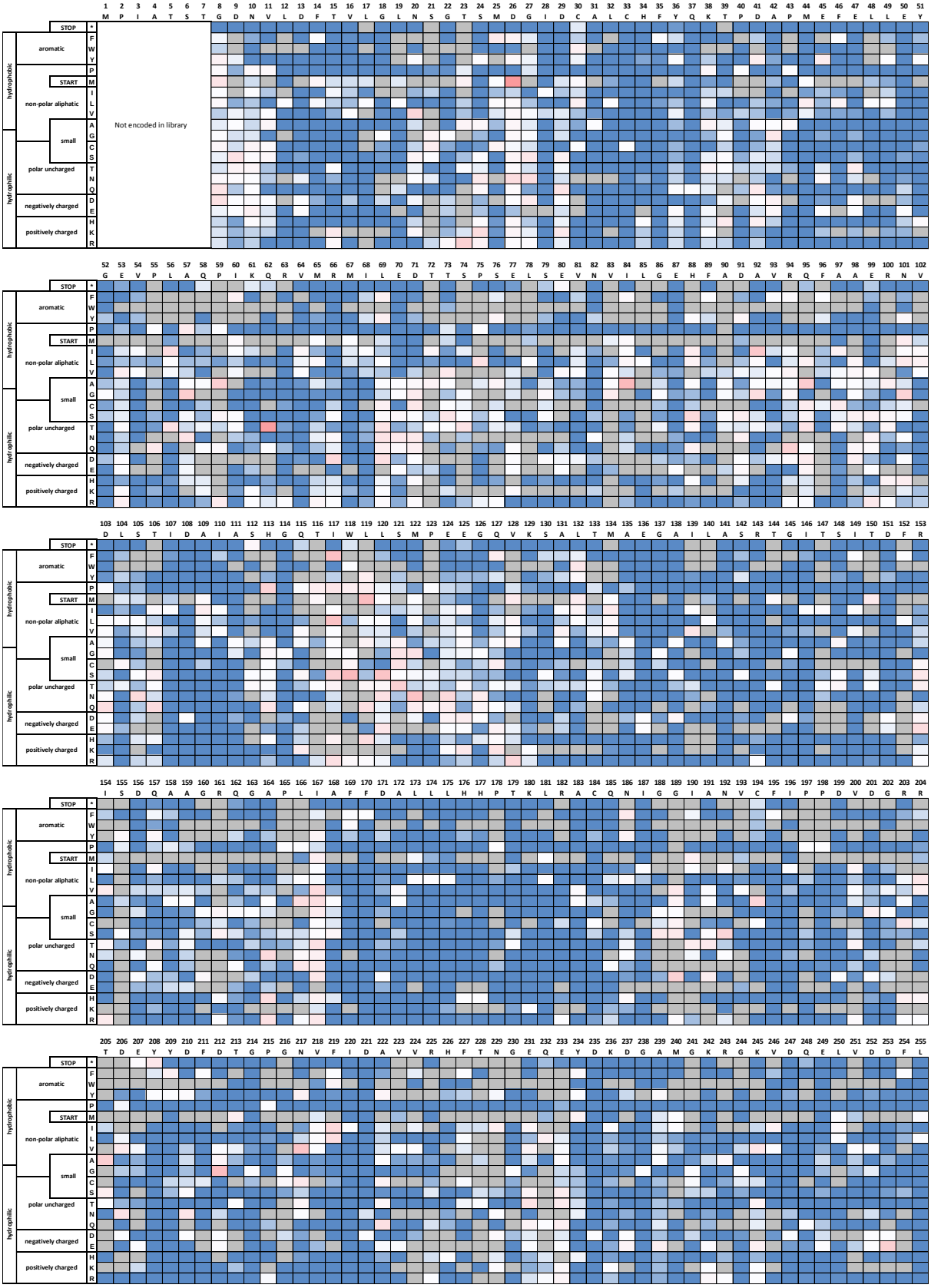
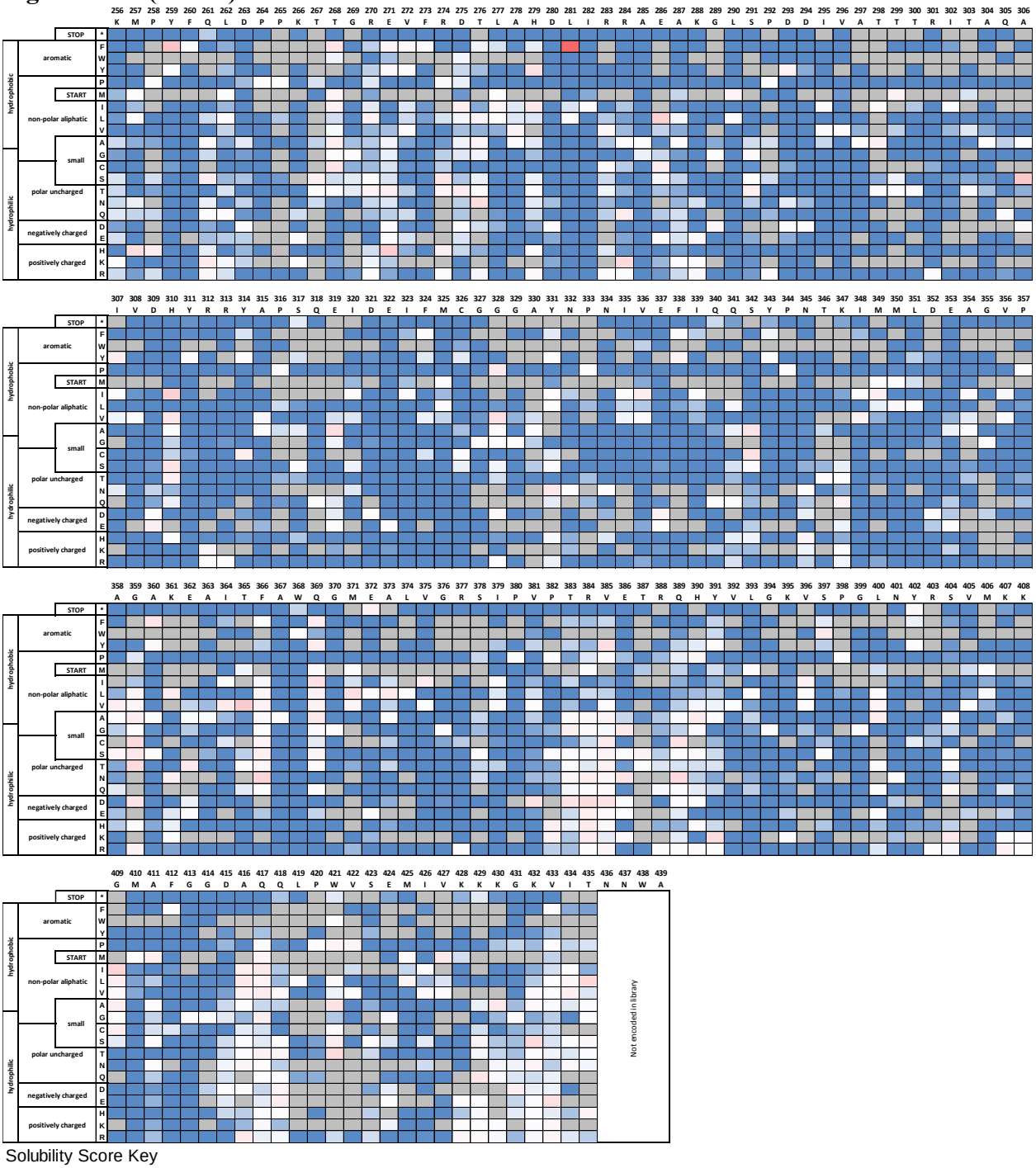


Figure B 3.4: Heatmap of solubility score of LGK variants screened by yeast display.

Figure B 3.4 (cont'd)



Solubility Score Key

YSD	
≥2.00	Improved
1.00	
0.00	Neutral
-0.25	
≤-0.50	Reduced

<12 reference counts:

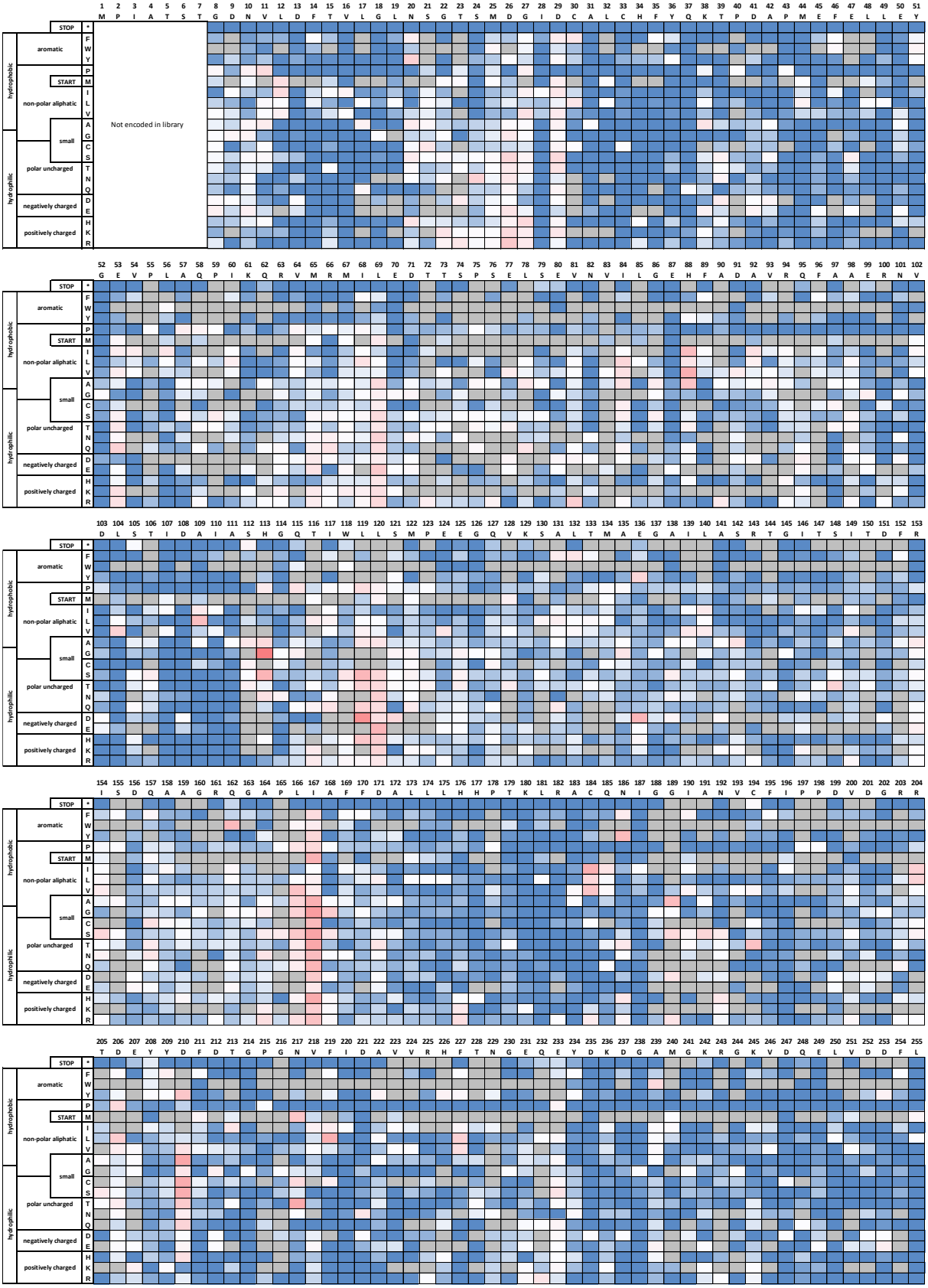
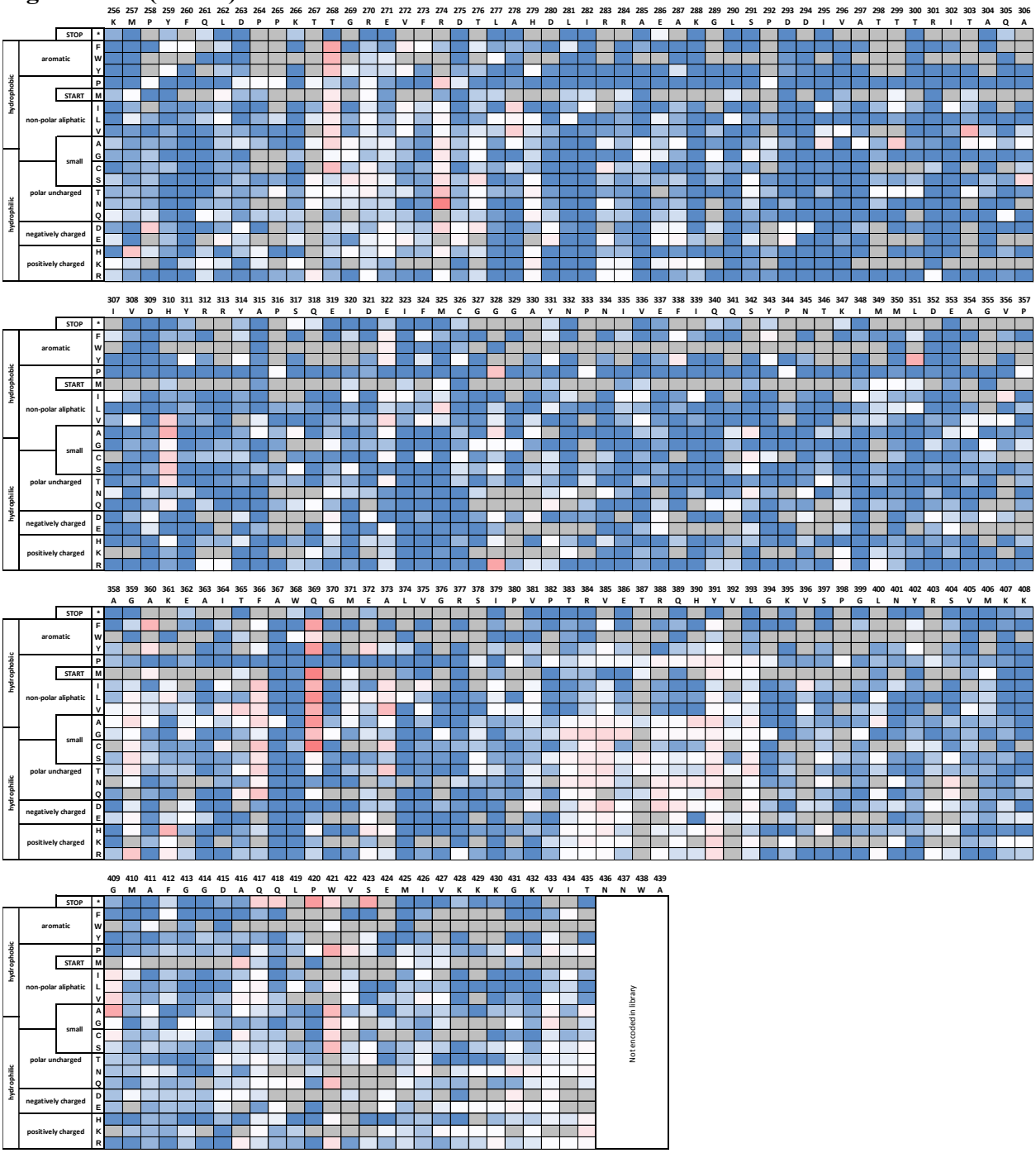


Figure B 3.5: Solubility score heatmap of LGK variants screened by GFP fusion.

Figure B 3.5 (cont'd)



Solubility Score Key

GFP	
≥2.00	Improved
1.00	
0.00	Neutral
-0.25	
≤-0.50	Reduced

<12 reference counts:

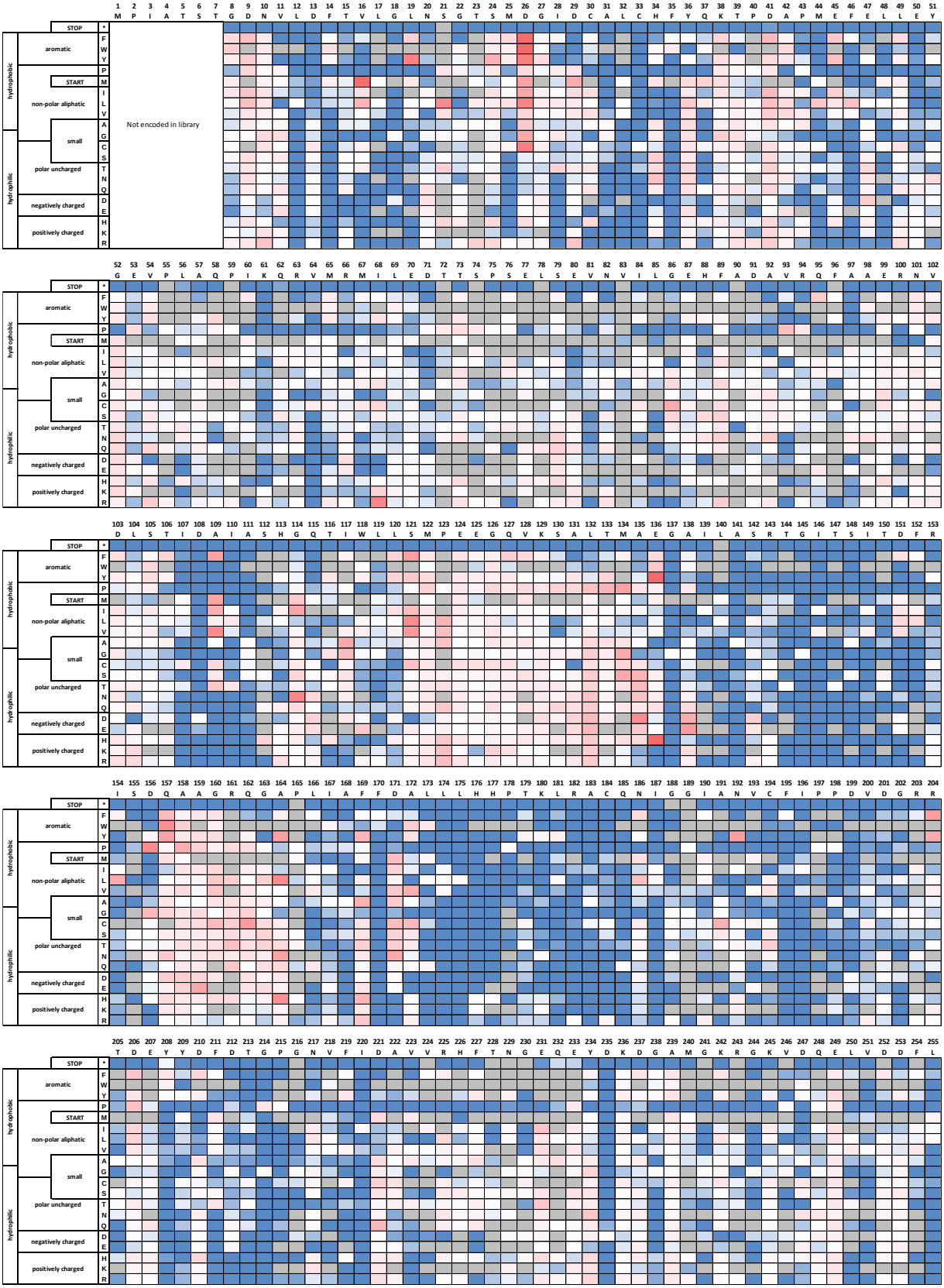
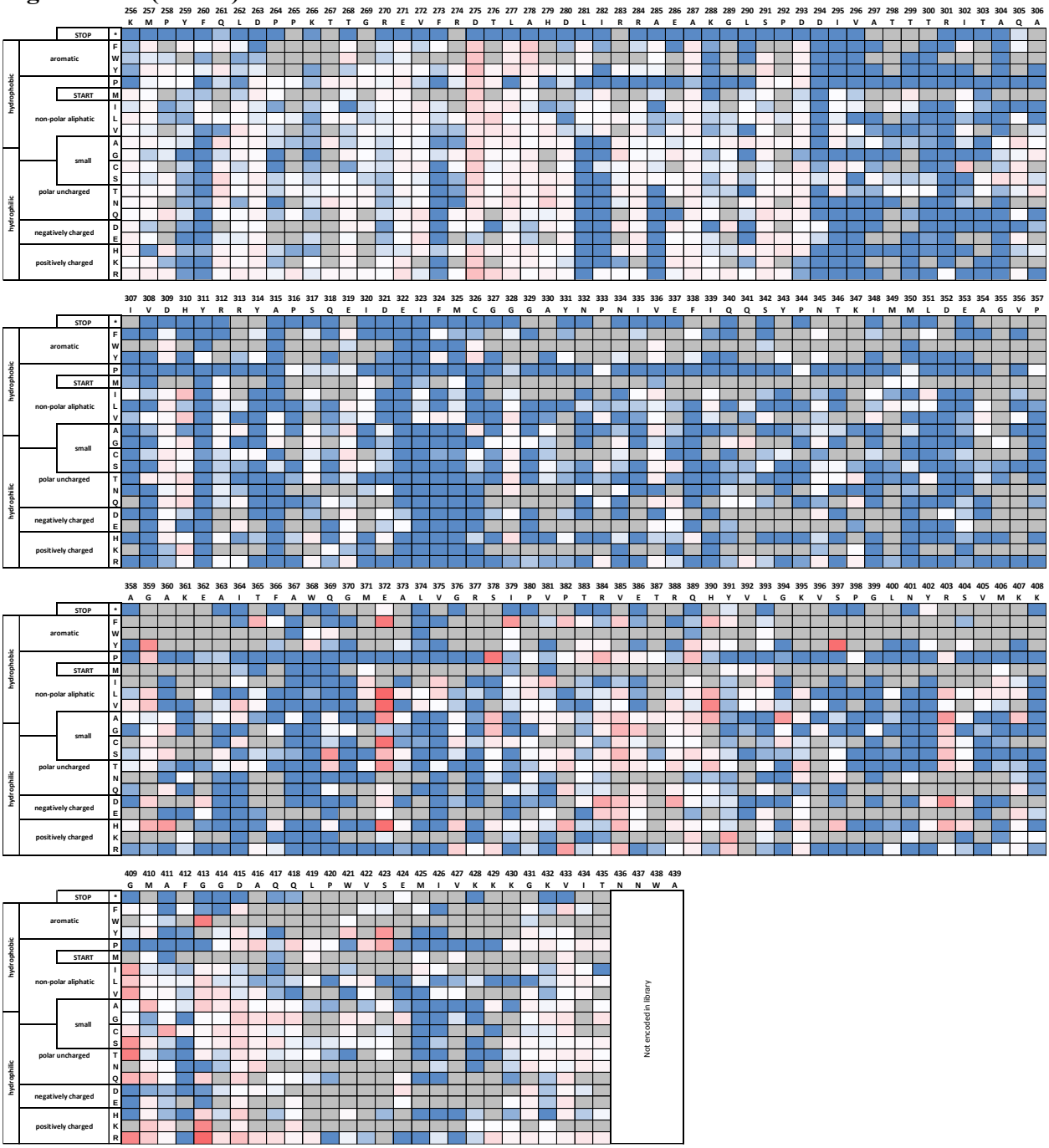


Figure B 3.6: Heatmap of solubility score of LGK variants selected by TAT export.

Figure B 3.6 (cont'd)



Solubility Score Key

TAT	
≥4.00	Improved
2.00	
0.00	Neutral
-0.50	
≤-1.00	Reduced

<12 reference counts:

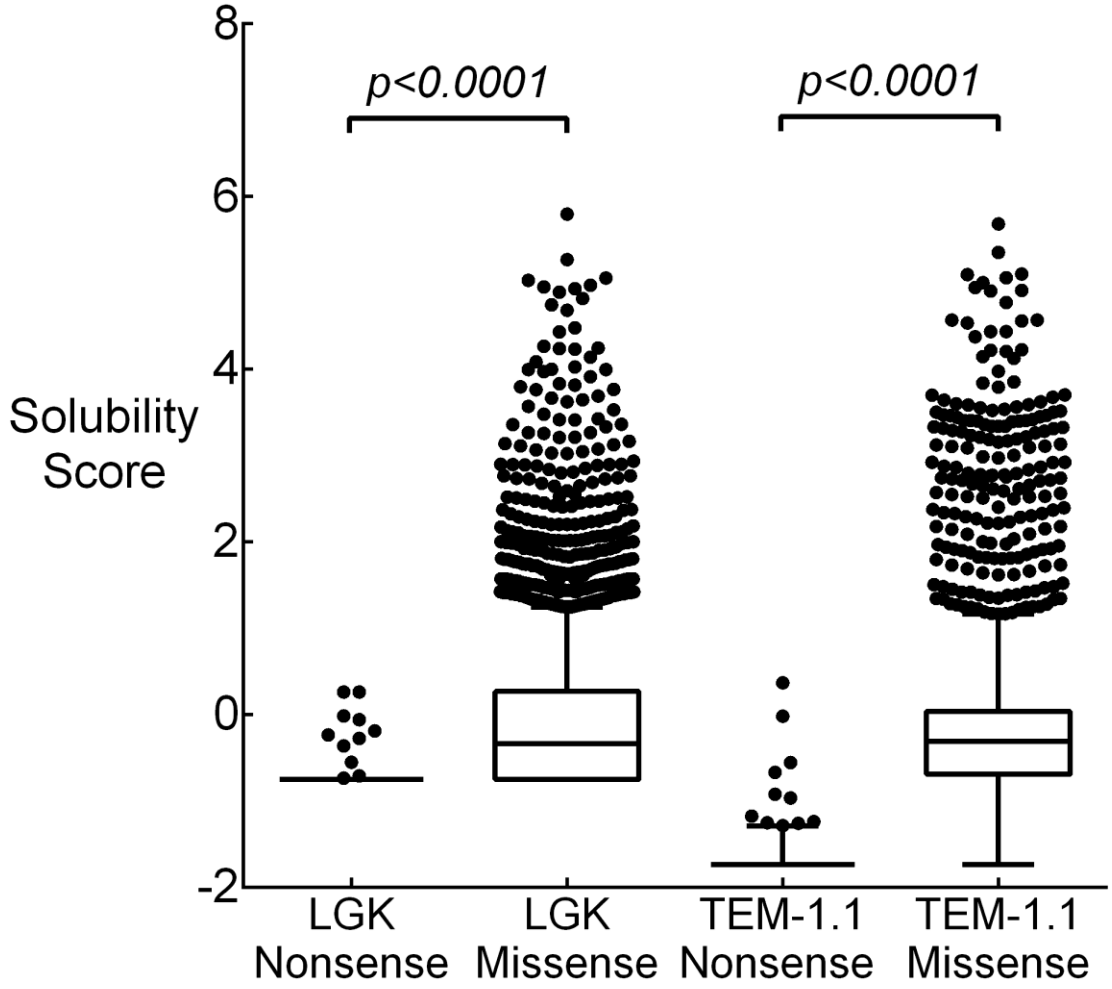


Figure B 3.7: Distribution of nonsense versus missense distributions for the TAT selection. An unpaired t-test with Welch's correction was performed between the fitness metrics for nonsense and missense mutations of each enzyme (n=331 and 6386 for nonsense and missense mutations in LGK respectively, n=227 and 3976 for nonsense and missense mutations in TEM-1.1 respectively).

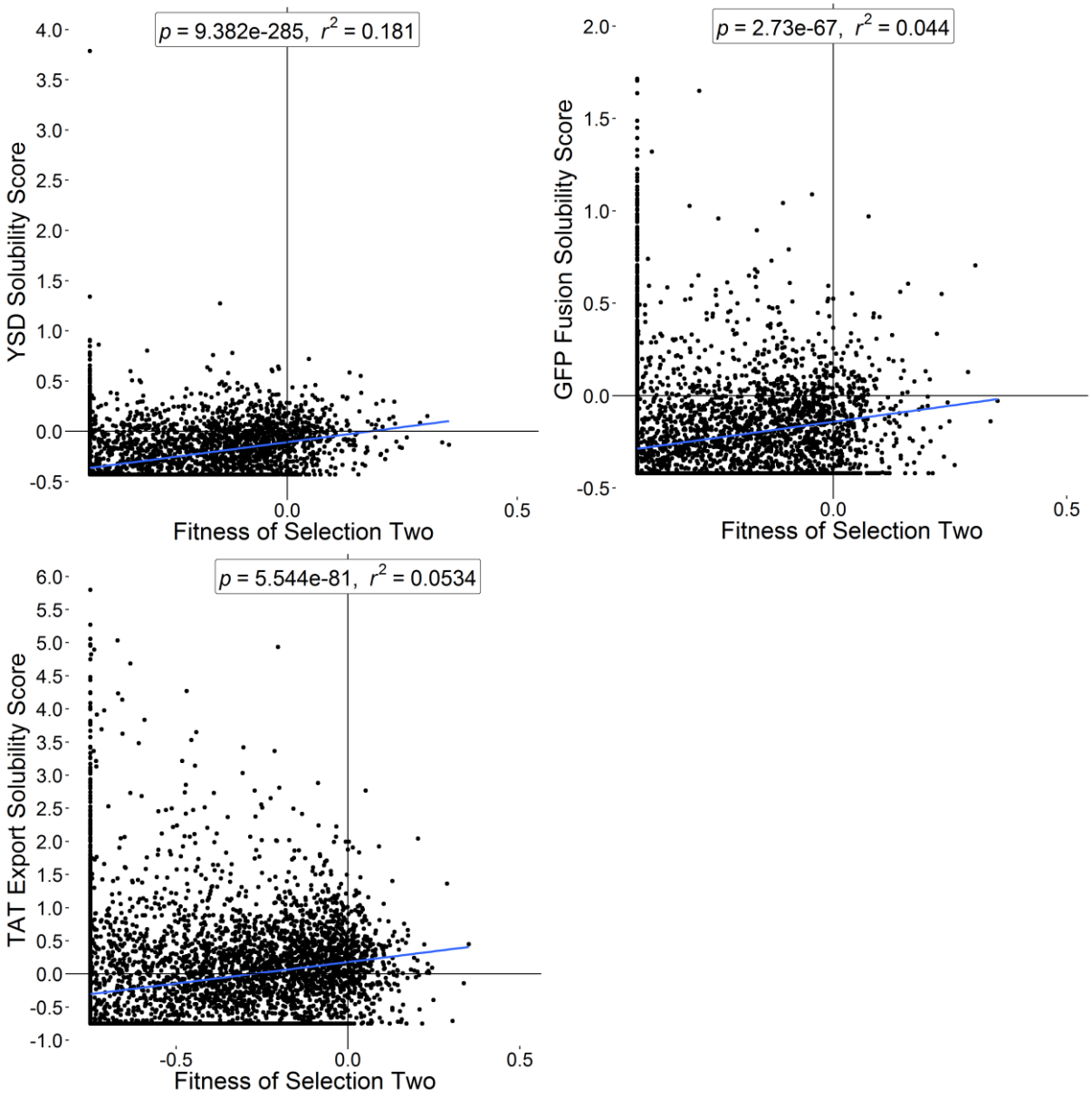


Figure B 3.8: Linear regressions of solubility versus functional datasets for LGK. The second selection using LGK.1 as the starting construct from Klesmith *et. al.* [18] was used as the functional dataset comparison for the solubility screens (denoted as "Selection Two" on the X-axis).

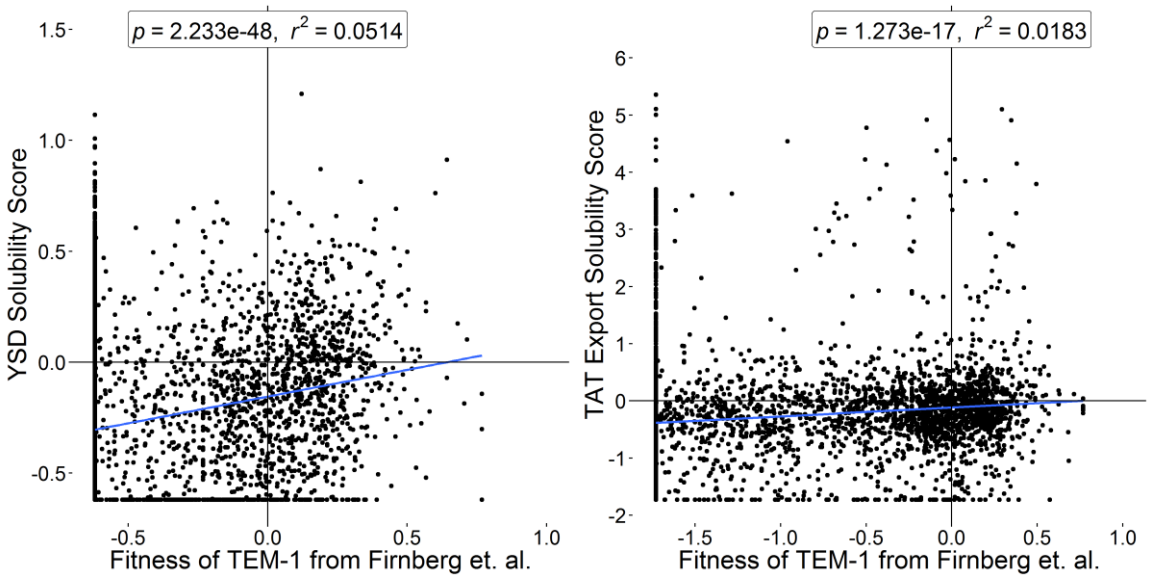


Figure B 3.9: Linear regressions of solubility versus functional datasets for TEM-1.1. Log2 transformed fitness values derived from Firnberg *et. al.*[36] were used for the functional dataset comparison for the solubility screens.

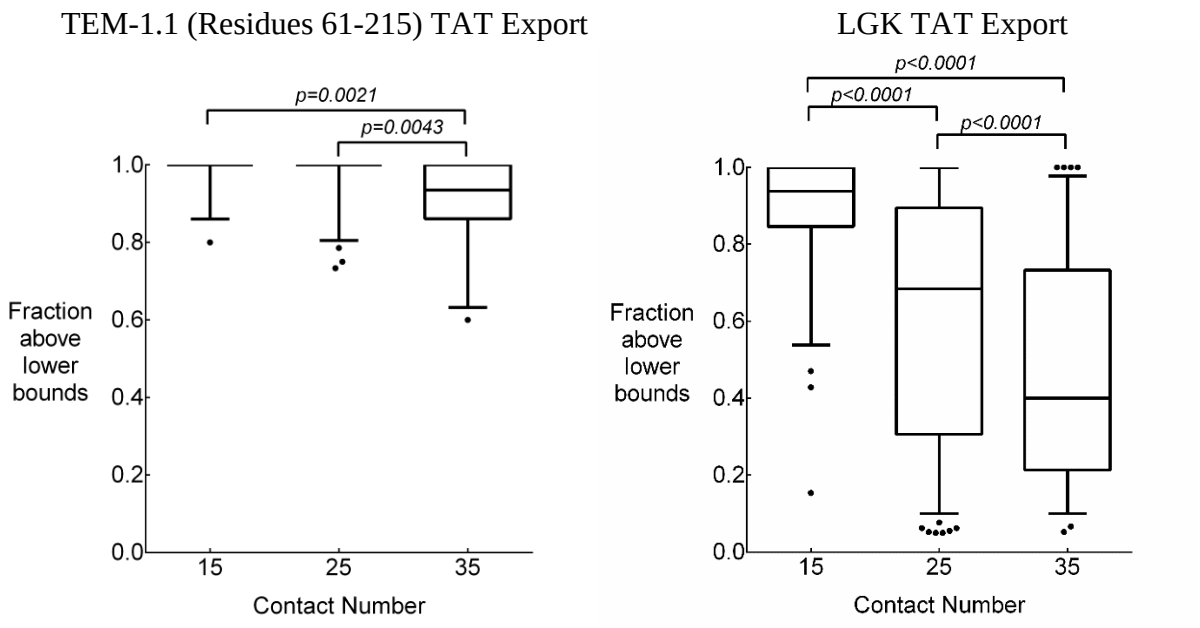


Figure B 3.10: Fraction of mutations above lower bounds versus contact number for TAT export. An unpaired t-test with Welch's correction was performed between bins.

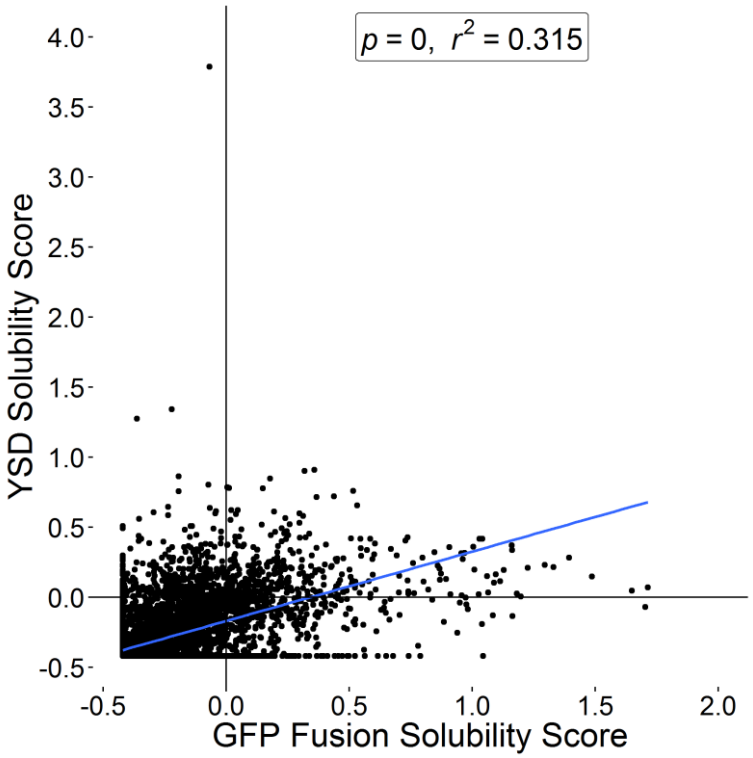


Figure B 3.11: Linear regression of LGK YSD versus GFP solubility screens.

Table B 3.1: Sorting statistics for LGK and TEM-1.1 libraries.

Enzyme	Tile Number	Method	Tile Length (AA)	Events Collected	Percent Sorted (Display)	Percent Sorted (Top)	Theoretical DNA Library Diversity	Fold Oversampling
LGK	1	Yeast Display	103	600,000	23.7	7.7	6,592	91
LGK	2	Yeast Display	110	700,000	25.8	5.2	7,040	99
LGK	3	Yeast Display	110	700,000	21.8	6.8	7,040	99
LGK	4	Yeast Display	105	700,000	19.6	4.6	6,720	104
TEM-1.1	1	Yeast Display	87	500,000	43.0	5.4	5,568	90
TEM-1.1	2	Yeast Display	88	500,000	48.4	6.2	5,632	89
TEM-1.1	3	Yeast Display	88	500,000	50.3	6.0	5,632	89
LGK	1	GFP Fusion	103	700,000	68.6	4.8	6,592	106
LGK	2	GFP Fusion	110	700,000	80.6	4.8	7,040	99
LGK	3	GFP Fusion	110	700,000	80.7	4.3	7,040	99
LGK	4	GFP Fusion	105	700,000	82.7	4.3	6,720	104

Table B 3.2: Deep sequencing library statistics for the yeast display screens.

Screen	Yeast Display													
Enzyme	LGK								TEM-1.1					
Tile Number	1		2		3		4		1		2		3	
Sort Population	Display	Top	Display	Top	Display	Top	Display	Top	Display	Top	Display	Top	Display	Top
Number of mutated codons	103		110		110		105		87		88		88	
Reference sequencing reads post quality filter	607,904		469,478		396,561		259,784		307,817		417,079		413,919	
Selected sequencing reads post quality filter	363,962	512,699	322,740	254,035	329,005	333,191	199,535	288,638	288,082	355,306	491,648	660,166	531,726	441,713
Percent of mutant codons with:														
1-bp substitution	100.0		100.0		99.7		99.7		99.6		99.7		99.7	
2-bp substitution	69.1		79.8		78.6		82.0		81.2		88.0		83.4	
3-bp substitution	63.8		69.2		71.4		75.5		75.0		79.5		77.4	
All substitutions	71.3		78.1		78.5		81.7		81.2		86.0		83.1	
Percent of reads with:														
No nonsynonymous mutations	46.5		45.3		40.1		35.1		43.4		29.2		30.2	
One nonsynonymous mutation	47.3		41.7		51.5		53.6		50.0		58.3		60.1	
Multiple nonsynonymous mutations	6.1		13		8.3		11.3		6.6		12.5		9.7	
Coverage of possible single nonsynonymous mutations	89.5		90.5		90.5		91.8		91.2		96.9		96.3	

Table B 3.3: Deep sequencing library statistics for the TAT pathway selections.

Screen	Tat pathway						
Enzyme	LGK				TEM-1.1		
Tile Number	1	2	3	4	1	2	3
Number of mutated codons	103	110	110	105	87	88	88
Reference sequencing reads post quality filter	512,321	453,663	491,970	167,053	458,755	490,299	402,030
Selected sequencing reads post quality filter	437,570	367,820	514,209	110,898	469,768	450,252	525,271
Percent of mutant codons with:							
1-bp substitution	99.9	100.0	99.9	99.4	99.7	99.6	99.1
2-bp substitution	88.0	96.2	92.0	84.2	86.0	83.6	79.4
3-bp substitution	84.2	93.7	86.4	78.4	80.8	75.0	72.3
All substitutions	88.1	95.7	90.7	83.8	85.8	82.2	79.2
Percent of reads with:							
No nonsynonymous mutations	38.7	35.1	37.6	35.0	28.2	28.3	27.1
One nonsynonymous mutation	51.8	53.9	52.2	52.4	61.2	63.9	62.1
Multiple nonsynonymous mutations	9.5	11.1	10.2	12.6	10.5	7.8	10.7
Coverage of possible single nonsynonymous mutations	96.6	99.4	94.8	85.4	93.3	92.6	90.7

Table B 3.4: Deep sequencing library statistics for the GFP fusion screens.

Screen	GFP Fusion							
Enzyme	LGK							
Tile Number	1		2		3		4	
Sort Population	Display	Top	Display	Top	Display	Top	Display	Top
Number of mutated codons	103		110		110		105	
Reference sequencing reads post quality filter	395,070		305,924		432,015		289,306	
Selected sequencing reads post quality filter	327,144	351,624	350,090	394,152	358,669	469,545	329,904	370,472
Percent of mutant codons with:								
1-bp substitution	99.5		99.9		100.0		99.9	
2-bp substitution	82.1		91.9		92.2		88.9	
3-bp substitution	78.2		85.1		86.9		83.8	
All substitutions	82.9		90.1		91.0		88.3	
Percent of reads with:								
No nonsynonymous mutations	35.3		38.6		36.3		40.8	
One nonsynonymous mutation	55.9		52.6		54.9		51.8	
Multiple nonsynonymous mutations	8.8		8.8		8.8		7.3	
Coverage of possible single nonsynonymous mutations	93.0		94.5		95.6		93.3	

Table B 3.5: Known stabilizing mutations in TEM-1.

Position	Mutation	ΔTm (°C)	Solubility Score YSD	Solubility Score TAT	Reference
31	V31R	3.2	0.19	-0.31	[62]
60	F60Y	2.6	0.64	0.21	[62]
62	P62S	1	0.27	-0.01	[63]
78	G78A	1.5	0.29	1.42	[62]
82	S82H	2.2	0.32	-0.21	[62]
92	G92D	4.1	0.41	0.00	[62]
104	E104K	1.7	0.07	-0.39	[64]
120	R120G	1.8	-0.09	-0.94	[29]
147	E147G	2.6	0.12	-0.43	[29]
153	H153R	3.3	0.23	0.09	[29]
182	M182T	5	0.43	5.10	[29]
201	L201P	1.4	0.26	-0.19	[29]
208	I208M	1.1	0.22	-0.32	[65]
224	A224V	3.1	0.21	-0.38	[63]
235	S235A	1.7	-0.06	-0.42	[66]
265	T265M	1.6	0.30	-0.70	[65]
275	R275L	5	0.59	-0.04	[63]
275	R275Q	2	0.50	0.21	[65]
276	N276D	1.3	0.29	0.24	[65]

Table B 3.6: Known stabilizing mutations in LGK. All mutations and associated biophysical data come from[18]. NS – mutation is not seen in the dataset.

Position	Mutation	ΔTm (°C)	Solubility Score YSD	Solubility Score GFP	Solubility Score TAT
75	P75L	1.4	0.36	-0.05	-0.50
94	R94H	1.9	-0.10	0.09	-0.72
113	H113G	4.9	0.07	1.71	-0.75
135	A135G	2.6	-0.17	0.32	-0.75
140	L140I	2.2	NS	0.26	-0.55
167	I167H	9.8	0.16	1.10	-0.58
194	C194T	6.0	-0.43	0.79	1.50
212	D212A	1.4	0.11	-0.39	-0.67
268	T268C	4.0	0.32	0.85	-0.75
306	A306S	1.1	0.72	0.44	-0.25
359	G359R	1.1	0.15	0.70	-0.71
369	Q369L	3.4	0.15	1.49	-0.75

Table B 3.7: PSSM classifier probabilities independent of a solubility screen.

		Classifier Probabilities		
	n	Neutral	Slightly Deleterious	Deleterious
PSSM (TEM-1.1)				
TOTAL	4997	32%	12%	56%
≥3	187	69%	11%	20%
≥0	1076	66%	14%	20%
PSSM (LGK)				
TOTAL	7701	28%	45%	27%
≥3	377	57%	33%	10%
≥0	1966	52%	37%	12%

Table B 3.8: Classifier probabilities for LGK GFP fusion screen.

LGK GFP Classifier Probabilities

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	421	37%	41%	22%
PSSM				
≥ 3	49	71%	22%	6%
< 3 & ≥ 0	125	49%	41%	10%
< 0	247	24%	45%	31%
Distance				
< 10 Å	159	20%	49%	31%
10-14 Å	114	32%	45%	23%
15-19 Å	87	53%	33%	14%
> 20 Å	61	66%	26%	8%
Contact Number				
≤ 16	65	75%	22%	3%
17-24	189	37%	44%	19%
≥ 25	167	22%	46%	32%
Mutation				
To/From PRO	21	43%	19%	38%

Table B 3.9: Classifier probabilities for chemical changes and size changes.

Classifier Probabilities (TEM-1 YSD)

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	637	37%	8%	55%
Chemical Change				
Polar/Charged to Polar/Charged	195	52%	11%	37%
Charge Reversal	25	40%	16%	44%
Polar/Charge to Hydrophopic/Aromatic	121	44%	6%	50%
Hydrophobic/Aromatic to Polar/Charged	170	16%	6%	78%
To/From Proline	64	13%	5%	83%
Hydrophobic/Aromatic to Hydrophobic/Aromatic	62	63%	8%	29%
Size Change				
Big to Big	184	41%	5%	54%
Big to Small	175	30%	11%	59%
To/From Proline	64	13%	5%	83%
Small to Big	120	49%	8%	43%
Small to Small	94	47%	9%	45%

Classifier Probabilities (LGK-YSD)

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	309	57%	28%	15%
Chemical Change				
Polar/Charged to Polar/Charged	132	70%	19%	11%
Charge Reversal	9	33%	56%	11%
Polar/Charge to Hydrophopic/Aromatic	69	59%	25%	16%
Hydrophobic/Aromatic to Polar/Charged	42	33%	38%	29%
To/From Proline	14	29%	29%	43%
Hydrophobic/Aromatic to Hydrophobic/Aromatic	43	53%	42%	5%
Size Change				
Big to Big	78	51%	32%	17%
Big to Small	82	55%	26%	20%
To/From Proline	14	29%	29%	43%
Small to Big	57	60%	26%	14%
Small to Small	78	69%	26%	5%

Classifier Probabilities (LGK-GFP)

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	421	37%	41%	22%
Chemical Change				
Polar/Charged to Polar/Charged	150	43%	35%	22%
Charge Reversal	11	27%	45%	27%
Polar/Charge to Hydrophopic/Aromatic	89	26%	46%	28%
Hydrophobic/Aromatic to Polar/Charged	102	28%	52%	20%
To/From Proline	21	43%	19%	38%
Hydrophobic/Aromatic to Hydrophobic/Aromatic	48	54%	40%	6%
Size Change				
Big to Big	96	38%	45%	18%
Big to Small	125	36%	40%	24%
To/From Proline	21	43%	19%	38%
Small to Big	83	29%	43%	28%
Small to Small	96	43%	43%	15%

Table B 3.10: Filters and Bayes analyses for LGK YSD and GFP screens.

LGK - YSD					
	Basal	PSSM ≥3	PSSM Filter	Naïve Bayes	Bayes + Filter
n =	309	39	58	242	125
Neutral	57%	82%	90%	66%	77%
Slightly Deleterious	28%	13%	7%	26%	19%
Deleterious	15%	5%	3%	8%	4%

LGK - GFP					
	Basal	PSSM ≥3	PSSM Filter	Naïve Bayes	Bayes + Filter
n =	421	49	34	265	125
Neutral	37%	71%	71%	51%	60%
Slightly Deleterious	41%	22%	26%	40%	33%
Deleterious	22%	6%	3%	9%	7%

Table B 3.11: Inner PCR tile primers. Illumina outer PCR attach point sequences are underlined.

Name	Sequence
pETCONNKFWD	GTTCAGAGTTCTACAGTCCGACGATCAGGGTCGGCTAGC
pETCONNKREV	CCTTGGCACCCGAGAATTCCAAAGCTTTTGTTCGGATC
pSALECTFWD	GTTCAGAGTTCTACAGTCCGACGATCACGTGCGACTGCG
pSALECTREV	CCTTGGCACCCGAGAATTCCATTAACCAGGGTCTCCG
pET29BGFPFWD	GTTCAGAGTTCTACAGTCCGACGATCTTAACTTTAAGAAGGAGATATACAT
pET29BGFPREV	CCTTGGCACCCGAGAATTCCATTCTCCTTTACGCTCGAG
LGKTILE1REV	CCTTGGCACCCGAGAATTCCAGCCGTGCGAAGC
LGKTILE2FWD	GTTCAGAGTTCTACAGTCCGACGATCACCATTGACGCAATC
LGKTILE2REV	CCTTGGCACCCGAGAATTCCACGAACCACTGCGTC
LGKTILE3FWD	GTTCAGAGTTCTACAGTCCGACGATCGGCAACGTGTTCATC
LGKTILE3REV	CCTTGGCACCCGAGAATTCCACCACAATATTCGGGTTATA
LGKTILE4FWD	GTTCAGAGTTCTACAGTCCGACGATCCGGTGGCGCC
TEMTILE1REV	CCTTGGCACCCGAGAATTCCACATGCCATCCGTAAG
TEMTILE2FWD	GTTCAGAGTTCTACAGTCCGACGATCCCAGTCACAGAAAAGCAT
TEMTILE2REV	CCTTGGCACCCGAGAATTCCATGCCGGGAAGCTAG
TEMTILE3FWD	GTTCAGAGTTCTACAGTCCGACGATCATTAACTGGCGAACTACTTACT

Table B 3.12: Crystallographic data processing and refinement statistics for LGK G359R crystallographic structure (values in parentheses refer to the high-resolution shell).

Data Collection	
Space group	$P4_12_12$
Unit cell (Å)	$a = b = 70.06$, $c = 261.77$ $\alpha = \beta = \gamma = 90.00$
Wavelength (Å)	0.9795
Resolution range (Å)	46.33 – 1.80 (1.90 – 1.80)
Total observations	411319
Total unique observations	61638
I/σ_I	9.4 (1.7)
Completeness (%)	99.9 (100.0)
R_{merge}	0.133 (1.045)
R_{pim}	0.056 (0.428)
Redundancy	6.7 (6.9)
Refinement Statistics	
Resolution (Å)	43.08-1.80
Reflections (total)	61556
Reflections (test)	3097
Total atoms refined	3832
Solvent	467
R_{work} (R_{free})	0.17 (0.21)
RMSDs Bond lengths (Å) / angles (º)	0.007/0.828
Ramachandran plot (Favored/allowed(%))	97.4/2.3
Average B, all atoms ($Å^2$)	24.0

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Tartaglia, G. G., and Vendruscolo, M. (2009) Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations, *Molecular BioSystems 5*, 1873-1876.

[2] de Groot, N. S., and Ventura, S. (2010) Protein Aggregation Profile of the Bacterial Cytosol, *PLoS ONE 5*, e9383.

[3] Kellogg, E. H., Leaver‐Fay, A., and Baker, D. (2011) Role of conformational sampling in computing mutation‐induced changes in protein structure and stability, *Proteins: Structure, Function, and Bioinformatics 79*, 830-838.

[4] Potapov, V., Cohen, M., and Schreiber, G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details, *Protein Engineering Design and Selection 22*, 553-560.

[5] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005) The FoldX web server: an online force field, *Nucleic acids research 33*, W382-W388.

[6] Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015) The CamSol method of rational design of protein mutants with enhanced solubility, *Journal of molecular biology 427*, 478-490.

[7] Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., and Sheffler, W. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, *Methods in Enzymology 487*, 545.

[8] Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., and Ventura, S. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures, *Nucleic acids research*, gkv359.

[9] Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B. L. (2009) Design of therapeutic proteins with enhanced stability, *Proceedings of the National Academy of Sciences 106*, 11937-11942.

[10] Goldenzweig, A., Goldsmith, M., Hill, Shannon E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, Raquel L., Aharoni, A., Silman, I., Sussman, Joel L., Tawfik, Dan S., and Fleishman, Sarel J. (2016) Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability, *Molecular Cell 63*, 337-346.

[11] Waldo, G. S. (2003) Genetic screens and directed evolution for protein solubility, *Current Opinion in Chemical Biology 7*, 33-38.

[12] Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein, *Nat Biotech 17*, 691-695.

[13] Cabantous, S., and Waldo, G. S. (2006) In vivo and in vitro protein solubility assays using split GFP, *Nat Meth 3*, 845-854.

[14] Park, S., Xu, Y., Stowell, X. F., Gai, F., Saven, J. G., and Boder, E. T. (2006) Limitations of yeast surface display in engineering proteins of high thermostability, *Protein Engineering Design and Selection 19*, 211-217.

[15] Ellgaard, L., and Helenius, A. (2003) Quality control in the endoplasmic reticulum, *Nat Rev Mol Cell Biol 4*, 181-191.

[16] Burns, M. L., Malott, T. M., Metcalf, K. J., Hackel, B. J., Chan, J. R., and Shusta, E. V. (2014) Directed Evolution of Brain-Derived Neurotrophic Factor for Improved Folding and Expression in Saccharomyces cerevisiae, *Applied and Environmental Microbiology 80*, 5732-5742.

[17] Traxlmayr, M. W., and Obinger, C. (2012) Directed evolution of proteins for increased stability and expression using yeast display, *Archives of Biochemistry and Biophysics 526*, 174-180.

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

[19] Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability, *Proceedings of the National Academy of Sciences 103*, 5869-5874.

[20] Bloom, J. D., and Arnold, F. H. (2009) In the light of directed evolution: Pathways of adaptive protein evolution, *Proceedings of the National Academy of Sciences 106*, 9995-10000.

[21] Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability, *Current Opinion in Structural Biology 19*, 596-604.

[22] Tawfik, O. K., and Dan, S. (2010) Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective, *Annual Review of Biochemistry 79*, 471-505.

[23] Schlinkmann, K. M., Hillenbrand, M., Rittner, A., Künz, M., Strohner, R., and Plückthun, A. (2012) Maximizing Detergent Stability and Functional Expression of a GPCR by Exhaustive Recombination and Evolution, *Journal of Molecular Biology 422*, 414-428.

[24] Schlinkmann, K. M., Honegger, A., Türeci, E., Robison, K. E., Lipovšek, D., and Plückthun, A. (2012) Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations, *Proceedings of the National Academy of Sciences 109*, 9810-9815.

[25] Boock, J. T., King, B. C., Taw, M. N., Conrado, R. J., Siu, K.-H., Stark, J. C., Walker, L. P., Gibson, D. M., and DeLisa, M. P. (2015) Repurposing a Bacterial Quality Control Mechanism to Enhance Enzyme Production in Living Cells, *Journal of molecular biology 427*, 1451-1463.

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

[27] Wang, X., Minasov, G., and Shoichet, B. K. (2002) Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs, *Journal of Molecular Biology 320*, 85-95.

[28] Beadle, B. M., and Shoichet, B. K. (2002) Structural Bases of Stability–function Tradeoffs in Enzymes, *Journal of Molecular Biology 321*, 285-296.

[29] Bershtein, S., Goldin, K., and Tawfik, D. S. (2008) Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins, *Journal of Molecular Biology 379*, 1029-1044.

[30] Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L., and van Loon, A. P. G. M. (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase, *Protein Engineering 13*, 49-57.

[31] Steipe, B., Schiller, B., Plückthun, A., and Steinbacher, S. (1994) Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain, *Journal of Molecular Biology 240*, 188-192.

[32] Araya, C. L., and Fowler, D. M. (2011) Deep mutational scanning: assessing protein function on a massive scale, *Trends in biotechnology 29*, 435-442.

[33] Fowler, D. M., and Fields, S. (2014) Deep mutational scanning: a new style of protein science, *Nature methods 11*, 801-807.

[34] Fowler, D. M., Stephany, J. J., and Fields, S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning, *Nat. Protocols 9*, 2267-2284.

[35] Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate, *Journal of molecular biology 425*, 1363-1377.

[36] Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape, *Molecular Biology and Evolution 31*, 1581-1592.

[37] Stiffler, Michael A., Hekstra, Doeke R., and Ranganathan, R. (2015) Evolvability as a Function of Purifying Selection in TEM-1 beta-Lactamase, *Cell 160*, 882-892.

[38] Hietpas, R., Roscoe, B., Jiang, L., and Bolon, D. N. A. (2012) Fitness analyses of all possible point mutations for regions of genes in yeast, *Nat. Protocols 7*, 1382-1396.

[39] Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2011) Experimental illumination of a fitness landscape, *Proceedings of the National Academy of Sciences 108*, 7896-7901.

[40] Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D. S. (2007) The Stability Effects of Protein Mutations Appear to be Universally Distributed, *Journal of Molecular Biology 369*, 1318-1332.

[41] Kiss, C., Temirov, J., Chasteen, L., Waldo, G. S., and Bradbury, A. R. M. (2009) Directed evolution of an extremely stable fluorescent protein, *Protein Engineering Design and Selection 22*, 313-323.

[42] Wrenbeck, E. E., Klesmith, J. R., Stapleton, J. A., Adeniran, A., Tyo, K. E. J., and Whitehead, T. A. Plasmid-based single-pot saturation mutagenesis, *Under Review*.

[43] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N., and Whitehead, T. A. (2015) High-resolution sequence-function mapping of full-length proteins, *PLoS One 10*, e0118193.

[44] Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M., and Wittrup, K. D. (2006) Isolating and engineering human antibodies using yeast surface display, *Nat. Protocols 1*, 755-768.

[45] Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E., Tomek, K. J., and Whitehead, T. A. (2014) The interrelationship between promoter strength, gene expression, and growth rate, *PLoS One 9*, e109105.

[46] Fisher, A. C., Kim, W., and Delisa, M. P. (2006) Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway, *Protein Science 15*, 449-458.

[47] Fleishman, S. J., Whitehead, T. A., Strauch, E.-M., Corn, J. E., Qin, S., Zhou, H.-X., Mitchell, J. C., Demerdash, O. N. A., Takeda-Shitaka, M., Terashi, G., Moal, I. H., Li, X., Bates, P. A., Zacharias, M., Park, H., Ko, J.-s., Lee, H., Seok, C., Bourquard, T., Bernauer, J., Poupon, A., Azé, J., Soner, S., Ovalı, Ş. K., Ozbek, P., Tal, N. B., Haliloglu, T., Hwang, H., Vreven, T., Pierce, B. G., Weng, Z., Pérez-Cano, L., Pons, C., Fernández-Recio, J., Jiang, F., Yang, F., Gong, X., Cao, L., Xu, X., Liu, B., Wang, P., Li, C., Wang, C., Robert, C. H., Guharoy, M., Liu, S., Huang, Y., Li, L., Guo, D., Chen, Y., Xiao, Y., London, N., Itzhaki, Z., Schueler-Furman, O., Inbar, Y., Potapov, V., Cohen, M., Schreiber, G., Tsuchiya, Y., Kanamori, E., Standley, D. M., Nakamura, H., Kinoshita, K., Driggers, C. M., Hall, R. G., Morgan, J. L., Hsu, V. L., Zhan, J., Yang, Y., Zhou, Y., Kastritis, P. L., Bonvin, A. M. J. J., Zhang, W., Camacho, C. J., Kilambi, K. P., Sircar, A., Gray, J. J., Ohue, M., Uchikoga, N., Matsuzaki, Y., Ishida, T., Akiyama, Y., Khashan, R., Bush, S., Fouches, D., Tropsha, A., Esquivel-Rodríguez, J., Kihara, D., Stranges, P. B., Jacak, R., Kuhlman, B., Huang, S.-Y., Zou, X., Wodak, S. J., Janin, J., and Baker, D.

(2011) Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology, *Journal of Molecular Biology 414*, 289-302.

[48] Klesmith, J. R., and Whitehead, T. A. (2016) High-throughput evaluation of synthetic metabolic pathways, *TECHNOLOGY 04*, 9-14.

[49] Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes, *Nucleic Acids Research 42*, e112-e112.

[50] Ambler, R. P., Coulson, A. F., Frère, J. M., Ghuysen, J. M., Joris, B., Forsman, M., Levesque, R. C., Tiraby, G., and Waley, S. G. (1991) A standard numbering scheme for the class A beta-lactamases, *Biochemical Journal 276*, 269-270.

[51] Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., and Baker, D. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing, *Nat Biotechnol 30*, 543-548.

[52] Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants, *Bioinformatics 27*, 3430-3431.

[53] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool, *Journal of Molecular Biology 215*, 403-410.

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

[55] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research 32*, 1792-1797.

[56] Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577-2637.

[57] Altschul, S. F., Gertz, E. M., Agarwala, R., Schäffer, A. A., and Yu, Y.-K. (2009) PSI-BLAST pseudocounts and the minimum description length principle, *Nucleic Acids Research 37*, 815-824.

[58] Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. In Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography, *Newsletter on Protein Crystallography 26*.

[59] Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics, *Acta Crystallogr D Biol Crystallogr 60*, 2126-2132.

[60] Chen, V. B., Arendall, W. B., III, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom

structure validation for macromolecular crystallography, *Acta Crystallographica Section D 66*, 12-21.

[61] DeLano, W. L. (2002) The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, USA.

[62] Deng, Z., Huang, W., Bakkalbasi, E., Brown, N. G., Adamski, C. J., Rice, K., Muzny, D., Gibbs, R. A., and Palzkill, T. (2012) Deep Sequencing of Systematic Combinatorial Libraries Reveals β-Lactamase Sequence Constraints at High Resolution, *Journal of Molecular Biology 424*, 150-167.

[63] Kather, I., Jakob, R. P., Dobbek, H., and Schmid, F. X. (2008) Increased Folding Stability of TEM-1 β-Lactamase by In Vitro Selection, *Journal of Molecular Biology 383*, 238-251.

[64] Raquet, X., Vanhove, M., Lamotte-Brasseur, J., Goussard, S., Courvalin, P., and Frère, J. M. (1995) Stability of TEM β-lactamase mutants hydrolyzing third generation cephalosporins, *Proteins: Structure, Function, and Bioinformatics 23*, 63-72.

[65] Brown, N. G., Pennington, J. M., Huang, W., Ayvaz, T., and Palzkill, T. (2010) Multiple Global Suppressors of Protein Stability Defects Facilitate the Evolution of Extended-Spectrum TEM β-Lactamases, *Journal of Molecular Biology 404*, 832-846.

[66] Dubus, A., Wilkin, J. M., Raquet, X., Normark, S., and Frère, J. M. (1994) Catalytic mechanism of active-site serine β-lactamases: role of the conserved hydroxy group of the Lys-Thr(Ser)-Gly triad, *Biochemical Journal 301*, 485-494.

CHAPTER 4

Interpreting deep mutational scanning data resulting from selections on solid media

Sarah Thorwall was an undergraduate researcher who performed plate assays in this work.

ABSTRACT

Deep mutational scanning is now used in directed evolution experiments to quantify the change in frequency of a cellular variant in a mixed population. A key concern is the extent to which the enrichment of a variant in a population corresponds to a fitness metric like relative growth rate or survival percentage. Analytical equations converting the enrichment of a variant to fitness metrics for plate-based selections are presented here. Using isogenic and mixed cultures I show that growth rates and survival percentages correlate for antibiotic plate-based selections. These results are important for proper interpretation of data resulting from deep sequencing.

INTRODUCTION

Deep sequencing has emerged as a powerful, enabling tool for protein engineering[1-4]. Deep sequencing-based measurements allow one to estimate the frequency of each mutational variant in a population to be screened or selected[5, 6]. The end-point measurement is an enrichment ratio (ε_i), defined as the base 2 logarithm of the frequency change of the variant i in the selected population compared to a reference population. A key question in such deep mutational scanning experiments is the extent to which this enrichment ratio corresponds to a fitness value or phenotype[7, 8].

Many directed evolution experiments involve selections on solid media. Typically, a population of cells expressing the mutated protein of interest is plated on a solid support containing selective media, and the selected hits are colonies that survive or are larger than other colonies after a set amount of time. A major problem with plate-based selections is that the output is binary – a hit or not a hit. To remedy this, several groups have used deep sequencing to determine an analog fitness for each variant[9-12] where the resulting fitness metric is usually normalized to the enrichment ratio of the starting sequence. However, the fitness metric calculated from enrichment ratios should depend strongly on whether the relative growth rates between variants differ, whether a variant survives the initial plating condition differentially, and the initial and final biomass concentrations. To that end, the objective of this work is to present analytical equations converting enrichment ratios to unambiguous fitness metrics for plate-based selections, and to supply experimental validation using an existing genetic selection with TEM-1 *bla* for an antibiotic-based selection.

THEORY

On solid media, growth of bacterial biomass follows exponential behavior during the growth phase[13]. Importantly, the specific growth rate during exponential phase and the final density is independent of the plating density[13]. Therefore, for our model it can be assumed that the growth model for each variant i can be written as:

$$X_i = X_o f_i e^{\mu_i (t - t_{lag})} \quad (1)$$

where X_o is the initial biomass for variant i, X_i is the biomass of variant i evaluated at time t, t_{lag} is the lag time, μ_i is the specific growth rate, and f_i is the fraction of variant i that survives the initial selection at t=0. Based on work from van Heerden et al, it can be assumed that the lag time is the same for all variants[14]. Equation (1) does not capture the characteristic sigmoidal shape of microbial growth curves – this is neglected for simplicity. Note that the inflection point on microbial growth curves typically occurs within 1 or 2 average population doublings of the maximum biomass concentration. Thus, the error resulting from this simplification is likely to be minimal.

Using the above exponential growth model, it can be shown for conditions where t is greater than t_{lag} and X_i is less than X_{max} that the enrichment ratio of a given clone in the population can be determined by μ_i and f_i:

$$\varepsilon_i = g_p \left(\frac{\mu_i}{\bar{\mu}} - 1 \right) + log_2 \left(\frac{f_i}{\bar{f}} \right) \quad (2)$$

here $\bar{\mu}$ and $\bar{f}$ are the population-averaged values, and g_p is the population-averaged number of doublings. Importantly, these three parameters are experimentally measurable and, furthermore, common to all variants. Equation (2) shows that a given variant is enriched in the population if the specific growth rate is faster than the population average and/or if the fraction of surviving colonies is higher than the population average.

There are two limiting cases. If the fraction of growing cells is the same for each variant, as may be the case for selections coupling growth with flux through primary metabolism[2] then the fitness equation becomes[8]:

$$log_2\left(\frac{\mu_i}{\mu_{wt}}\right) = log_2\left(\frac{\frac{\varepsilon_i}{g_p}+1}{\frac{\varepsilon_{wt}}{g_p}+1}\right) \quad (3)$$

At the other limit, if the exponential growth rates are equivalent between all variants, the enrichment ratios can be normalized to a fitness metric defined as:

$$log_2\left(\frac{f_i}{f_{wt}}\right) = \left(\frac{\varepsilon_i}{\varepsilon_{wt}}\right) \quad (4)$$

Here the wt subscript refers to the values from the wild-type sequence. While it is often implicitly assumed that equation (4) is the appropriate fitness metric for selections[10], it could be the case that both the fraction of surviving variants and the relative growth rates vary between variants in the selection. In this case, time points must be taken over a time interval in order to calculate the specific growth rate and the surviving fraction for each variant.

To evaluate the appropriate form of the normalization expressions for antibiotic-based selections, I evaluated a genetic selection exploiting the twin-arginine translocation (TAT) pathway in Gram-negative bacteria[15]. In this selection a protein of interest is fused between an N-terminal ssTorA Tat periplasmic export signal peptide and a truncated, active C-terminal TEM-1 beta-lactamase. Because the TAT pathway is thought to export only folded proteins, TEM-1 will be differentially exported to the periplasm based on the protein of interest in the fusion construct. Thus, mutations conferring enhanced fusion protein periplasmic localization can be selected on solid media containing different amounts of beta-lactam antibiotics.

RESULTS

I first evaluated selection-specific growth parameters for a ΔS4-A25 TEM-1 *bla* with the activity abrogating mutation S70A. I modified the TAT selection plasmid pSALECT-EcoBam (Addgene: #59705)[9] by fusing a codon-swapped TEM-1 *bla* ΔS4-A25 in-frame after the XhoI restriction site. This active codon-swapped TEM-1 *bla* is used to avoid recombination with the N-terminal TEM-1 *bla* S70A. *E. coli* strain MC4100 harboring this plasmid was plated on LB-agar plates containing 50, 100, or 200 µg/mL carbenicillin.

Colonies from a fresh transformation of *E. coli* MC4100 with the plasmid pSALECT-TEM-1(S70A)/csTEM-1 were used to start a culture in liquid LB media with 34 µg/mL chloramphenicol. This culture was grown at 30ºC at 250 rpm for 10 hours. Fresh LB agar plates with 50, 100, or 200 µg/mL carbenicillin were made the day of the transformation and poured at a constant volume of agar per plate. The OD_{600} of the culture was measured after the 10 hour growth period to determine the cell density. The cells were diluted such that between 12-21 hours the cells were in exponential growth phase and plated. This initial plating density is different for each antibiotic concentration and was determined by control experiments. 150 µL of the diluted culture was spread onto the desired LB agar + carbenicillin plates. Plates were then placed into a humidified 30ºC incubator and grown for 12, 15, 18, and 21 hours. At these time points a plate was taken out, and 1 mL of phosphate buffered saline was added onto the plate. All of the cells were scraped off of the agar plate and resuspended in this phosphate buffered saline. The final OD_{600} was then measured to determine the final cell mass. The specific growth rate was calculated from the natural log transformed cell densities at different time points. This experiment was repeated at least twice.

The fraction of surviving clones was determined by making serial dilutions of a culture with an OD_{600} of 1.0 ranging from 10^{-3} to 10^{-7} onto the selective LB agar + carbenicillin plates and LB agar + 34 µg/mL chloramphenicol plates. Plates were incubated at 30ºC in a humidified incubator and the number of distinct colonies recorded at each dilution. The fraction of surviving clones was calculated by dividing the number of distinct colonies on each carbenicillin plate by the number on the chloramphenicol plate. This was done with at least three biological replicates per antibiotic concentration.

For TEM-1 (S70A) the specific growth rate decreases at higher antibiotic concentrations (**Figure 4.1a**). This decreased growth rate correlates with low viability, although the growth rate plateaus at approximately half of the specific growth rate under conditions of high viability (**Figure 4.1a**). I performed the same experiment on a destabilized, catalytically inactive variant TEM-1 *bla* with mutations S70A, D179G[16]. As expected, the negative change in cellular viability is much greater than with TEM-1(S70A) (**Figure 4.1b**). Similar to TEM-1(S70A), the growth rate decreases and plateaus to around half that of the high viability growth rate (**Figure 4.1b**).

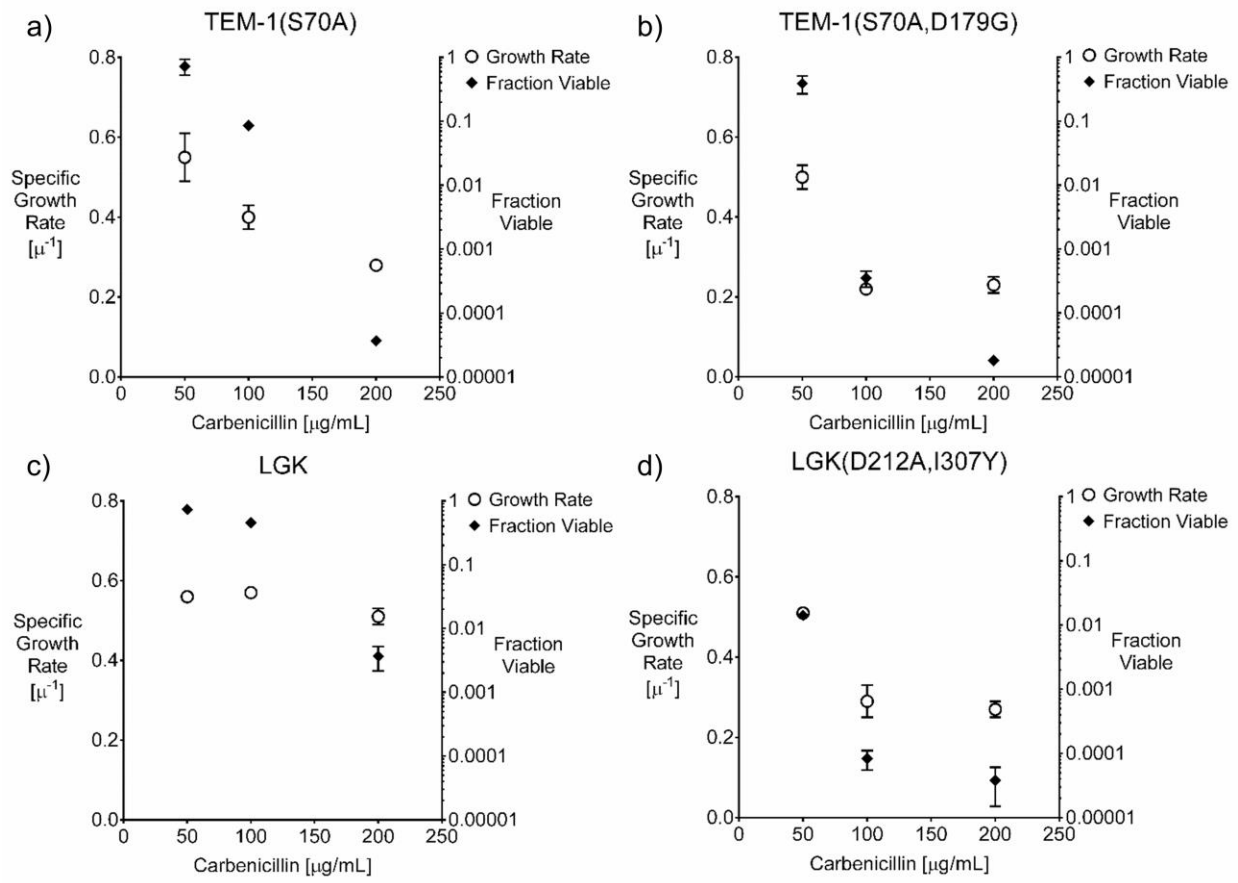


Figure 4.1: Specific growth rate (circles) and fraction viable (diamonds) of *E. coli* MC4100 expressing TEM-1 or LGK variants. a) TEM-1(S70A), b) TEM-1(S70A,D179G), c) LGK, and d) LGK(D212A,I307Y). Error bars are 1 standard deviation of biological replicates (growth rates) or triplicates (fraction viable).

I performed the same set of experiments on a different enzyme system to confirm my initial observations. I used a codon optimized levoglucosan kinase (LGK) from *L. starkeyi*[2] (**Figure 4.1c**); and a destabilized, catalytically inactive variant LGK with mutations D212A, I307Y (**Figure 4.1d**). Similar results were seen for both strains where the growth rate decrease is independent of viability, the growth rate plateaus to half maximum for the destabilized variant, and viability has a more substantial decrease than growth rate. Therefore, both the growth rate and cellular viability impact enrichment ratios for plate based selections. From these results I predict that within a population the enrichment ratios for beneficial variants should increase with

population size, and this effect should be accounted for in the analysis of deep mutational scanning experiments.

To evaluate gain of function mutations on a larger scale, nicking mutagenesis[17] was used to create a single-site saturation mutagenesis library for residues 331-435 in LGK. A TAT selection was performed where 100 µL of a culture of MC4100 *E. coli* at an OD_{600} of 1.0 with the LGK library was plated on two 100 mm diameter Petri plates with LB agar and 200 µg/mL carbenicillin per time point. Dilutions were also plated at 200 µg/mL carbenicillin and 34 µg/mL chloramphenicol to measure library viability. The plates were incubated in parallel at 30ºC in a humidified incubator, and the cellular mass was scraped and collected at time-points of 12, 14, 16, and 18 hours. The two replicates at each time point were pooled in equal volumes. The plasmids extracted from each time point were deep sequenced using an Illumina MiSeq in 300 bp paired-end mode using previously developed library preparation procedures[8]. Enrich[18] was then used to calculate enrichment ratios for each variant at a given time-point relative to t = 0 hours. 1,220 single point mutants with at least 15 read counts in the reference were observed (unselected; t = 0 hours) population. Library statistics for the selections are shown in **Table 4.1**. Processed deep sequencing datasets are freely available at figshare (www.figshare.com).

Table 4.1: Library statistics, cellular densities, and fraction viable of time points.

Enzyme	LGK	
Number of mutated codons	105	
Reference sequencing reads post quality filter	167,053	
Selected sequencing reads post quality filter		
12 hours	110,898	
14 hours	184,418	
16 hours	186,130	
18 hours	182,375	
Percent of mutant codons with:		
1-bp substitution	99.4	
2-bp substitution	84.2	
3-bp substitution	78.4	
All substitutions	83.8	
Percent of reads with:		
No nonsynonymous mutations	35.0	
One nonsynonymous mutation	52.4	
Multiple nonsynonymous mutations	12.6	
Coverage of possible single nonsynonymous mutations	85.4	
Biomass (OD_{600}-mL):		
0 hours	0.1	0.1
12 hours	26.2	28.4
14 hours	43.1	42.8
16 hours	52.0	58.0
18 hours	75.4	85.2
Fraction viable:		
12 hours	0.0074	
14 hours	0.0126	
16 hours	0.0145	
18 hours	0.0153	

For each variant the enrichment ratio as a function of the observed average number of population doublings was plotted (**Figure 4.2a**). From this plot two parameters were calculated: an enrichment ratio slope ($slope_{er}$) as well as an average enrichment ratio ($average_{er}$), defined here as the enrichment ratio at the midpoint of the best-fit linear regression line joining the four

experimental data-points. The correlation coefficient between $slope_{er}$ and $average_{er}$ is only 0.38 (**Figure 4.2b**). However, there is a statistically significant relationship between the sign of $slope_{er}$ and $average_{er}$ (Fisher's exact test binary classification $p<0.0001$). That is, if the $slope_{er}$ is positive, the $average_{er}$ is much more likely to be positive, and vice versa.

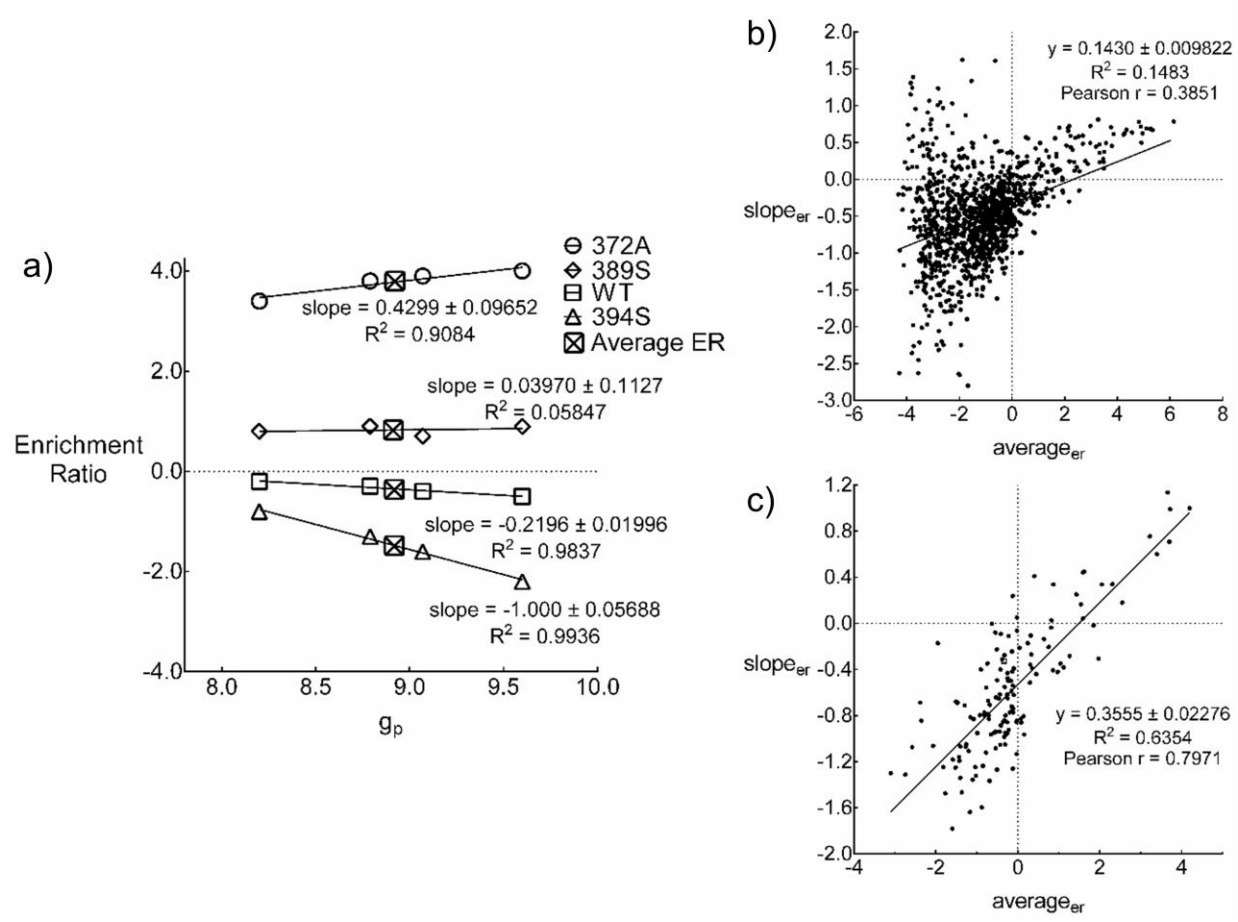


Figure 4.2: **Enrichment ratio versus average population doublings and the relationship between the change in enrichment ratio and average enrichment ratio.** a) Enrichment ratio versus average population doublings of example variants showing an increase, neutral, or decrease in their enrichment over time. b,c): relationship between the change in enrichment ratio ($slope_{er}$) and average enrichment ratio ($average_{er}$) for (b) all variants above 15 read counts; and (c) all the subset of "high confidence" variants. Wild-type is indicated with an open square.

I reasoned that intrinsic counting error resulting from counting small numbers of variants could impact accurate determination of slopes. To test this assumption, the above analysis was performed again on the subset of variants with over 100 read counts in the reference population

(t = 0 hours) and at least 50 read counts on average in the four subsequent timepoints. For these resulting 142 variants, a much stronger relationship emerged between $slope_{er}$ and $average_{er}$ (**Figure 4.2c**), with the correlation coefficient now 0.80.

CONCLUSION

The above results show that enrichment ratios vary with respect to average number of population doublings for the plate-based TAT genetic selection in *E. coli*. I speculate enrichment ratios will vary for most coupled selections involving beta-lactam antibiotic resistance. For this particular antibiotic-based selection, these results are consistent with growth rate being a non-linear function of cell viability as shown here for isogenic cultures. While viability and growth rate may or may not be coupled for other types of plate-based selections, the above results have strong implications in the interpretation of deep mutational scanning data resulting from selections on solid media. In particular, implicit assumptions about the conversion of enrichment ratios to a fitness metric should be experimentally demonstrated. Furthermore, accurate determination of the slope is only apparent with variants well sampled in the population. As a practical matter, more depth of coverage for many deep mutational scanning experiments may be warranted.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., and Baker, D. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing, *Nat Biotechnol 30*, 543-548.

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

[3] Jardine, J. G., Kulp, D. W., Havenar-Daughton, C., Sarkar, A., Briney, B., Sok, D., Sesterhenn, F., Ereño-Orbea, J., Kalyuzhniy, O., Deresa, I., Hu, X., Spencer, S., Jones, M., Georgeson, E., Adachi, Y., Kubitz, M., deCamp, A. C., Julien, J.-P., Wilson, I. A., Burton, D. R., Crotty, S., and Schief, W. R. (2016) HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen, *Science 351*, 1458.

[4] Koenig, P., Lee, C. V., Sanowar, S., Wu, P., Stinson, J., Harris, S. F., and Fuh, G. (2015) Deep Sequencing-guided Design of a High Affinity Dual Specificity Antibody to Target Two Angiogenic Factors in Neovascular Age-related Macular Degeneration, *Journal of Biological Chemistry 290*, 21773-21786.

[5] Fowler, D. M., Stephany, J. J., and Fields, S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning, *Nat. Protocols 9*, 2267-2284.

[6] Hietpas, R., Roscoe, B., Jiang, L., and Bolon, D. N. A. (2012) Fitness analyses of all possible point mutations for regions of genes in yeast, *Nat. Protocols 7*, 1382-1396.

[7] Boucher, J. I., Bolon, D. N. A., and Tawfik, D. S. (2016) Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature, *Protein Science 25*, 1219-1226.

[8] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N., and Whitehead, T. A. (2015) High-resolution sequence-function mapping of full-length proteins, *PLoS One 10*, e0118193.

[9] Hsiau, T. H., Sukovich, D., Elms, P., Prince, R. N., Strittmatter, T., Ruan, P., Curry, B., Anderson, P., Sampson, J., and Anderson, J. C. (2015) A method for multiplex gene synthesis employing error correction based on expression, *PLoS One 10*, e0126078.

[10] Elazar, A., Weinstein, J., Biran, I., Fridman, Y., Bibi, E., and Fleishman, S. J. (2016) Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane, *eLife 5*, e12125.

[11] Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape, *Molecular Biology and Evolution 31*, 1581-1592.

[12] Kim, I., Miller, C. R., Young, D. L., and Fields, S. (2013) High-throughput Analysis of in vivo Protein Stability, *Molecular & Cellular Proteomics 12*, 3370-3378.

[13] Fujikawa, H., and Morozumi, S. (2005) Modeling Surface Growth of Escherichia coli on Agar Plates, *Applied and Environmental Microbiology 71*, 7920-7926.

[14] van Heerden, J. H., Wortel, M. T., Bruggeman, F. J., Heijnen, J. J., Bollen, Y. J. M., Planqué, R., Hulshof, J., O'Toole, T. G., Wahl, S. A., and Teusink, B. (2014) Lost in Transition: Start-Up of Glycolysis Yields Subpopulations of Nongrowing Cells, *Science 343*.

[15] Fisher, A. C., Kim, W., and Delisa, M. P. (2006) Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway, *Protein Science 15*, 449-458.

[16] Wang, X., Minasov, G., and Shoichet, B. K. (2002) Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs, *Journal of Molecular Biology 320*, 85-95.

[17] Wrenbeck, E. E., Klesmith, J. R., Stapleton, J. A., Adeniran, A., Tyo, K. E. J., and Whitehead, T. A. Plasmid-based single-pot saturation mutagenesis, *Accepted*.

[18] Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants, *Bioinformatics 27*, 3430-3431.

CHAPTER 5

Concluding Remarks

PERSPECTIVE

My dissertation provides an approach to comprehensively study the effect of thousands of mutations on the function of enzymes in a designed metabolic pathway. It also provides an approach to improve their soluble expression in the absence of a growth selection. These studies provide a framework for additional investigations by other researchers. Below, I provide a summary of my results with greater perspective and describe broader impact of the work.

At the time of developing the method described in Chapter 2, this work was the first example of applying population-based deep sequencing to study an enzyme in a designed pathway. The only other examples were enzymes that confer antibiotic resistance through a plate-based selection[1, 2]. The present work studied levoglucosan kinase from *L. starkeyi* which converts levoglucosan, a anhydrosugar from fast pyrolysis, into glucose-6-phosphate[3]. This is a potential route to convert woody biomass deconstructed by fast pyrolysis into biochemicals of choice via fermentation[3-5]. To introduce this enzyme into *E. coli* a synthetic promoter collection was created and searched[6, 7] such that growth was possible on minimal media with levoglucosan as the sole carbon source. Because of the poor kinetics and thermodynamic stability of levoglucosan kinase, only strong synthetic promoters were able to support weak growth at 37ºC in *E. coli* showing a clear need to improve the enzyme itself.

This work detailed the development and the requirements that a growth selection needs to satisfy for a population-based growth study: 1) that an active enzyme is required for growth, 2) the selection can resolve changes in growth rate, and 3) the change in enzyme activity changes the cellular growth rate. These requirements are generally applicable to other enzymes and should serve as guidelines for further research by outside labs. Using this growth selection combined with deep mutational scanning[8, 9], it was possible to quantify the enrichment of over

8,000 single point mutations of levoglucosan kinase in this designed pathway. Using analytical equations that I provided assistance on[10], the enrichments were converted into growth rates to be used in a fitness metric. The method was validated by deep sequencing biological replicates of the population, and testing that the fitness metric correlates with growth rates and lysate enzyme activity of single isogenic cultures. From the first selection, over 200 beneficial mutations were identified scattered spatially throughout the enzyme structure. This number would not be typically found from studies using current practices by the field as outlined in Chapter 1.

Computational design combined with deep sequencing datasets enabled the creation of enzyme designs that incorporated 35 to 57 beneficial mutations which increased the $T_{m,app}$ by 47.8ºC in the best case. However, all designs were inactive and the solved crystal structure of one was globally alignable to the wild-type structure and no change was observed to the active site residues. Single variants were tested *in vitro* for their catalytic efficiency and thermodynamic stability which indicated that there is a general tradeoff between the two. Furthermore, improvement of stability was the primary reason for being enriched in the population as the wild-type enzyme has a T_m below the selection temperature. A more modest enzyme design, LGK.1, was made with only three mutations and had wild-type catalytic fitness and improved stability. Cells expressing this enzyme supported higher growth rates and metabolic flux which required weaker expression. Comprehensive datasets starting from this enzyme compared to the original datasets predicted beneficial mutations that did not trade off catalytic fitness for stability. This predictive ability was used to make the final enzyme design, LGK.9, which supported a 15-fold improvement in growth rate and over 24-fold improvement in lysate activity over wild-type levels. A key take-away from this work is that metabolic pathway performance can be significantly improved just by increasing the soluble expression of the enzyme alone (in the case

of LGK.1 and partially LGK.9). A second lesson from this project is that more modest additions of mutations increase the chance of not inactivating the enzyme, however a non-trivial amount is still required to be incorporated to see significant performance improvements. These two takeaways drive the project outlined in Chapter 3.

Chapter 3 addressed the problem of improving stability while retaining catalytic activity. The method outlined in Chapter 2 works extremely well for enzymes in primary metabolism that are required for growth. However, there is a particular interest to improve enzymes from plants or other organisms that are incorporated into secondary metabolism in production hosts. For these enzymes a growth selection that requires active enzyme may not exist. The project in Chapter 3 set out to address this problem utilizing the lessons from the project in Chapter 2. Three complementary screens that have been used to improve soluble expression of proteins were tested on two enzymes with comprehensive functional datasets[1, 11]. In particular, yeast surface display[12], GFP fusion screening[13], and Tat export[14] were used for TEM-1 beta-lactamase[1] and LGK[11]. Problems of each screening method were diagnosed and are outlined for future applications. In particular, accessibility of the c-myc epitope for yeast surface display and the fusion to the c-terminal beta-lactamase for Tat export lead to false positives in the dataset.

Classifiers were developed and trained on the functional datasets to allow identification of mutations beneficial for solubility improvements that retain wild-type catalytic fitness. Furthermore, these classifiers were built such that a rough homology model could be used as many difficult to express enzymes do not have solved crystal structures. It was found that a combination of PSSM (a method using existing genomic datasets)[15], contact number (the average number of amino acid contacts in a defined distance)[16], distance to active site, and mutation type was able to identify solubility enhancing mutations that retained wild-type catalytic fitness with

90% accuracy. Alternately, the rate of picking a mutation that is strongly deleterious for catalytic fitness is 2%, which is a low enough for 10 to 20 mutations to be incorporated into an enzyme design and maintain activity. However, the problem with strict classification filters leads to culling Pareto optimal variants as introduced in Chapter 1. One example of this was LGK G359R which was near the active site and the solved structure indicates it interacts with the ATP/ADP moiety. Future work from this method should be to apply the solubility screens and classifiers on enzymes that do not have solved structures and are part of important secondary metabolic pathways.

The final project in Chapter 4 is an extension of the solubility screening project. Early in my career I contributed to another project to create fitness metrics for liquid media growth selections and FACS (used here in Chapter 2 and Chapter 3 respectively)[10]. However, equations did not exist for plate-based selections like the Tat export screen. Therefore, the final project set out to develop equations to convert the enrichment of a variant to a fitness metric. Isogenic cultures expressing variants of TEM-1 beta-lactamase and LGK were measured for their specific growth rate on solid media and their fraction of survival. And a population-based time course study was performed with LGK. Both the population study and isogenic cultures are in agreement that the specific growth rate and surviving fraction after plating are important when converting the enrichment of a variant to a fitness metric. This study should provide a framework for other plate-based deep mutational scanning studies when calculating fitness of variants.

In total, my results have led to significant advances in methodological approaches to study enzymes and discovery of general biophysical features to improve enzyme performance. Nevertheless, many questions about enzyme function remain to be answered and new research to

enhance our understanding of these vital proteins remain an exciting area for many generations to come.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape, *Molecular Biology and Evolution 31*, 1581-1592.

[2] Stiffler, Michael A., Hekstra, Doeke R., and Ranganathan, R. (2015) Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase, *Cell 160*, 882-892.

[3] Bacik, J.-P., Klesmith, J. R., Whitehead, T. A., Jarboe, L. R., Unkefer, C. J., Mark, B. L., and Michalczyk, R. (2015) Producing Glucose 6-Phosphate from Cellulosic Biomass: STRUCTURAL INSIGHTS INTO LEVOGLUCOSAN BIOCONVERSION, *Journal of Biological Chemistry 290*, 26638-26648.

[4] Jarboe, L. R., Wen, Z. Y., Choi, D. W., and Brown, R. C. (2011) Hybrid thermochemical processing: fermentation of pyrolysis-derived bio-oil, *Appl Microbiol Biot 91*, 1519-1523.

[5] Bennett, N. M., Helle, S. S., and Duff, S. J. B. (2009) Extraction and hydrolysis of levoglucosan from pyrolysis oil, *Bioresource Technology 100*, 6059-6063.

[6] Davis, J. H., Rubin, A. J., and Sauer, R. T. (2011) Design, construction and characterization of a set of insulated bacterial promoters, *Nucleic Acids Research 39*, 1131-1141.

[7] Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E., Tomek, K. J., and Whitehead, T. A. (2014) The interrelationship between promoter strength, gene expression, and growth rate, *PloS one 9*, e109105.

[8] Araya, C. L., and Fowler, D. M. (2011) Deep mutational scanning: assessing protein function on a massive scale, *Trends in biotechnology 29*, 435-442.

[9] Fowler, D. M., and Fields, S. (2014) Deep mutational scanning: a new style of protein science, *Nature methods 11*, 801-807.

[10] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N., and Whitehead, T. A. (2015) High-resolution sequence-function mapping of full-length proteins, *PloS one 10*, e0118193.

[11] Klesmith, J. R., Bacik, J.-P., Michalczyk, R., and Whitehead, T. A. (2015) Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli, *ACS Synthetic Biology 4*, 1235-1243.

[12] Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M., and Wittrup, K. D. (2006) Isolating and engineering human antibodies using yeast surface display, *Nat. Protocols 1*, 755-768.

[13] Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein, *Nat Biotech 17*, 691-695.

[14] Boock, J. T., King, B. C., Taw, M. N., Conrado, R. J., Siu, K.-H., Stark, J. C., Walker, L. P., Gibson, D. M., and DeLisa, M. P. (2015) Repurposing a Bacterial Quality Control Mechanism to Enhance Enzyme Production in Living Cells, *J Mol Biol 427*, 1451-1463.

[15] Goldenzweig, A., Goldsmith, M., Hill, Shannon E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, Raquel L., Aharoni, A., Silman, I., Sussman, Joel L., Tawfik, Dan S., and Fleishman, Sarel J. (2016) Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability, *Molecular Cell 63*, 337-346.

[16] Fleishman, S. J., Whitehead, T. A., Strauch, E.-M., Corn, J. E., Qin, S., Zhou, H.-X., Mitchell, J. C., Demerdash, O. N. A., Takeda-Shitaka, M., Terashi, G., Moal, I. H., Li, X., Bates, P. A., Zacharias, M., Park, H., Ko, J.-s., Lee, H., Seok, C., Bourquard, T., Bernauer, J., Poupon, A., Azé, J., Soner, S., Ovalı, Ş. K., Ozbek, P., Tal, N. B., Haliloglu, T., Hwang, H., Vreven, T., Pierce, B. G., Weng, Z., Pérez-Cano, L., Pons, C., Fernández-Recio, J., Jiang, F., Yang, F., Gong, X., Cao, L., Xu, X., Liu, B., Wang, P., Li, C., Wang, C., Robert, C. H., Guharoy, M., Liu, S., Huang, Y., Li, L., Guo, D., Chen, Y., Xiao, Y., London, N., Itzhaki, Z., Schueler-Furman, O., Inbar, Y., Potapov, V., Cohen, M., Schreiber, G., Tsuchiya, Y., Kanamori, E., Standley, D. M., Nakamura, H., Kinoshita, K., Driggers, C. M., Hall, R. G., Morgan, J. L., Hsu, V. L., Zhan, J., Yang, Y., Zhou, Y., Kastritis, P. L., Bonvin, A. M. J. J., Zhang, W., Camacho, C. J., Kilambi, K. P., Sircar, A., Gray, J. J., Ohue, M., Uchikoga, N., Matsuzaki, Y., Ishida, T., Akiyama, Y., Khashan, R., Bush, S., Fouches, D., Tropsha, A., Esquivel-Rodríguez, J., Kihara, D., Stranges, P. B., Jacak, R., Kuhlman, B., Huang, S.-Y., Zou, X., Wodak, S. J., Janin, J., and Baker, D. (2011) Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology, *Journal of Molecular Biology 414*, 289-302.