ABSTRACT

MCARTHUR, JOHN. *In Vitro* Ultra-High Throughput Screens for Directed Evolution. (Under the direction of Dr. Gavin J. Williams).

Biocatalyst development is an important area of interest in synthetic biology. The complexity of enzyme structure strongly impedes our ability to design enzymes de novo, forcing biocatalyst development to build from natural, functional enzymes through enzyme engineering. The evolvability of these natural enzymes can be exploited to improve or alter enzyme function without any understanding of structure-function relationships through a process called directed evolution. Directed evolution mimics natural evolution; iterative rounds of mutagenesis and screening or selection yield mutants with the desired activity.

New screening formats that minimize sample size and increase analysis rate (throughput) greatly reduce the cost and time burden of directed evolution. However, to date each of these formats lack broad utility. Here, a potentially broad utility ultra-high throughput screening platform was developed that uses fluorescence activated cell sorting (FACS), *in vitro* compartmentalization (IVC) in emulsion microdroplets, a previously unexploited fluorogenic natural product glycosylation reaction, and an extensive, modular coupled enzyme biosensor system enabling detection of a variety of metabolites and cofactors.

In Chapter 2, an IVC-FACS screen was developed based on a natural reaction catalyzed by a natural product glycosyltransferase UGT72B1: the selective glucosylation of esculetin to yield esculin. Prior to this advance, screening platforms based on IVC-FACS suffered from poor protein expression yields. Here, this problem is resolved using a strategy of emulsion compartmentalization of *E. coli* cells followed by cell lysis. Robust biocatalyst expression permits exploitation of otherwise sluggish fluorogenic enzymes, such as UGT72B1. The

crucial genotype-phenotype link necessary for directed evolution was verified by enriching model populations of UGT72B1 and an inactive mutant. Confirmation of enrichment provided strong evidence for a fully functional directed evolution screen.

In Chapter 3, the UGT72B1 catalyzed fluorogenic reaction was coupled to several upstream enzymatic reactions to produce coupled enzyme biosensors (CEBs) capable of detecting several natural biomolecules. Endpoint fluorescence correlated with analyte concentration in both the IVC-FACS lysate format as well as in microtiter plates. This CEB system is highly modular and engineerable, providing the potential for customized detection of nearly any metabolite of interest.

In Chapter 4, mathematical models were derived which can be used to predict optimal mutation rates in error-prone polymerase chain reaction (epPCR) for directed evolution. Modeling of a process called neutral drift, which can be performed with our CEB-IVC-FACS screens or other ultra-high throughput screens, including a previously published universal FACS screen for protein folding and stability, improves the guidelines for epPCR library mutation rates for neutral drift. The results of these models provide strong support for neutral drift as an advantageous preliminary step in all directed evolution experiments and suggest improved library quality compared to other prominent library quality enhancing techniques.

In Chapter 5 a mutant library of UGT72B1 was generated by epPCR and was subjected to neutral drift using the IVC-FACS screen from Chapter 2. Subsequent analysis of the enriched library by FACS failed to indicate the presence of a mutant displaying activity with non-native donor substrates. Further efforts to identify mutants with improved catalytic properties from the enriched library by screening mutants in 96-well microtiter plates is described.

In Chapter 6, the potential for a CEB-IVC-FACS screen to be used for engineering polyketide synthase (PKS) activities is discussed. The work of Dr. Irina Koryakina to better understand PKS specificities toward non-natural substrates has helped place the Williams lab at the forefront of PKS engineering, but the lack of a suitable screen hinders efforts to apply this knowledge toward altering PKS activities.

Progress in the field of synthetic biology is highly dependent on biocatalyst engineering and is therefore hindered or enhanced according to advances in directed evolution. The CEB-IVC-FACS screening format has great potential as a broadly applicable directed evolution tool, and the mathematical models presented in this work enhance our understanding of strategic choices in directed evolution library preparation.

In Vitro Ultra-High Throughput Screens for Directed Evolution

by
John McArthur

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Chemistry

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
Dr. Gavin J. Williams Committee Chair	Dr. Christian Melander
_____	_____
Dr. Reza Ghiladi	Dr. Edmond Bowden

BIOGRAPHY

The author, John Barron McArthur, Jr., was born on April 29th 1987 in Columbia, South Carolina. John has a younger brother, Will, and loving parents, John and Barbara. John began studying science in high school at the South Carolina Governor's School for Science and Mathematics, and after graduating John enrolled at the University of South Carolina where he earned his B.S. in Chemistry in 2009. While at USC, John performed research under the guidance of Prof. Qian Wang. The focus of this research was the synthesis of water-soluble, sugar-sensitive polymers for applications in real-time glucose monitoring. After graduating, John joined the Department of Chemistry at NCSU and began conducting research under the supervision of Prof. Gavin Williams. During his doctorate, John was involved in a successful collaboration with Irina Koryakina that resulted in three manuscripts. After the completion of his PhD, John will pursue his postdoctoral studies in the lab of Prof. Xi Chen at the University of California, Davis. The objective of this research will be engineering enzymes as biocatalysts for complex therapeutic oligosaccharides.

ACKNOWLEDGMENTS

I would first like to thank my research advisor, mentor, and friend, Prof. Gavin Williams. Thank you so much for showing me the powerful potential of biocatalysis and directed evolution. Thank you for providing me with the chance to develop my own sometimes absurd ideas. Thank you for your tolerance and advice when progress was slow, and for your enthusiasm when progress was less slow.

I would also like to thank my colleagues in the Williams Lab. Thank you for tolerating my quirks, patiently listening to my ideas, and diligently advancing progress in our field through your own hard work and innovation. Dr. Irina Koryakina especially deserves my thanks. You have been inspirational and supportive beyond description, and I can't imagine the last five years without you. I would like to thank the undergraduate researchers I trained and mentored, especially Hemant Desai for his hard work and thought provoking ideas. Thank you Edward Kalkreuter for continuing the work on our fascinating project; I know the work is left in good hands.

I would like to thank my family for your unconditional love and support. I could not have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Mutagenesis and Screening in Synthetic Biology

1.1. Introduction to Synthetic Biology and Directed Evolution

Living organisms have been intentionally altered to better serve our needs for thousands of years. Selective breeding programs began before written history and have yielded modern corn, dairy cows, and highly ethanol tolerant yeast strains.[1,2] The rise of molecular biology and recombinant DNA technology provide the means for reengineering living organisms for exploitation in ways unachievable by traditional breeding.[3] This new field, in which molecular biology tools are used to create or engineer biology rather than study it, is known as synthetic biology. Applications of synthetic biology include therapeutic proteins[4], biocatalysts for the chemical and pharmaceutical industries[5], and genetically modified crops.[6] In 2012, the total revenues from these industries were at least $350 billion in the US, approximately 2.5% of GDP, and these domestic revenues are growing at rates exceeding 10% per annum.[7]

Biocatalyst development is a rapidly expanding area in synthetic biology with potential to resolve a number of problems faced by the chemical and pharmaceutical industries.[8] Photosynthetic biofuels, for example, are liquid fuels compatible with our existing oil infrastructure and automobiles that can be burned with net zero carbon emissions.[9] Chiral compounds, particularly fine chemicals and natural products, can often be produced biocatalytically at lower cost, higher yield, and higher stereopurity.[10] Bulk chemicals can be prepared biocatalytically from renewable starting materials without the need for toxic metals or solvents. For example, 1,3-propanediol, mainly used as a building block for synthetic

copolyesters, is produced by Dupont from corn syrup using a genetically modified strain of *E. coli*.[11]

In many cases, enzymes catalyzing the desired reactions must be engineered to optimize the biocatalytic system. Enzymes can be engineered stability under non-natural conditions[12], to alleviate pathway bottlenecks[13], or to eliminate feedback inhibition.[14] Fully realizing the potential of biocatalysis will depend on our ability to quickly reengineer natural enzymes to catalyze entirely unnatural reactions.[8] Due to our near complete ignorance of how enzyme structure encodes function, rationally designing a biocatalyst to perform a reaction that does not naturally occur is particularly difficult.[15]

Directed evolution is a powerful enzyme engineering strategy that can require no knowledge of structure-function relationships to be successful.[16] Inspired by the power of natural evolution as an efficient optimization algorithm, directed evolution uses random or targeted mutagenesis and screening or selection to find mutant enzymes with improved or altered properties (Fig. 1). Importantly, the selective pressure should be defined according to the desired changes by carefully establishing proper assay conditions.

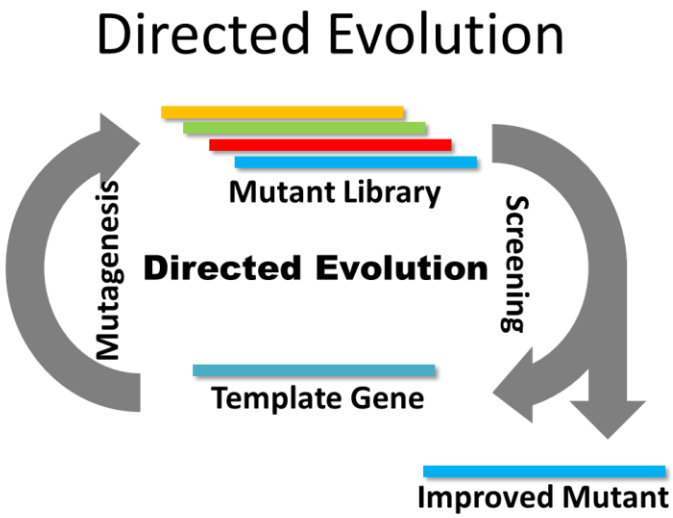

Figure 1. Scheme of directed evolution. A mutant library is created from a template gene, and the subsequent application of a screen or selection provides an improved mutant. If the engineering goal has not yet been reached, the improved mutant is the template for the next round of mutagenesis.

Directed evolution has resulted in many notable successes and is an established technology within industrial biocatalysis. Prominent examples of laboratory-evolved enzymes include an efficient propane monooxygenase from a fatty acid P450 hydroxylase[17], T7 RNA polymerases with non-natural promoter sequence recognition[18], and a stereoselective transaminase capable of accepting ketones with two bulky substituents for the preparation of diabetes drug Sitagliptin.[19]

The field of directed evolution has advanced rapidly from both conceptual and technological breakthroughs.[20] The focus of my doctoral work has been the development of both conceptual and technological tools to improve the scope and efficiency of directed enzyme evolution.

1.2 Choosing a Template for Directed Evolution

The first step in a directed evolution cycle is the choice of a template gene for mutagenesis. In the first round of directed evolution, an ideal template choice would be an enzyme that expresses well in the desired host (most commonly *E. coli*), has detectable levels of the desired activity, and has a substrate-bound crystal structure available in the Protein Data Bank (PDB).[21] Typically, the template gene is codon optimized for improved heterologous expression[22], although this is not always necessary.

In subsequent rounds of directed evolution, one or more genes from earlier rounds of directed evolution may be used as the template. When multiple improved variants have been found, often the most efficient way to further improve the enzyme is to recombine these variants through site-directed mutagenesis, gene shuffling, or multisite saturation mutagenesis.[23]

1.3 Mutagenesis in Directed Evolution

The mutant library generation step in each round of directed evolution relies on robust technologies.[24] Commercial kits are now available for multi-site saturation mutagenesis, in which specific codons are simultaneously randomized in a one-pot reaction.[25] Commercial error-prone polymerase chain reaction (epPCR) kits use engineered polymerases with lower mutational bias toward transition mutations and other improved mutagenic properties.[26] A number of techniques now exist for chimeragenesis and gene shuffling.[27,28] Reductions in the cost of gene synthesis have made customized library synthesis feasible.[29] Clever combinations of primers generated from restricted nucleotide alphabets provide greatly reduced redundancy

for multisite saturation mutagenesis.[30] Additionally, cloning procedures have been developed to allow libraries to be prepared without the artifacts of traditional ligation based cloning.[31]

Perhaps more important are the advancements in resources for sequence and structural analysis that allow targeted mutagenesis approaches to enjoy high success rates. These include the 450% growth of the number of protein crystal structures deposited in the PDB over the last decade[8], improved bioinformatics software for multiple sequence alignments, ProSAR analysis of sequence and activity data[32], and SCHEMA sequence and structure analysis for contiguous and noncontiguous chimeragenesis.[33] Additionally, natural enzymes are being more thoroughly analyzed for properties potentially valuable to biotechnology, such as temperature and solvent tolerance and substrate specificity with unnatural substrates. This information can guide choices of sequence and structural alignments.

The conceptual advancements in mutagenesis for directed evolution have been significant. In particular, many in the field no longer consider simple error-prone PCR to be the optimal approach, despite the ability to find beneficial mutations which can not currently be predicted rationally or computationally.[34] Instead, "smaller and smarter" libraries are preferred.[20]

There is considerable interest in mathematical models that predict optimal mutagenesis and screening strategy.[35] Chapter 4 expands on prior theoretical work to develop improved mathematical models for directed evolution.[34-36] The work in Chapter 4 is focused on modeling libraries generated by epPCR, but further understanding of the "hidden variables" of directed evolution, evolvability and mutational robustness, will allow the models to be adapted to other methods of mutagenesis.

1.4 Screens and Selections in Directed Evolution

While mutagenesis technologies are essentially universal, assays used as screens and selections in directed evolution must be customized and validated for each enzymatic activity of interest.[37] Unfortunately, most enzyme reactions do not lead to a detectable change in absorbance or fluorescence.[38] In addition, improvements to many enzyme activities do not provide any advantage to host cell reproduction or a detectable phenotype change, a problem even more significant when attempt to engineer enzyme activities toward non-natural or non-native substrates. The development of one-size-fits-all screens and selections continues to lag behind the development of similarly broadly applicable mutagenesis methods due to the inherent difficulty in detecting inconspicuous molecules at high throughput and low cost.

The terms 'screen' and 'selection' are easily confused, as both are tools used to apply 'selective' evolutionary pressures upon a mutant population. The best distinguishing feature between 'screens' and 'selections' is that when screening, mutants can be kept or discarded after phenotype testing, while with selections, mutants with activity below a pre-set threshold are discarded automatically.

Interestingly, some misnomers exist within the literature, most notably the ubiquitous labeling of Fluorescence Activated Cell Sorting (FACS) assays as screens.[39] Whereas FACS works through fluorescence detection of single cells, particles, or droplets, and all other spectrophotometric assays to date are termed as screens, FACS is definitely a *selection* technology.[40] The fluorescence threshold used to judge whether a mutant enzyme in or on a cell, droplet, or particle should be collected or discarded must be set before a sample's fluorescence is measured. Subsequently adjusting the threshold only affects the portion of the

sample waiting to be assayed. Additionally, although the fluorescence readout of each mutant is measured individually, there is no link between an individual data point and a collected 'hit'; all one can say about the performance of any individual 'hit' is that it performed well enough to cross the pre-set threshold.

Ultimately, the argument over the labels 'screen' or 'selection' for FACS is a predominantly semantic one, as each label fails to capture the unique properties of FACS as a tool in directed evolution. Defined as a selection, FACS is unparalleled in terms of its detection of fluorescence; most selections are highly specific to a single activity, while fluorescence is a somewhat more general molecular property. Defined as a screen, FACS reaches unparalleled throughput and low sample cost; up to 10^8 mutants can be assayed in a single round of directed evolution using smaller quantities of substrates and reagents than would be used in a single well of a 96-well microtiter plate. Technically, FACS assays are selections, but due to the inadequacy of this label and for the sake of consistency with decades of prior work, I will refer to FACS assays as 'screens' and the use of these assays as 'screening' for the duration of this document.

The most common screening format in directed evolution is microtiter plate screening of *E. coli* lysates.[37] There are many benefits to this format; assays are conducted *in vitro*, soluble reaction products from separate mutants are unable to mix, and a wide variety of detection systems are commercially available including fluorescence and absorbance monitoring plate readers and specialized, automated LC-MS technology. However, ultimately the cost and time burden of microtiter plate screening, as well as the modest throughput of typically less than 10,000 mutants per round, limits the strength of microtiter plate screening.

The development of a screening or selection technology with similar benefits but higher throughput and lower cost has been a long-standing goal.

The ultra-high throughput and low cost of FACS make it a top contender as a tool for next generation screening platforms. However, how can the production of inconspicuous molecules be tied to the fluorescence properties of an *E. coli* cell? One suggested solution is genetic biosensors, which exploit natural mechanisms (transcription factors, riboswitches, etc.) that alter expression of an easily detectable gene product, such as Green Fluorescent Protein, dependent on the concentration of a metabolite of interest.[38] Unfortunately, when considering the limitations of conducting assays *in vivo*, it becomes clear that any truly robust screening or selection platform must test the mutant enzymes *in vitro.*

In vivo assays are limited by a number of factors (Fig. 2). The substrates of the desired enzymatic activity must be available within the cell. This can be problematic for non-natural substrates of anabolic enzymes, since most metabolic building block molecules are charged, cell impermeable, and often impossible to produce *in vivo*. Heterologous overexpression of the mutant library as well as the desired level of activity are often toxic to the host cell. Endogenous enzymes and metabolites prevent many assays viable with purified components from being successful *in vivo*. Finally, many product molecules are cell permeable and diffuse from the cell into the media, breaking the required link between genotype and phenotype. All of these problems limit the broad utility of genetic biosensors.

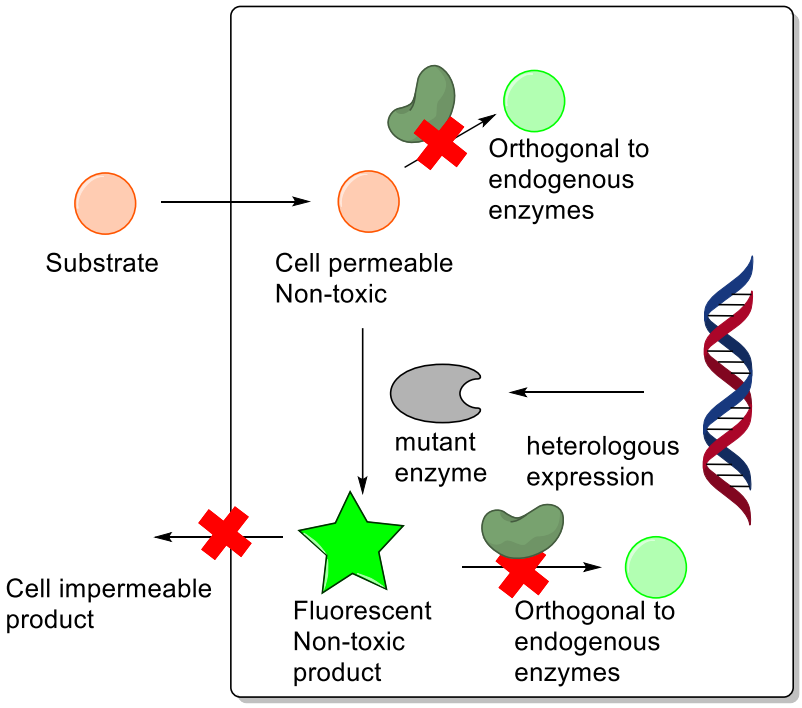

Figure 2. Cells are poor compartments for conducting assays. For a reaction to be useful for *in vivo* FACS screening, the substrates must be non-toxic, cell permeable, and orthogonal to endogenous enzymes. Additionally, the mutant enzyme must be non-toxic and should be orthogonal to endogenous metabolites (not shown). The desired product must be fluorescent, cell impermeable, non-toxic, and orthogonal to endogenous enzymes.

Fortunately, two robust technologies provide *in vitro* reaction conditions compatible with FACS sorting. The first is yeast surface display (Fig. 3).[41] This technology relies on genetically fusing the enzyme of interest to a small yeast surface protein. A substrate of the enzyme of interest is transferred to a separate surface protein which associates on the cell surface with the enzyme-fused surface protein. The mutant enzyme is challenged to catalyze the desired reaction on the immobilized substrate, and the change in chemical structure upon successful catalysis can be detected via bioorthogonal labeling reactions or fluorophore-tagged antibodies. There are considerable limits on which enzymatic properties can be evolved, it is

difficult to apply selective pressures for k_{cat} or K_M of the immobilized substrate. Additionally, enzymes often do not accept substrates with large structural changes, so investing in the development of a yeast surface display screen can be quite risky. The enzyme of interest must be active as a monomer to be successfully screened via yeast surface display. Like many FACS screens, yeast surface display is not amenable to medium or high throughput formats which allow more accurate measurements of activity for individual mutants after initial enrichment, necessitating the development of a separate secondary screen. Even for enzymes which can be screened by yeast surface display, creating the genetic constructs, synthesizing the substrate analogues, and validating the screen requires substantial investments in time and resources.

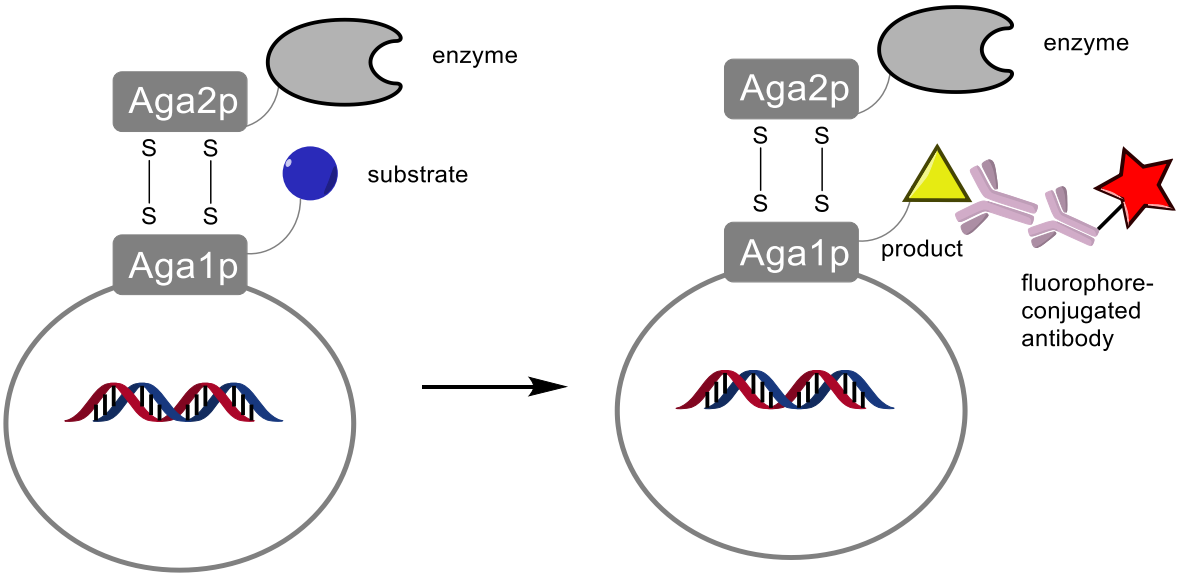

Figure 3. Yeast surface display is a general *in vitro* screening format compatible with FACS. Intrinsic problems, such as the limits of modified enzyme substrates and the difficulty and expense involved in setting up this complex system, have limited applications of this technology.

The other technology that provides *in vitro* reaction conditions compatible with FACS is called *in vitro* compartmentalization (IVC, Fig. 4).[42,43] IVC involves the formation of emulsion microdroplets from an oil layer and an aqueous layer. The aqueous layer contains the DNA library, substrates for the desired reaction, and components of a commercial *in vitro* expression kit. Since the emulsion droplets are generated from this mixture, the components are equally dispersed amongst the droplets and cell-impermeable charged molecules can easily be included. Some enzymes that can modify their own DNA such as DNA polymerases, methyltransferases, and restriction endonucleases, have been directly selected for with IVC.[44-46] However, these approaches have little potential for screening other enzymatic activities. Reemulsification of a water-in-oil (w/o) emulsion with a second aqueous phase results in a water-in-oil-in-water (w/o/w) emulsion compatible with FACS.[43] Recovery of 'hit' DNA is achieved through PCR amplification of miniscule quantities of DNA extracted and purified from pooled droplets.

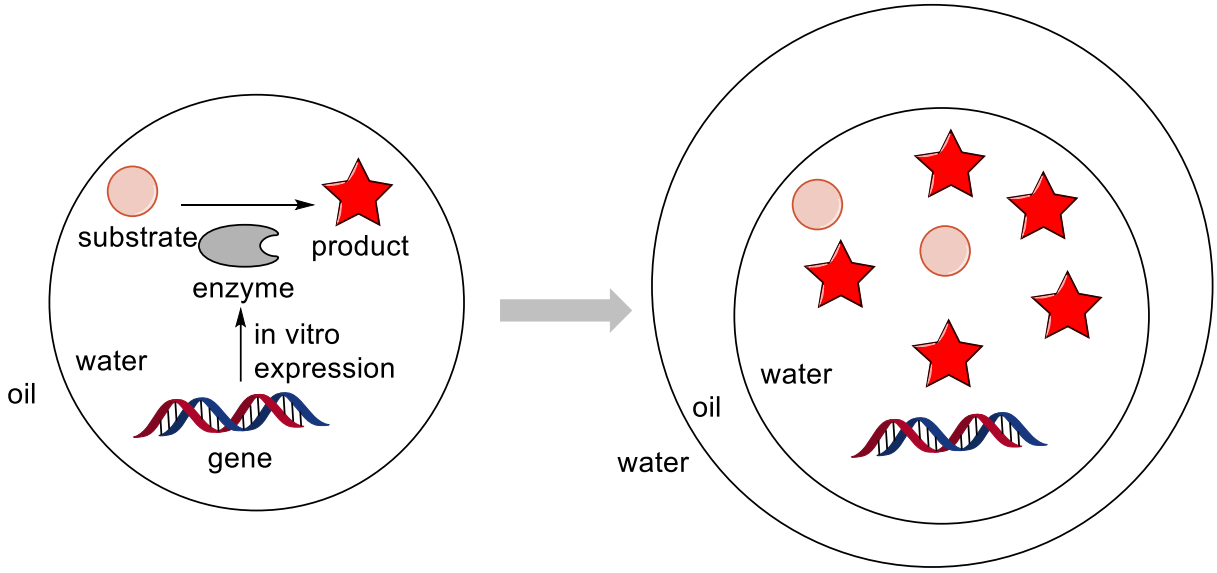

Figure 4. IVC provides FACS compatible compartments for *in vitro* directed evolution. A droplet within the w/o emulsion (left) contains an *in vitro* expression kit, a molecule of mutant DNA, and the substrate of the fluorogenic reaction. The most important features of the w/o/w emulsion are the fluorescent product detected by FACS and the mutant DNA which can be recovered and amplified.

The combination of IVC and FACS is a potentially powerful screening format, but due to technical challenges there have been examples of successful directed evolution with this technology. Two challenges in particular must be addressed: poor *in vitro* protein expression[42] and few compatible fluorescent assays. Chapters 2 and 3 describe the development a more robust bulk emulsion IVC-FACS screening technology useful for directed evolution of a wide range of enzymatic reactions. The screen was used to detect the activity of a number of different enzymatic activities through coupled enzyme biosensors (CEBs), and it was also used to improve the heterologous expression of a key component of these biosensors, glycosyltransferase UGT72B1, via neutral drift and secondary screening. Importantly these CEBs are suitable for microtiter plate screening as well as bulk emulsion IVC-FACS. The

solutions developed herein to address the major problems of IVC-FACS are perfectly compatible with microfluidic technology[47], providing a route to further improvements in sensitivity. Databases of enzyme catalyzed reactions enable the rational design of customized CEBs for potentially any IVC-FACS compatible metabolites and cofactors of interest.

CHAPTER 2

Development of an IVC-FACS Screen

2.1. Introduction

One of the primary reasons the potential of IVC-FACS as a versatile ultra-high throughput screening platform for directed evolution has not yet been realized is low sensitivity caused by poor *in vitro* protein expression. When the mutant library is expressed from DNA in droplets by *in vitro* expression kits, the DNA must be diluted so that few droplets contain more than one molecule of DNA. Expression yields under these conditions have been estimated to be as low as 10 to 100 *molecules* of mutant enzyme *per* droplet.[42] Additionally, many common emulsion oils and surfactants inhibit *in vitro* protein expression.[48] Only one enzymatic activity has been reengineered using the basic IVC-FACS strategy shown in Fig. 4. Mastrobattista et al. screened a mutant library of Ebg, an *E. coli* protein of unknown function but known to be evolvable for β-galactosidase activity[49], for activity against fluorescein di(β-galactoside).[43]

Others have circumvented the poor *in vitro* expression problem by manipulating droplets within microfluidic chips.[50] In one such study, horseradish peroxidase mutants were displayed on the surface of yeast cells and compartmentalized with a fluorogenic substrate.[51] 10^4 molecules of enzyme were displayed per cell, and compartmentalization prevented the need to covalently attach the fluorogenic substrate to a cell surface protein (Fig. 3). This device used in this directed evolution campaign used a custom-built fluorescence detection system that signals microelectrodes imbedded within the device for dielectrophoretic sorting. In another study, microfluidic droplets are subjected to on-chip PCR, resulting in a 30,000-fold

increased DNA copy number per droplet.[52] After amplification, these droplets were merged one to one with droplets containing *in vitro* expression components. Unfortunately, these sophisticated approaches are costly and technically challenging to adopt.

Still other examples of IVC-FACS use *E. coli* or yeast cells compartmentalized within droplets.[53] These are *in vivo* FACS assays that use microfluidic droplets to provide a genotype-phenotype link when the cell membrane is not sufficient for product entrapment. Inspired by this, we proposed to address the poor *in vitro* expression levels of IVC-FACS by generating single-cell lysates in emulsion droplets (Fig. 5). By compartmentalizing and lysing whole *E. coli* cells within droplets, expression levels as high as 10^6 molecules of protein per droplet could theoretically be achieved. Unlike preceding cell in droplet approaches, this format would allow *in vitro* screening analogous to screening lysates in microtiter plates. Compared with *in vivo* FACS screening, this approach would achieve the same protein expression levels, but cell membrane permeability of the substrate would not be a restricting factor. Furthermore, competition with intracellular metabolites would not be problematic, due to vast dilution of endogenous metabolites upon cell lysis within the emulsion droplet. During the course of this study, proof of concept of this approach was reported by Kintses et al., albeit using cells compartmentalized on microfluidic chips.[54] Interestingly, the 10^3 to 10^5 -fold increase in protein expression enabled by cell lysis instead of *in vitro* expression was not exploited or mentioned; rather the purpose for this innovation was the large (up to 10^3-fold) increase in molecules of DNA per droplet.

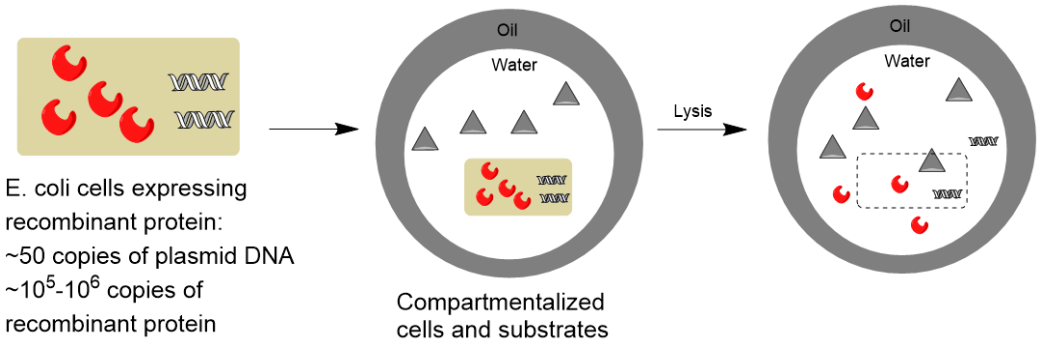

Figure 5. Generating cell lysates in emulsion microdroplets provides large improvements in enzyme expression compared with *in vitro* expression kits.

The improved protein expression levels potentially offered by this strategy allowed us to consider fluorogenic enzymatic reactions that were otherwise far too slow to be previously screened by IVC-FACS and which can not be screened by *in vivo* FACS. We proposed a novel fluorogenic reaction for the cell lysate IVC-FACS strategy based on a natural product glycosyltransferase. Regioselective glycosylation of the coumarin natural product esculetin by *Arabidopsis thaliana* glycosyltransferase UGT72B1 with UDP-glucose produces the natural product glycoside esculin (Fig. 6).[55] UGT72B1 is poorly expressed in *E. coli* and its reaction is 2000-fold slower than the IVC-FACS evolved β-galactosidase reaction.[43] The reaction is only moderately fluorogenic, but it was hypothesized that upon reemulsification and dilution, esculetin would leak from the internal droplets while esculin would be retained due to its higher polarity, thus improving the change in fluorescence for droplets containing an active glycosyltransferase. Notably, this reaction provides the first opportunity to engineer a natural product glycosyltransferase through directed evolution with an ultra-high throughput screen.

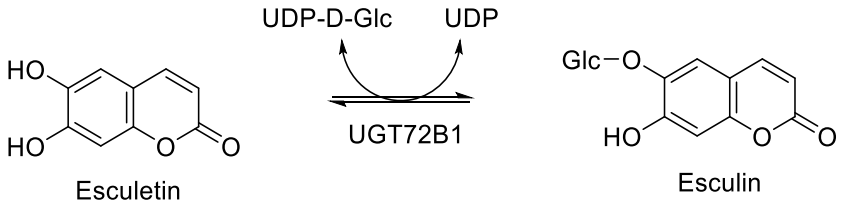

Figure 6. Scheme of UGT72B1 catalyzed glycosylation of esculetin. Esculin, the natural product glycoside, is more fluorescent than esculetin.

2.2. Results and Discussion

2.2.1 IVC Cell Lysis Validation

Cell lysis is a fundamental step in microtiter plate screening of *E. coli* lysates, and therefore several efficient methods have been developed. One common strategy for cell lysis is the use of a cell lysis reagent, as was used by Kintses et al.[54] However, for the IVC format it is critical that lysis begins only after compartmentalization, otherwise the essential genotype-phenotype link could be broken. The addition of polar reagents into emulsion microdroplets has been achieved,[56] but uniform delivery into primary emulsions is difficult without microfluidic manipulation due to the high viscosity of the primary emulsions. Another strategy for cell lysis is freezing and thawing, often in the presence of lysozyme.[23] Since emulsions have been formulated which are stable to the large temperature changes that take place during PCR,[52] we reasoned that an emulsion might also be formulated for stability through freezing and thawing.

E. coli cells were compartmentalized in w/o emulsions with propidium iodide (PI), a membrane impermeable DNA intercalating dye commonly used to differentiate necrotic, apoptotic, and normal cells by fluorescence microscopy or flow cytometry.[57] Some of the

emulsion was subjected to rapid freezing in a dry ice ethanol bath (-72 °C) followed by thawing at room temperature. The untreated emulsion and that portion subjected to freeze-thaw were then analyzed by fluorescence microscopy. Fortunately, the droplets subjected to the freeze-thaw process retained their structural integrity and showed significant enhancement in PI fluorescence (Fig. 7), providing initial evidence that a simple freeze-thaw procedure could be used to liberate expressed protein from within compartmentalized bacterial cells.

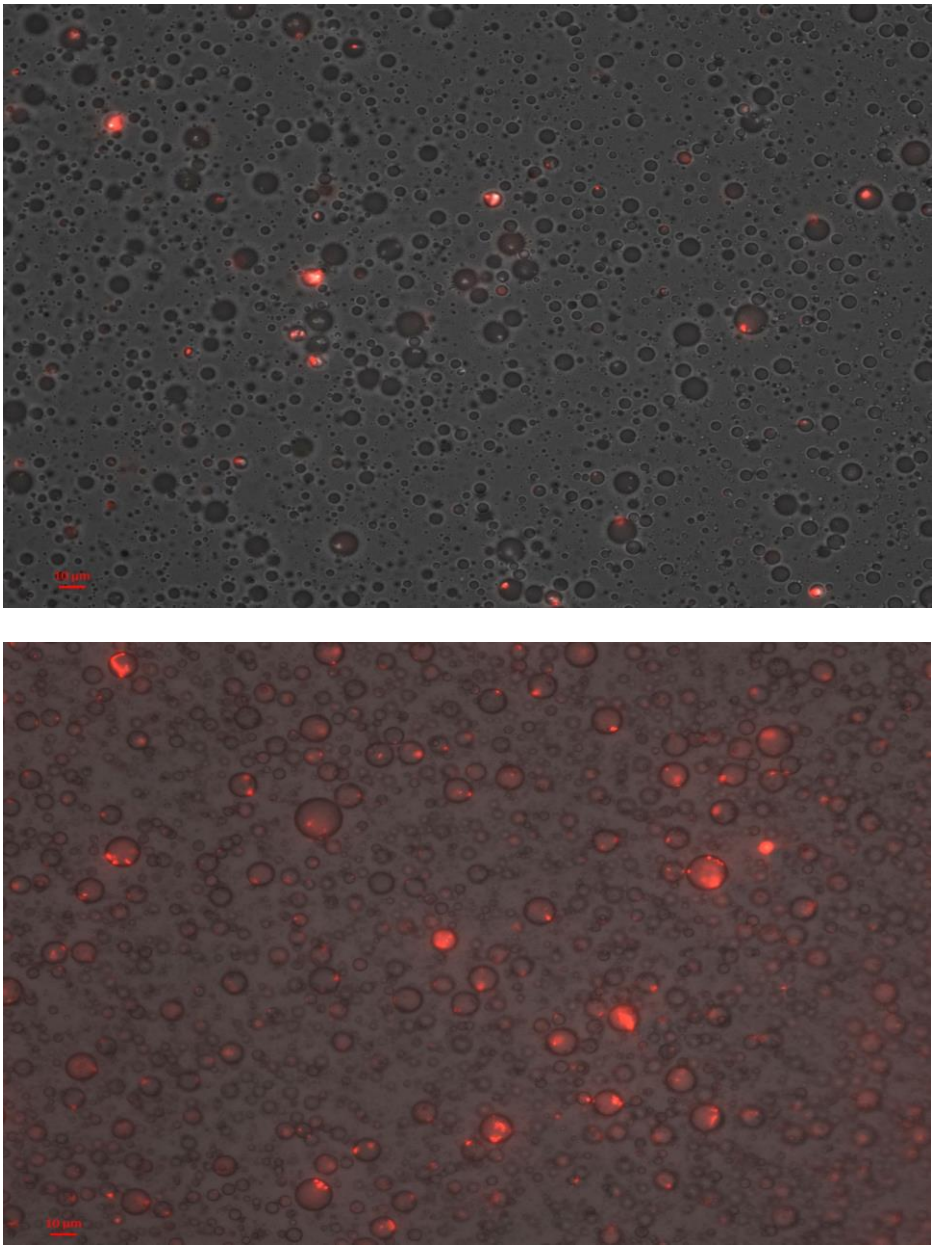

Figure 7. Fluorescence microscopy image of primary emulsions containing *E. coli* cells and PI without (top) and with (bottom) freeze-thaw treatment.

2.2.2 UGT72B1 Fluorogenic Reaction Development

Enzymes which transfer charged moieties, such as sialyltransferases and kinases,[58,59] or those which can release a charged moiety through hydrolysis, such as esterases,[60] can be readily screened for some properties by FACS *in vivo* because they render fluorescent probes impermeable upon successful catalysis. In such cases, the fluorescent product is trapped within the cell which expressed the active enzymes, thus ensuring the genotype-phenotype link. We reasoned that the oil layer separating the internal and external aqueous phases in a double emulsion would be sufficiently non-polar to enable the entrapment of uncharged natural product glycosides.

7-hydroxycoumarins are natural product fluorophores which are commonly isolated as glycosides.[55] However, glycosylation at the 7-OH quenches fluorescence.[23] Therefore we sought a fluorescent coumarin natural product glycosylated at a position other than the 7-OH. Esculin is a fluorescent coumarin glycoside commonly used in esculin-bile agar plates for the separation and identification of bacteria of the genus *Enterococcus*.[61,62] Members of this genus are able to grow in the presence of 4% bile and hydrolyze the glucose from esculin, producing esculetin, resulting in a decrease in coumarin fluorescence. This suggested that glycosylation of esculetin to form esculin would be fluorogenic, although to our knowledge this fluorogenic glycosylation reaction had not yet been exploited as an enzyme assay, despite multiple previous efforts to engineer UGT72B1.[63,64]

Compartmentalization of esculetin and esculin in w/o/w emulsions and subsequent analysis by FACS showed that esculin was faithfully retained within the droplets, while esculetin was not. At pH 8 and with the emulsions diluted 10-fold in PBS prior to sorting,

droplets made with 1 mM esculin were over 60-fold more fluorescent that those made with esculetin (Fig. 8).

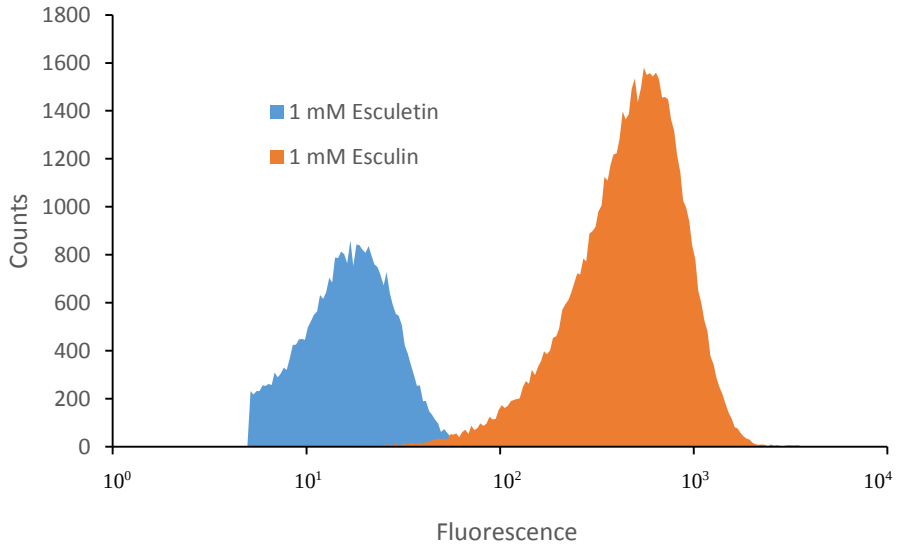

Figure 8. Histogram overlay of FACS data from droplets constructed with 1mM esculetin or 1 mM esculin.

Arabidopsis thaliana UGT72B1 was selected as a potential esculetin glycosyltransferase to test based on the work published by the Bowles group, which showed that UGT72B1 produces the desired esculin regioisomer exclusively.[55] A crystal structure with bound substrate analogues is available for UGT72B1, and several promiscuous activities have been characterized.[63] The UGT72B1 gene was cloned into pET-28a, and UGT72B1 was expressed and purified to at least 90% homogeneity as a C-terminally His-tagged fusion by affinity chromatography on nickel NTA resin. The activity of the purified enzyme was tested using UDP-glucose and esculetin as substrates. Gratifyingly, the reaction in a microtiter plate

was fluorogenic (Fig. 9). Low levels of activity toward other UDP-sugars would suggest that UGT72B1 is highly evolvable toward these substrates and would present an opportunity to learn about natural product glycosyltransferase donor substrate evolvability through directed evolution. UGT72B1 was therefore tested with other UDP-sugars: UDP-galactose, UDP-N-acetylglucosamine, UDP-N-acetylgalactosamine, and UDP-glucuronic acid. No product was detected by HPLC (data not shown), and no fluorescence increase was observed in microtiter plates for any of these substrates (Fig. 9).

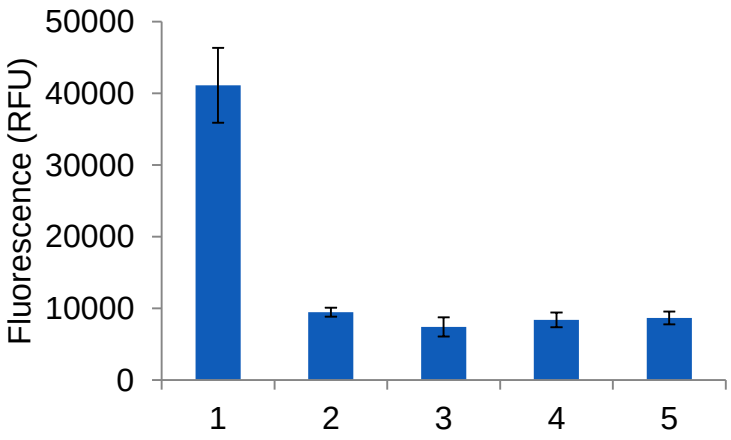

Figure 9. Microplate analysis of UGT72B1 with UDP-sugars: glucose (1), galactose (2), glucuronic acid (3), N-acetylglucosamine (4), and N-acetylgalactosamine (5). Error bars represent the standard deviation from the mean (n = 3).

With a fluorogenic glyosyltransferase catalyzed reaction established, the enzymatic activity needed to be reproduced and detected in microdroplet compartmentalized lysates. The gene encoding UGT72B1 was subcloned into the first multiple cloning site of pETduet after cloning a gene encoding the superfolder Green Fluorescent Protein (sfGFP) into the second

multiple cloning site of pETduet. It was expected that coexpression of sfGFP within the compartmentalized cell would allow the FACS sorter to trigger an event only when droplets that contain sfGFP (and thus a lysate) passed through the detector. Additionally, overlap extension PCR mutagenesis was used to construct a knockout (KO) mutant of UGT72B1 in which a conserved catalytic aspartate (D117) was replaced with alanine.[63] This mutant serves as an excellent negative control for use in demonstrating the ability of the fluorogenic assay to distinguish active UGT72B1 from inactive enzymes. The KO gene was constructed in such a way that it could be selectively digested with the restriction endonuclease, *Afe*I, releasing DNA fragments that could be distinguished from uncut wild-type (WT) UGT72B1 gene by gel electrophoresis. Cultures of *E. coli* BL21(DE3) harboring either the pETduet-*UGT72B1-sfGFP* or pETduet-*KO-sfGFP* plasmids were compartmentalized in a w/o emulsion in the presence of UDP-glucose and esculetin, subjected to freeze/thaw cycles, incubated at 37 °C for 20 minutes to allow the reaction to proceed, and reemulsified. Then, the WT and KO droplets were each analyzed by FACS. Gratifyingly, droplets containing lysed cells that had expressed the WT glycosyltransferase displayed a mean fluorescence intensity 16-fold higher than those droplets containing cells that expressed the inactive KO enzyme (Fig. 10). In the absence of exogenously supplied UDP-glucose, there was not a significant difference between the fluorescence intensity of droplets containing the WT or KO glycosyltransferase (Fig. 10). Crucially, this result suggests that any endogenous UDP-glucose within *E. coli* (which is diluted to the volume of the droplet upon cell lysis) does not significantly contribute to the UGT72B1 activity. Furthermore, this data suggests that other, non-native or non-natural NDP-

sugars could be used as target substrates in the droplet screen, thus enabling directed evolution of NDP-donor activity when screening for activity.

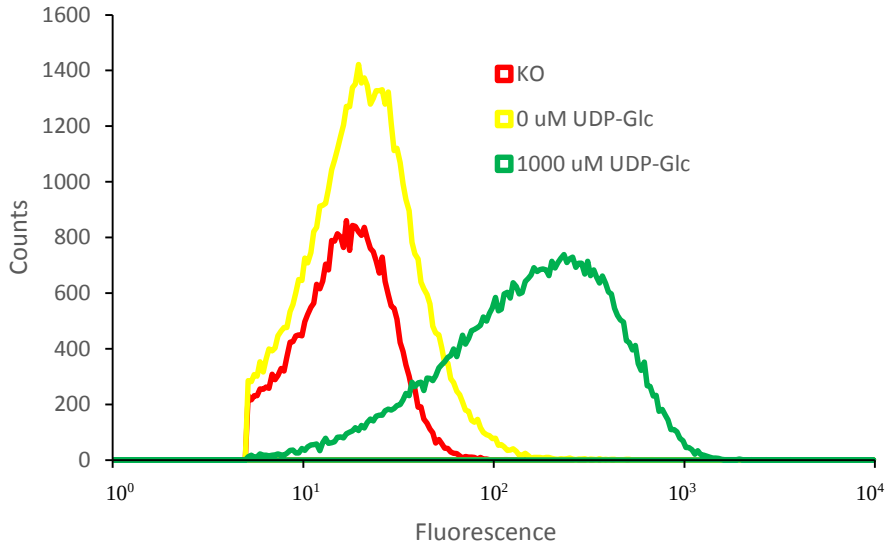

Figure 10. FACS analysis of droplets containing cells expressing WT UGT72B1 with (green) or without (yellow) UDP-glucose and cells expressing KO UGT72B1 (red).

2.2.3 FACS Validation of Freeze-Thaw

With this evidence that the fluorogenic assay faithfully reports UGT72B1 activity in the IVC-FACS format, this assay was used to determine the efficiency of multiple freeze-thaw cycles. Accordingly, *E. coli* BL21(DE3) cells expressing UGT72B1 were compartmentalized in the presence of UDP-glucose and esculetin and were then subjected to a varying number of freeze-thaw cycles. FACS analysis of the treated microdroplets revealed that a single freeze-thaw cycle is sufficient to maximize fluorescence resulting from UGT72B1 catalysis (Fig. 11).

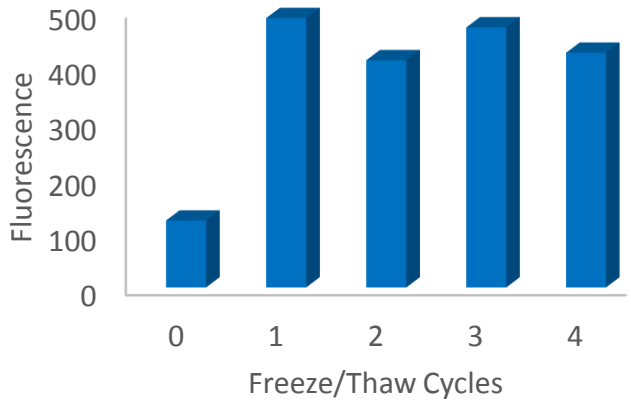

Figure 11. FACS analysis of w/o/w emulsions containing cells expressing UGT72B1 treated with varying number of freeze-thaw cycles. The fluorescence is the mean fluorescence intensity of the entire sorted sample.

2.2.4 Genotype-Phenotype Link and Model Enrichments

Every directed evolution project critically relies on a genotype-phenotype link.[65] With IVC, maintenance of the genotype-phenotype link depends on whether the detected product is retained within the droplet containing the enzyme responsible for its synthesis. Diffusion of the product between droplets would result in a broken genotype-phenotype link and a non-functional screen. Furthermore, because the proteins can not be sequenced from such small quantities or amplified, the plasmid DNA must remain within the same droplet as its encoded enzyme and the detected product. Otherwise, the genotype encoding the desirable phenotype can not be recovered. To test the genotype-phenotype link, cells harboring the UGT72B1 (WT) and KO genes were mixed in a 1:9 ratio, cultured together, compartmentalized, and analyzed by FACS. Gratifyingly, the distinction between the WT and KO populations was maintained (Fig. 12).

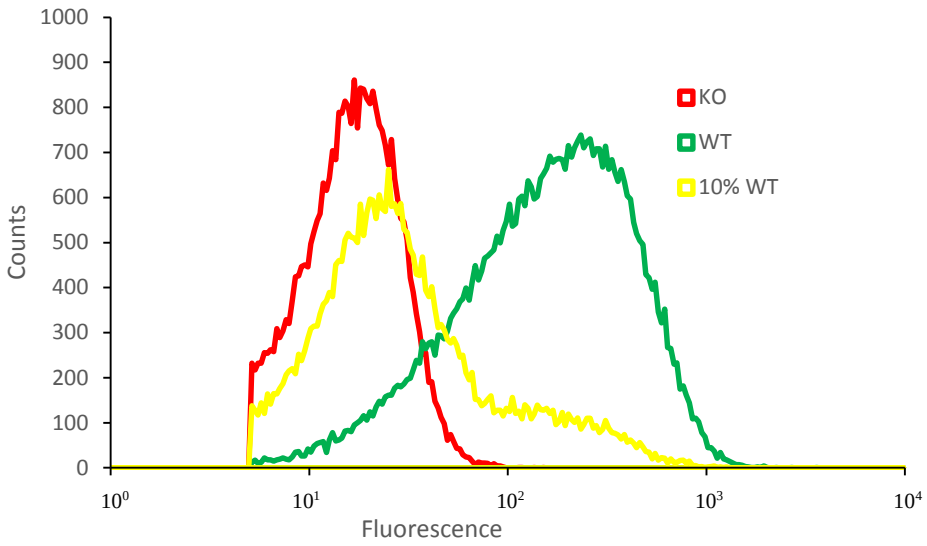

Figure 12. FACS analysis of droplets containing either WT (green), KO (red), or a WT/KO mixture (yellow). The shape of the WT and KO populations are maintained in the 10% WT mix.

To demonstrate the library enriching capabilities of this FACS screen, two model enrichment experiments were performed. In the first, a 1:19 WT/KO sample was prepared and sorted. More than 10^7 droplets were sorted, and the positive population was gated very liberally, with the top 7% of droplets collected without concern for the knowledge that only 5% of the total population is positive (Fig. 13). This gating strategy is suitable to mimic purifying selections where the inactive variants are removed but no selective advantage is provided for improved variants. The plasmid DNA was isolated from the collected droplets by isopropanol precipitation and the gene mixture was amplified by PCR. The nucleotide sequences of the WT and KO genes are 99.8% identical, but to control for amplification bias a control consisting of plasmid DNA isolated from the same 1:19 WT:KO culture used to make the emulsions was amplified under identical PCR conditions. The sorted sample and the

unsorted control were analyzed by *Afe*I restriction digestion and gel electrophoresis (Fig. 14). From densitometric analysis of the cut and uncut band densities, the enrichment factor (final ratio/initial ratio) was calculated to be 8.6-fold. Additionally the PCR mixture was analyzed qualitatively by DNA sequencing. The three nucleotides that distinguish the WT from the KO were significantly enriched (Fig. 15), as judged by inspection of the corresponding sequencing chromatogram.

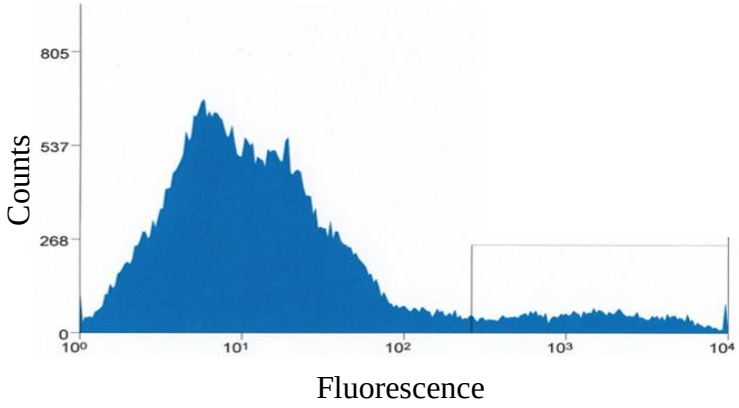

Figure 13. FACS histogram of the 1:19 WT:KO mixed population. The gated area shows the 7% of droplets which were collected.

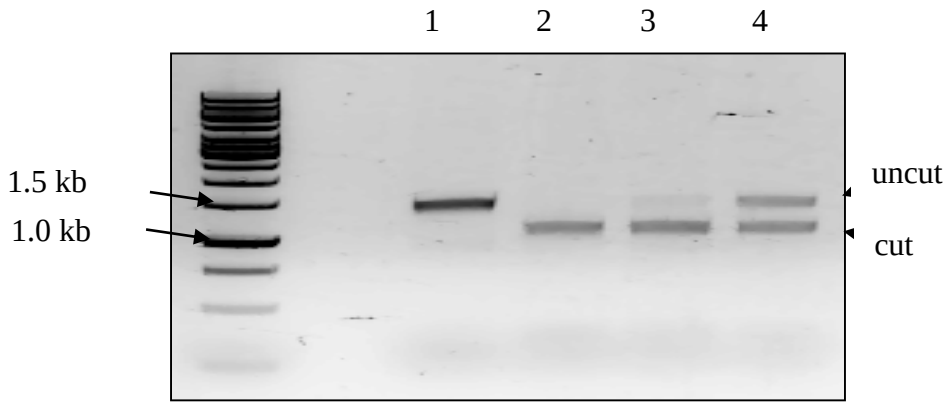

Figure 14. Agarose gel electrophoresis analysis of WT (1), KO (2), control (unsorted) 1:19 WT:KO mix (3), and FACS enriched (4) DNA after *Afe*I digestion. Only the KO gene includes an *Afe*I recognition sequence and is expected to be digested.

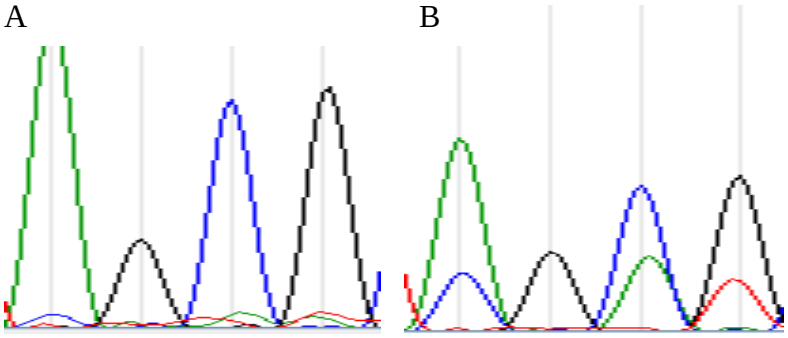

Figure 15. Sequencing of nucleotides 348-351 from the *UGT72B1* gene shown before (A) and after (B) enrichment of a 1:19 WT/KO mixture. The WT sequence, CGAT, is visible as a major contaminate from the KO sequence, AGCG, in the FACS enriched mixture.

Amongst published FACS model enrichments, enrichment factors improve for a given screen for samples in which the positive population is further diluted into the negative. The standard model enrichment initial ratio is 1% positive,[42] and enrichment factors are commonly between 50-fold (33% positive) and 200-fold (50% positive).[53] Accordingly, a model

enrichment with a 1% positive population was performed next. The positive population was gated conservatively, with just the top 0.7% of droplets collected (Fig. 16). The resulting population was approximately 50% WT (100-fold enrichment), as judged by *Afe*I digestion and gel electrophoresis analysis (data not shown). Qualitative analysis by sequencing confirmed substantial enrichment (Fig. 17).

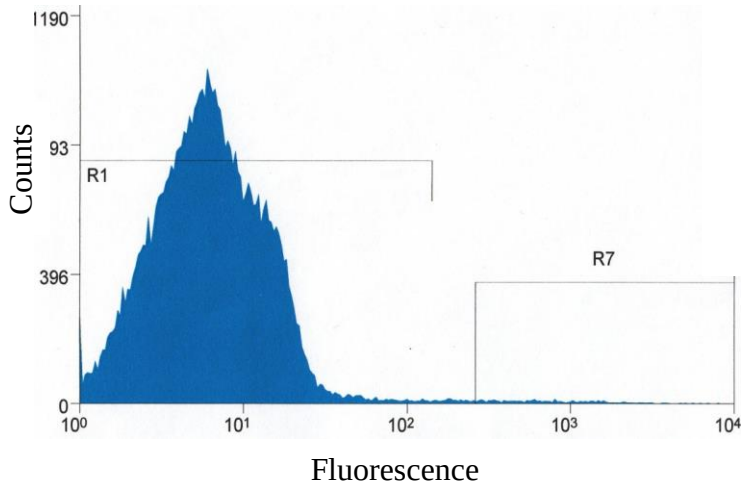

Figure 16. FACS histogram of the 1:99 mixed population. Gated region R7 shows the 0.7% of droplets which were collected.

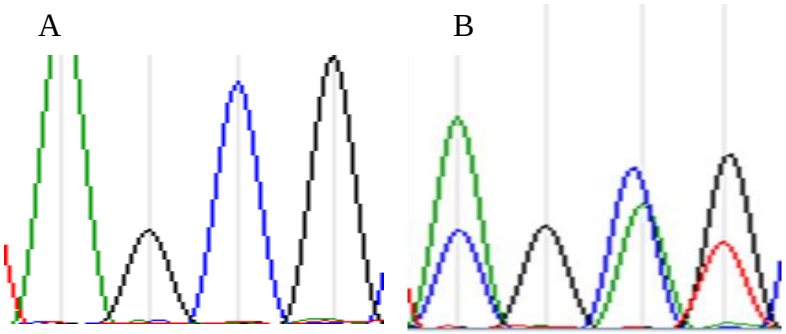

Figure 17. Sequencing chromatograms of nucleotides 348-351 of the UGT72B1 gene shown before (A) and after (B) enrichment of a 1:99 WT/KO mixture. The WT sequence, CGAT, is visible as a major contaminate from the KO sequence, AGCG, in the FACS enriched mixture.

2.3. Conclusions

We have developed the first high expression level bulk emulsion IVC-FACS screen through bacterial cell compartmentalization and lysis, effectively solving a decade old problem that drastically limited broad application of IVC-FACS for directed evolution. The transition from *in vitro* expression kits to IVC lysed cells results in a theoretical 10^3-10^5 fold improvement in expression. Importantly, this strategy can be adopted for any enzyme which can be expressed in *E. coli*. *In vitro* expression kits are inordinately expensive, and thus this advancement also serves to reduce the cost of screening, and, more significantly, screen validation. Additionally, high copy number plasmids could be used to assist in genotype recovery. While mid-copy number plasmids are better for expression of potentially toxic proteins, high copy number plasmids could increase the plasmid DNA isolated per droplet to as high as 10^3 molecules per droplet. The pETduet plasmid used here is maintained at approximately 50 copies per cell. The transition from *in vitro* expression kit strategies that use

one molecule of DNA per droplet to IVC lysed cells results in 50-fold more DNA per droplet isolated.

Compared with *in vivo* FACS screening, this strategy allows the use of exogenous cell impermeable metabolites and reagents and is negligibly interfered with by endogenous metabolites. This is particularly important for enzyme targets in the Williams Lab, which focuses on altering the substrate specificity of enzymes from natural product biosynthesis to accept synthetically modified substrates. It's currently not possible to screen these functions in the presence of their natural metabolites. Additionally, the permeability requirements of the fluorescent product are somewhat reduced compared to fluorescent products suitable for *in vivo* FACS screening. Uncharged small molecule glycosides are typically poorly retained by *E. coli*,[66] but esculin polarity is sufficient for retention within these emulsion microdroplets.

We have developed a new fluorogenic reaction for *in vitro* FACS screening. To the best of our knowledge, no previously reported glycosyltransferase reaction on its natural substrates is fluorogenic. Importantly, the inherent fluorogenic nature of the UGT72B1 reaction provides the basis for an inexpensive secondary screen in microtiter plates. All FACS screens are followed up by more accurate secondary screening, so the development of an assay that provides a distinct signal by FACS and in microtiter plates is distinctly advantageous. FACS screens that rely exclusively on the change in retention of a fluorophore within cells or droplets require validation of a separate secondary screen. The reaction is, to our knowledge, the slowest of all reactions screened by IVC-FACS, which is a testament to the effectiveness of the IVC-lysis strategy. Additionally the UGT72B1 reaction is the first bond-forming reaction to be screened by FACS with natural substrates.

The UDP-glucose dependence of the fluorogenic reaction provides the possibility of coupling UGT72B1 activity to enzymes which produce UDP-glucose. Chapter 3 discusses the construction and testing of coupled enzyme biosensors (CEBs) consisting of UGT72B1 and upstream enzymes involved in nucleotide sugar biosynthesis. These CEBs act as IVC-FACS compatible fluorogenic reporter systems that greatly expand the range of metabolites detectable by IVC-FACS.

2.4. Experimental Section

Unless otherwise stated, all materials and reagents were of the highest grade possible and purchased from Sigma (St. Louis, MO). Isopropyl β-D-thiogalactoside (IPTG) was from Calbiochem (Gibbstown, NJ). Primers were ordered from Integrated DNA Technologies (Coralville, IA). AbilEM90 was purchased from Tego (Germany).

Emulsification and Cell Lysis

The emulsification procedure was adapted from Aharoni et al.[53] *E. coli* BL21(DE3) cells were grown overnight to saturation at 37 °C in LB supplemented with 100 μg/mL ampicillin, and from this starter culture a fresh 3 mL culture of LB supplemented with 100 μg/mL was inoculated, grown at 37 °C to an optical density of 0.6-0.7, induced by a final concentration of 1mM IPTG, and shaken at 22 °C overnight. The cells were spun down at 2500 *g* for 5 min, rinsed twice with ice cold activity buffer (50 mM Tris, 10 mM $MgCl_2$, pH 8), and were resuspended in 0.8 mL ice cold activity buffer. A total volume of 80 μL internal aqueous phase, consisting of resuspended cells as well as any substrates, purified enzymes, and reagents, was placed in a 2 mL cryotube. Immediately 0.8 mL ice-cold oil mix (2.9% Abil EM90 in light

mineral oil) was added, and the phases were homogenized in an ice bath for 5 minutes at 9500 rpm using a Fisher Scientific Power Gen 125 homogenizer with a 7 x 65 mm Fisher Scientific PowerGen disposable plastic generator. The sample was either stored in a -80 °C freezer for storage or placed in an ethanol and dry ice bath for one minute for immediate cell lysis. Next the sample was thawed at room temperature and then incubated at 37 °C for 20 min. To the sample was added 0.8 mL ice-cold aqueous mix (1.5% v/v medium viscosity carboxy methyl cellulose and 1% v/v Triton X102) and the phases were homogenized in an ice bath for 3 min at 8000 rpm to give the double emulsion.

Microscopy

A water in oil emulsion was prepared with *E. coli* cells and 250 μM propidium iodide. The emulsion was aliquoted into two separate samples, one of which was subjected to a freeze-thaw cycle consisting of a one minute freeze in an ethanol and dry ice bath followed by thawing at room temperature for four minutes. 5μL of each sample was mounted on slides and analyzed by fluorescence microscopy using a Zeiss Observer.Z1 microscope.

DNA Manipulations

The *sfGFP* gene was purchased from Theranostech, Inc, and amplified using the following primers: 5'-GCATCATATGAGCAAAGGAGAAGAAC-3' and 5'-TGATGGTACCTTTGTAGAGCTCATCCA -3' (restriction sites underlined). The *sfGFP* gene was then digested with NdeI and KpnI, purified by gel electrophoresis from a 0.8% (w/v) agarose gel, and ligated with similarly treated pETduet-1 using T4 DNA ligase (New England Biolabs, MA), affording the plasmid pETduet-*sfGFP* in which the *sfGFP* gene was inserted into the second (downstream) multiple cloning site of the parent vector. The ligation mixture

was transformed into chemically competent *E. coli* DH5α cells, and single colonies were screened by restriction digestion analysis. The *UGT72B1* gene from *Arabidopsis thaliana* was a kind gift from the Bowles lab (University of York, England) and was amplified by PCR using the 72B1_F and 72B1_R primers: 5'-ATCGCCATGGAGGAATCCAAAACACC-3' and 5'-ATCGAAGCTTGTGGTTGCCATTTTG-3' (restriction sites underlined). The *UGT72B1* gene was first cloned into pET-28a via *Nco*I and *Hin*dIII restriction sites. Then, *UGT72B1* was subcloned into the first multiple cloning site of the pETduet-*sfGFP* construct via *Nco*I and *Hin*dIII restriction sites, affording the plasmid pETduet-*sfGFP-UGT72B1*. The knockout *UGT72B1* mutant D117A was prepared via overlap extension PCR using the following primers: 72B1_F and 5'- GAGCGCTACGACGAGCGCCGTTGGC-3' for the forward fragment and 5'- CGCTCGTCGTAGCGCTCTTCGGTACGGACGCT -3' and 72B1_R for the reverse fragment. The knockout gene was cloned into pETduet and pETduet-*sfGFP* via the *Nco*I and *Hin*dIII restriction sites.

Protein Expression and Purification

UGT72B1 was overexpressed and purified in *E. coli* BL21(DE3) cells. After transformation, sequence verification (data not shown), and replating on LB-agar plates supplemented with 30 μg/mL kanamycin, a single colony was transferred to 3 mL LB supplemented with 30 μg/mL kanamycin and grown at 37 °C and 250 rpm overnight. 1 mL of the saturated culture was transferred to 1 L LB supplemented with 30 μg/mL kanamycin and grown at 37 °C and 250 rpm until the optical density reached a value of 0.6-0.7. At this point, IPTG was added to a concentration of 1mM, the temperature was adjusted to 22 °C, and protein expression proceeded for 18 hours. The cells were collected by centrifugation at 5,000 *g* for 20 min and

resuspended in 10 mL of the wash buffer (20 mM phosphate, 500 mM NaCl, 20 mM imidazole, pH 7.4). Cell suspensions were subjected to sonication and then centrifuged at 14,000 rpm for 30 minutes. The soluble extract was purified by fast protein liquid chromatography using a 1 mL HisTrap HP column (GE Healthcare, NJ). The enzymes were eluted with 20 mM phosphate, 500 mM NaCl, 200 mM imidazole, pH 7.4. The purified proteins were concentrated using an Amicon Ultra 30,000 MWCO centrifugal filter (Millipore Corp., MA) and stored as 20% glycerol stocks at -20 °C. Protein purity was verified by SDS-PAGE, and the Bradford Protein Assay Kit from Bio-Rad was used to estimate protein concentration.

Glycosyltransferase Characterization

In vitro reactions were conducted with 50 μg UGT72B1, 2.5 mM UDP-sugar, and 1 mM esculetin in 100 μL reactions buffered by 50 mM Tris, 10 mM $MgCl_2$, pH 8.0. After 24 hours at room temperature, an equal volume of ice-cold methanol was added and the samples were centrifuged at 10,000 *g*. 25 μL volumes of the quenched reactions were analyzed by HPLC using a Pursuit XRs C18 column (250 x 4.6 mm, Varian Inc.). Esculetin, esculin, and cicchorin were separated by HPLC (detection wavelength of 334 nm) using a series of linear gradients from 0.1% TFA in water (A) to methanol (B): 0 min, 100% A; 0-18 min, 60% B; 18-25 min, 95% B; 25-32 min, 95% B; 32-35 min, 100% A. Additionally, samples of the supernatants were analyzed on a BioTek Hybrid Synergy 4 plate reader (Winooski, VT) for fluorescence emission at 454 nm following excitation at 336 nm.

FACS Assay

Double emulsions were diluted 10 fold in ice-cold phosphate buffered saline (PBS), pH 7.4, and then passed through a 30 μm disposable Celtric filter (Partec, Germany). The samples were

analyzed or sorted on a MoFlo XDP (Beckman Coulter, CA) cell sorter. Events were triggered on fluorescence excited by the 100 mW 488 nm laser in order to ignore droplets which do not contain a cell lysate. Esculin was excited by a 150 mW 355 nm laser. Events were appropriately gated for forward and side scattering to partially offset droplet polydispersity.

FACS Cell Lysis Analysis

E. coli BL21(DE3) cells expressing UGT72B1 and sfGFP were compartmentalized with final concentrations of 1 mM esculetin and 2.5 mM UDP-glucose. Immediately after emulsification the sample was subjected to a number of freeze-thaw cycles and then reemulsified, generating the w/o/w emulsion. In this way, 0 through 4 freeze-thaw cycles were tested by FACS.

Enrichment Factor Analysis

3 mL cultures of *E. coli* BL21(DE3) cells harboring pETduet-*sfGFP-UGT72B1* or pETduet-*sfGFP-KO* vectors were grown to saturation overnight at 37 °C and 250 rpm. 100 uL of these cultures were used to inoculate 10 mL LB which were grown to optical density 0.6. The cultures were combined to produce the 19:1 or 99:1 KO:WT mixture, and to this mixture was added IPTG to a final concentration of 1 mM. The mixture was cultured at 22 °C and 250 rpm for 18 hours. FACS samples were prepared with 1 mM esculetin and 2.5 mM UDP-glucose, and the samples were incubated for 30 minutes at 37 °C after emulsification and freeze-thaw treatment. Over 10^7 events were sorted for each sample, and positive events were collected in 10 mL polypropylene tubes. The collected droplets were diluted to 100 μL with nuclease free water, to which was added 70 μL isopropanol and 10 μL freshly prepared 3M sodium acetate. This solution was centrifuged at 10,000 *g* for 30 minutes, the supernatant was carefully removed and discarded, and the precipitate (not visible) washed twice with 70% ethanol. After

removing the aqueous ethanol, the sample was further dried for 10 minutes at 50 °C. The DNA was dissolved in 5 μL nuclease free water an amplified by PCR with the 72B1_F and 72B1_R cloning primers, and the DNA was sequenced by Genewiz. Digestion reactions with *Afe*I were performed for 3 hours at 37 °C on 200 ng of the DNA, and the bands were quantified after agarose gel electrophoresis with a Typhoon FLA 7000 scanner and its ImageQuant TL software from GE Healthcare Life Sciences (Piscataway, NJ, US).

CHAPTER 3

Broadening the Utility of IVC-FACS Via Coupled Enzyme Biosensors

3.1 Introduction

In addition to the limitations imposed by inadequate *in vitro* protein expression, the primary challenge in the development of a robust ultra-high throughput screening platform based on IVC-FACS is the limited number of fluorescent assays that are compatible with emulsion microdroplets. Traditional fluorescent probes for *in vivo* FACS screening were designed to overcome the permeability restrictions of that format and are mostly limited to hydrolytic reactions that yield an entrapped product.[43,60] Other fluorescent probes developed for enzyme assays were designed for use in microtiter plates and are too lipophilic for use in emulsion microdroplets.

The first rule of directed evolution is "you get what you screen for."[16] Therefore the substrates used for screening should ideally be the intended substrates of the desired mutant. However, most *in vitro* FACS screens use chemically modified substrates that can be appended to the surface of microbeads or cell surface proteins. The unintended consequences of such an approach are best demonstrated by an *in vivo* FACS screen developed for a sialyltransferase.[58] The transfer of the charged sialyl group onto a fluorophore conjugated acceptor substrate rendered the substrate entrapped within the cell. However, initial hits from this screen were found to transfer the sialyl group directly to the fluorophore, rather than to the intended substrate moiety.

The risk of evolving an unintended phenotype due to vague selective pressure means that specific detection of the desired product is preferred, when possible. Considerable efforts

are being made to achieve such detection through genetic biosensors, which regulate expression of an autofluorescent protein or reporter enzyme in a ligand inducible fashion.[38] The ligand binding components of these biosensors are evolvable, and reengineering ligand binding is typically an easier feat than reengineering catalysis due to a number of powerful screening technologies.[67] Although customized genetic biosensors may broaden the scope of *in vivo* enzyme FACS screening, our IVC-FACS platform decouples protein expression from the enzymatic reaction of interest and therefore does not stand to benefit from advances in genetic biosensor engineering. Ultimately, however, the potential of genetic biosensors is restricted by the inadequacies of whole cells as reaction compartments for directed evolution. The development of a custom genetic biosensor for an enzyme product is not of much use if the product is cell permeable or acted on by endogenous enzymes.

Another form of biosensor that is commonly employed for *in vitro* enzyme assays is the coupled enzyme biosensor (CEB). CEBs consist of enzymes which perform a cascade of reactions initiated by a metabolite of interest and concluding in an easily detectable reaction, such as reduction or oxidation of a nicotinamide cofactor yielding a change in absorbance.[68,69] CEBs are commonly developed for microtiter plate assays, and many CEBs are commercially available. Like genetic biosensors, CEBs offer high substrate selectivity and sensitivity. Additionally, like genetic biosensors, the enzymes which make up CEBs have the potential to be reengineered for custom applications.

CEBs are a particularly promising compliment to IVC-FACS. It is envisioned that retrosynthetic CEB design for this format could begin with a fluorogenic enzymatic reaction dependent on a natural, polar metabolite or cofactor. Databases of enzymatic reactions, such

as the Kyoto Encyclopedia of Genes and Genomes (KEGG), can then be used to find suitable reactions that produce that metabolite or cofactor. The UDP-glucose dependence of UGT72B1 represents a remarkable opportunity to create CEBs for a number of important classes of enzymes due to the large number of potentially detectable analytes and the charged phosphate moieties on all intermediates (Fig. 18). These phosphates provide the necessary polarity to ensure a strong genotype-phenotype link is maintained through every step of the CEBs.

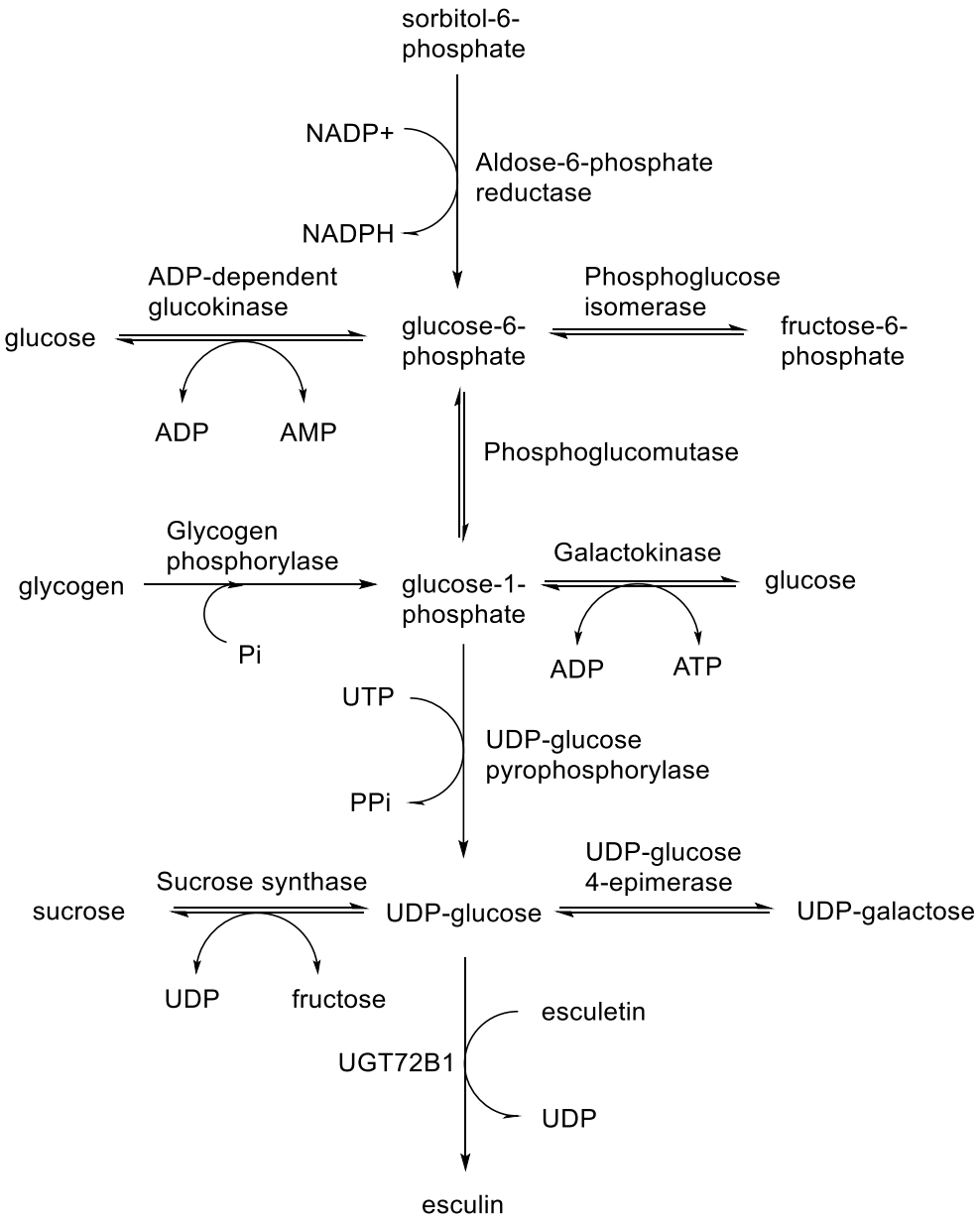

Figure 18. Examples of possible CEBs, designed with assistance from the KEGG database, leading to the UGT72B1 fluorogenic reaction.

Coupled enzyme assays are typically developed using extremely fast enzymes from primary metabolism, so that the reaction being assayed is rate limiting. If that reaction is not

rate limiting, the CEB can only report the yield of that reaction but not its kinetics.[70] While improving kinetic parameters of enzymes is frequently an enzyme engineering goal, improving the yield of the desired product is typically more important.[71] An enzyme might rapidly perform the desired reaction but suffer from low yields because it also catalyzes undesirable substrate or product degradation. Accordingly, sluggishness of UGT72B1 likely prevents its potential upstream CEBs (Fig. 18) from being useful for kinetic characterization. However, the endpoint fluorescence from potential UGT72B1 CEBs was expected to be dependent on the concentration of the limiting analyte.

In this chapter, to demonstrate the potential of CEBs with IVC-FACS (Fig. 19), a series of CEBs was designed that allows detection of uridine diphosphate glucose (UDP-glucose), uridine triphosphate (UTP), glucose-1-phosphate, glucose-6-phosphate, sorbitol-6-phosphate, and oxidized nicotinamide adenine dinucleotide phosphate ($NADP^+$). The relevance of these CEBs for directed evolution projects in the area of natural product synthetic biology will be discussed in Chapter 6.

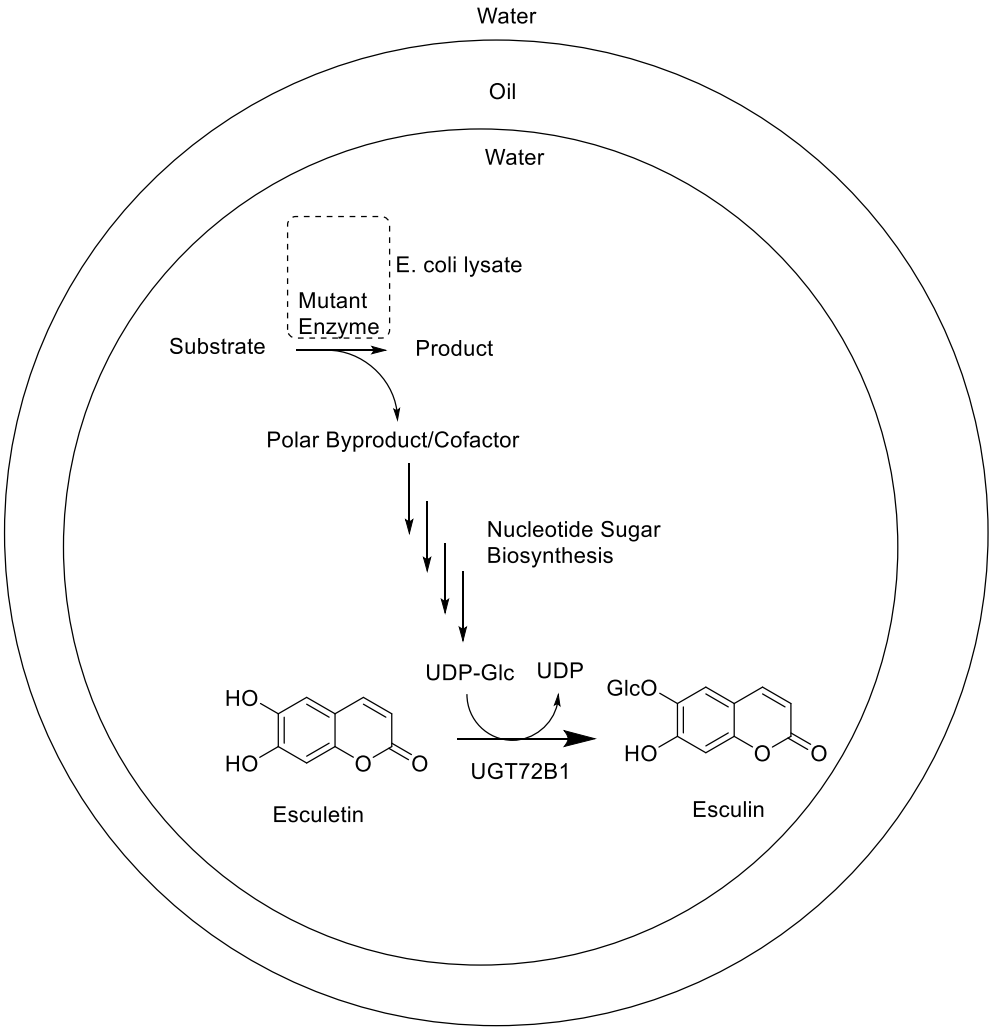

Figure 19. Scheme for a CEB-IVC-FACS platform using UGT72B1 for the fluorogenic step.

3.2. Results and Discussion

3.2.1 UGT72B1 Reaction Analysis

To be a useful enzyme to conclude the proposed sequence of CEBs, the fluorescence signal generated by the UGT72B1-catalyzed reaction should respond to varying initial concentrations of UDP-glucose. To test whether this criteria was met, purified UGT72B1 was

assayed in wells of a 96-well microtiter plate using concentrations of UDP-glucose that would be representative of those found in the proposed CEB-IVC-FACS screening platform. Gratifyingly, it was confirmed that the endpoint fluorescence of the UGT72B1-catalyzed reaction was dependent on initial UDP-glucose concentration (Fig. 20 and 21).

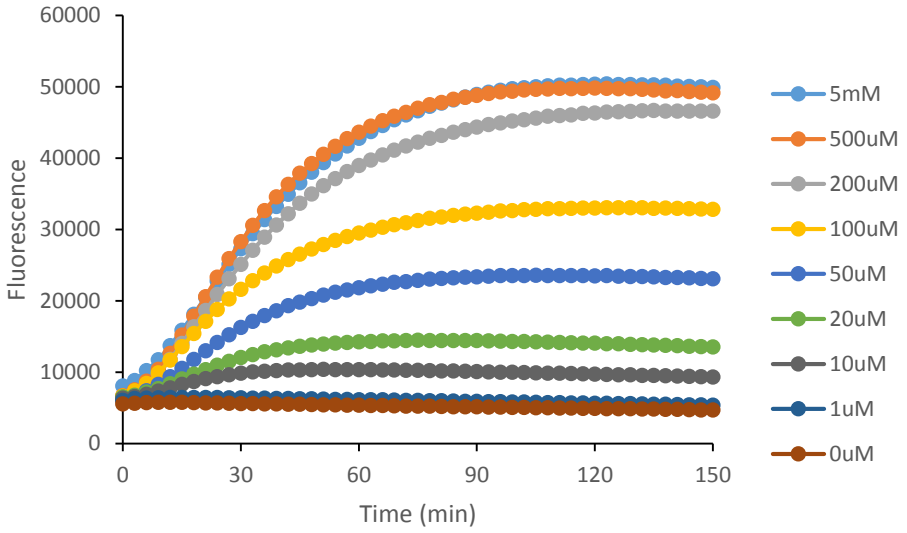

Figure 20. Timecourse of UGT72B1 generated fluorescence with varying concentrations of UDP-glucose and 100 μM esculetin.

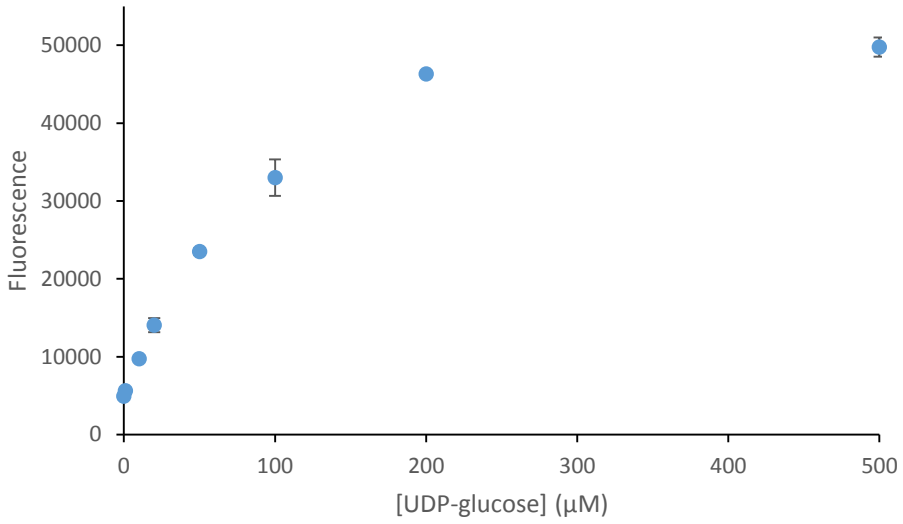

Figure 21. Fluorescence of UGT72B1 reactions with 100 μM esculetin and varying concentrations of UDP-glucose after two hours. Error bars represent standard deviation of the mean (n = 3).

Next the reaction conditions were altered for FACS analysis, because escape of excess esculetin through the double emulsion permits the use of significantly higher concentrations of esculetin (10-fold greater than on microtiter plates). *E. coli* BL21(DE3) pETduet-*sfGFP* was used to prepare w/o/w emulsions, as described in Chapter 2. Purified UGT72B1, varying amounts of UDP-glucose, and 1 mM esculetin were added to the buffered suspensions of *E. coli* BL21(DE3) pETduet-*sfGFP* and were homogenized with the oil layer to yield w/o emulsions. These emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 30 minutes at 37 °C to allow the enzymatic reaction to proceed. The droplets were reemulsified to generate the w/o/w emulsions and analyzed by FACS (Fig. 22). Across the entire UDP-glucose concentration range tested, a clear linear relationship between initial UDP-glucose and mean fluorescence was observed (Fig. 23).

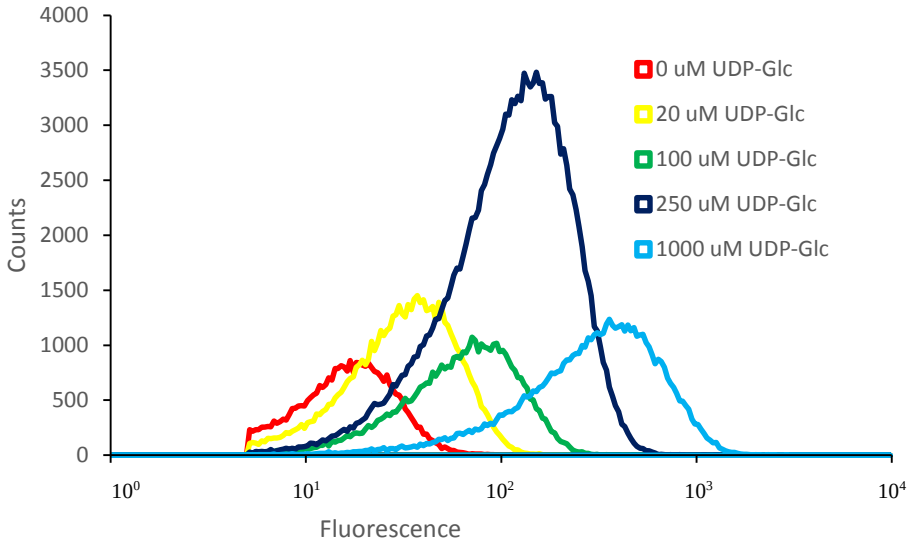

Figure 22. IVC-FACS populations of *E. coli* lysates supplemented with purified UGT72B1 and increasing concentrations of UDP-glucose.

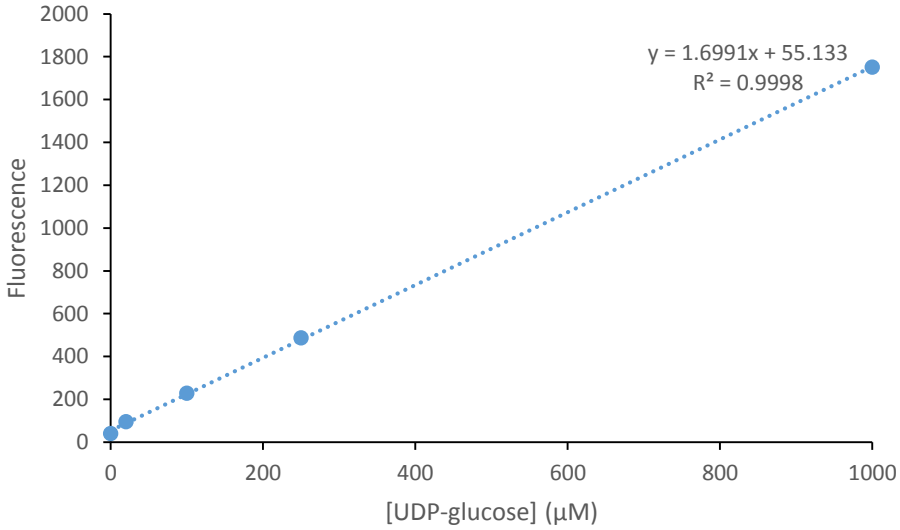

Figure 23. Calibration curve of UGT72B1 generated fluorescence with varying concentrations of UDP-glucose in *E. coli* lysates by IVC-FACS. A linear relationship between initial UDP-glucose concentration and mean fluorescence was observed.

Thus the UGT72B1-catalyzed fluorogenic glycosylation reaction can be used to report UDP-glucose concentrations in droplets containing lysed *E. coli* cells and assay components. Notably, UDP-glucose is produced by several interesting enzymatic reactions, any of which might be desirable templates for enzyme engineering. For example, UDP-glucose can be biosynthesized by nucleotidyltransferases, UDP-galactose-4-epimerase, and the reverse action of glycogen synthase or other glucosyltransferases. All of these reactions are likely compatible with IVC-FACS technology.

3.2.2 Coupling UGT72B1 to RmlA

Nucleotidyltransferase RmlA from *Salmonella enterica* was chosen as the second step in the CEB pathway (Fig. 24) due to its extensive prior characterization[72-75] and the value of its substrates as intermediates in more extensive CEBs according to the KEGG database. This enzyme is capable of producing UDP-glucose from UTP and glucose-1-phosphate. Both UTP and glucose-1-phosphate are presumably droplet impermeable and therefore represent potential intermediates for various CEBs.

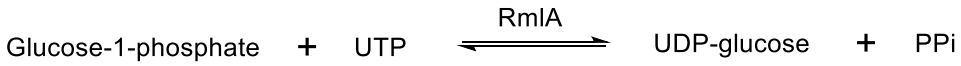

Figure 24. RmlA catalyzes the production of UDP-glucose from glucose-1-phosphate and UTP.

To test whether RmlA and UGT72B1 could be used together for fluorescence detection of glucose-1-phosphate and UTP, purified RmlA and UGT72B1 were tested with varying

initial concentration of UTP and glucose-1-phosphate in microtiter plates. It was confirmed that endpoint fluorescence was dependent on the initial concentration of the variable substrate (Fig. 25-27). No fluorescence increase was observed in the absence of UTP, glucose-1-phosphate or RmlA. Interestingly, the RmlA/UGT72B1 CEB displayed significantly lower activity when glucose-1-phosphate was the limiting substrate (Fig. 27).

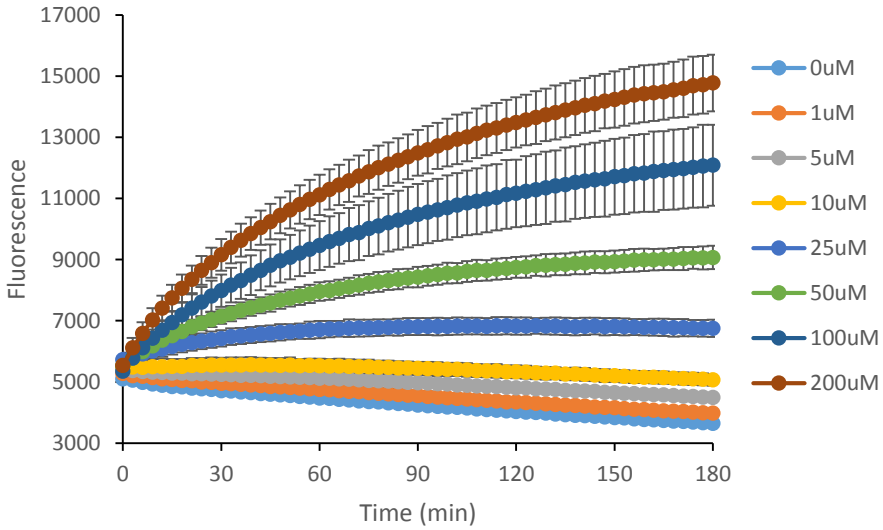

Figure 25. Timecourse of RmlA/UGT72B1 generated fluorescence with varying concentrations of glucose-1-phosphate. Error bars represent standard deviation of the mean (n = 3).

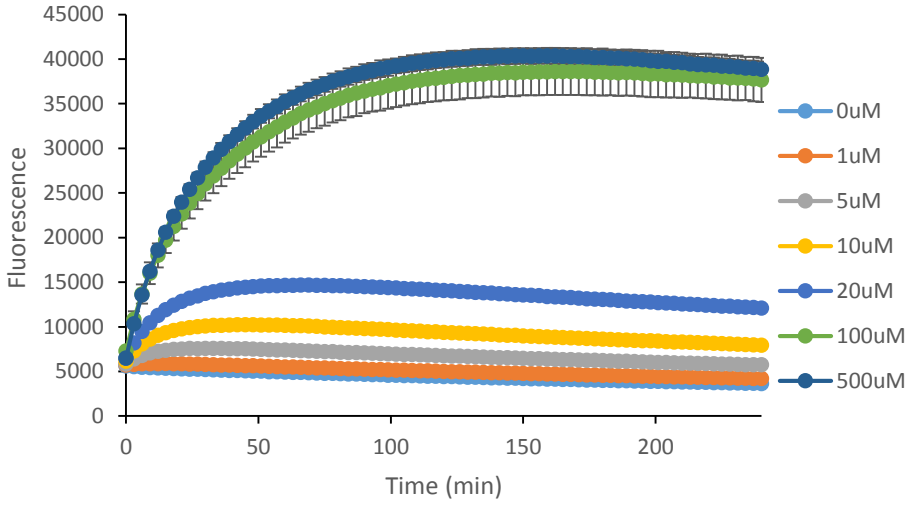

Figure 26. Timecourse of RmlA/UGT72B1 generated fluorescence with varying concentrations of UTP. Error bars represent standard deviation of the mean (n = 3).

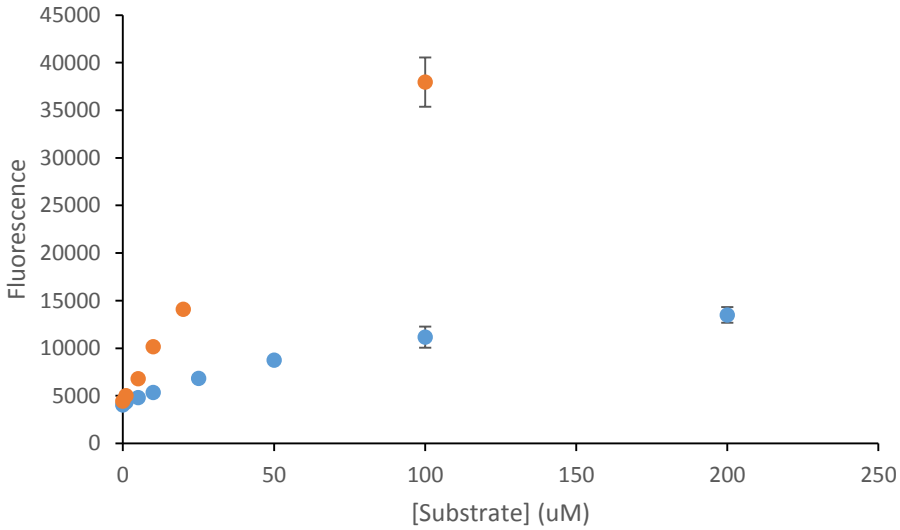

Figure 27. Fluorescence of RmlA/UGT72B1 reactions detecting varying concentrations of UTP (orange) or glucose-1-phosphate (blue) after two hours. Error bars represent standard deviation of the mean (n = 3).

W/o emulsions were then generated as previously described from *E. coli* BL21(DE3) pETduet-*sfGFP* supplemented with purified RmlA and UGT72B1, 1 mM esculetin, and 1 mM of either UTP or glucose-1-phosphate and the other RmlA substrate added at varying concentrations. These emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 60 minutes at 37 °C to allow the enzymatic reactions to proceed. The droplets were re-emulsified to generate the w/o/w emulsions and analyzed by FACS (Fig. 28 and 29). A clear linear relationship between initial variable analyte concentration and mean fluorescence was observed, indicating the suitability of the RmlA/UGT72B1 as a biosensor for directed enzyme evolution by IVC-FACS (Fig. 30 and 31).

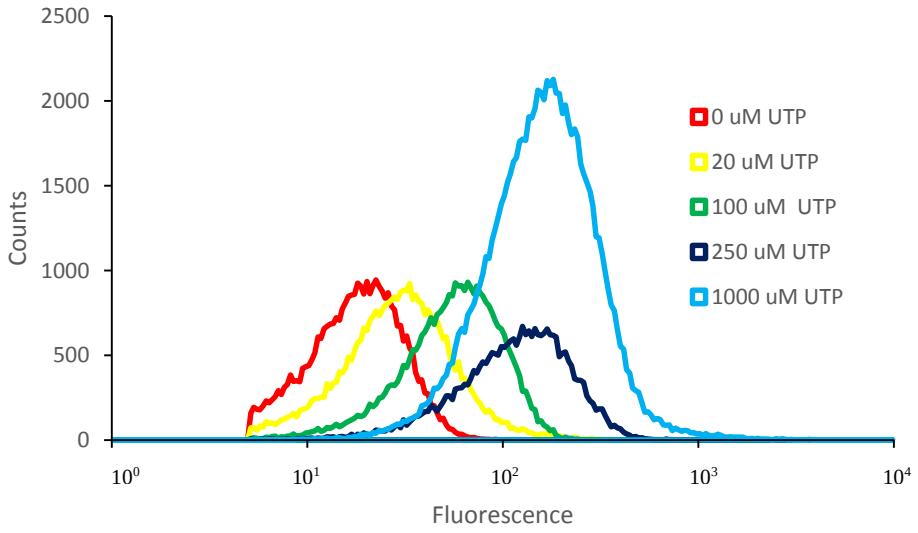

Figure 28. IVC-FACS populations of *E. coli* lysates supplemented with purified RmlA and UGT72B1 and varying UTP concentrations.

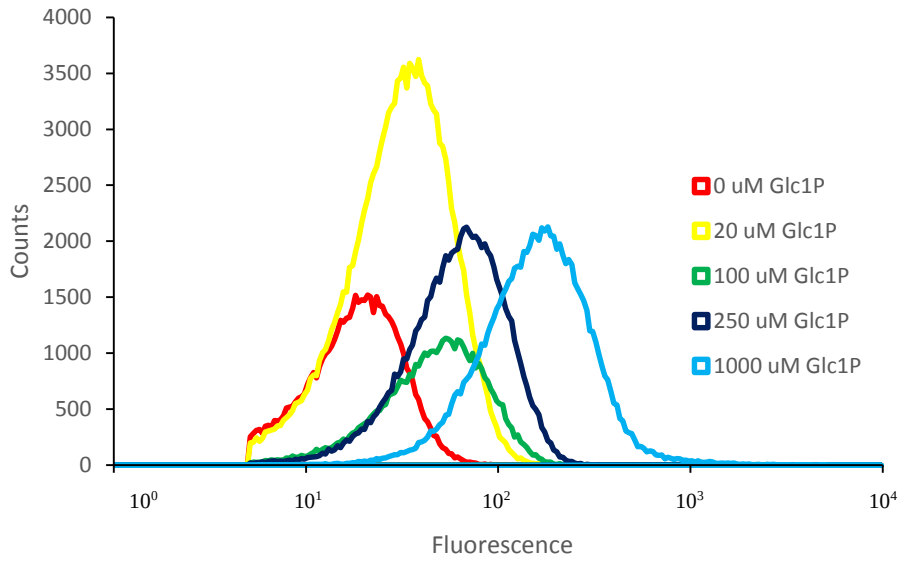

Figure 29. IVC-FACS populations of *E. coli* lysates supplemented with purified RmlA and UGT72B1 and varying glucose-1-phosphate concentrations.

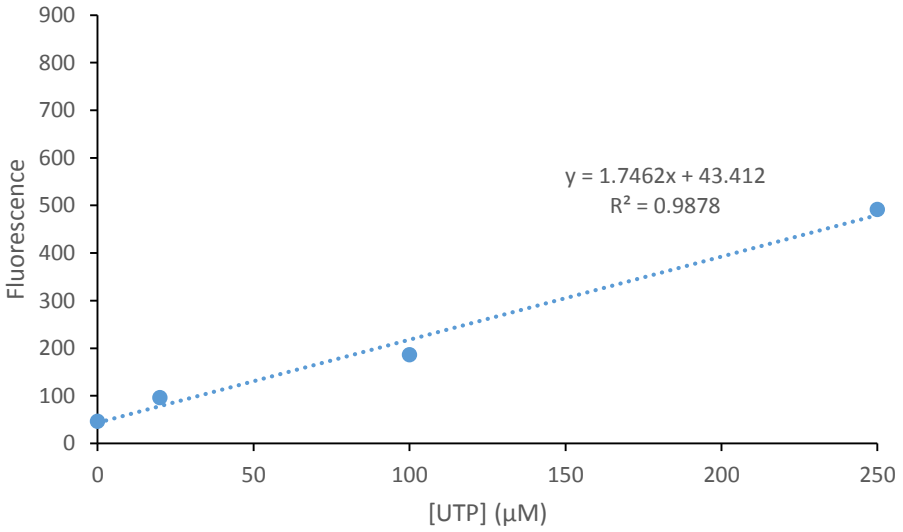

Figure 30. Calibration curve of fluorescence generated by purified RmlA and UGT72B1 with varying concentrations of UTP in *E. coli* lysates by IVC-FACS. A linear relationship between initial UTP concentration and mean fluorescence was observed for concentrations between 0 and 250 μM.

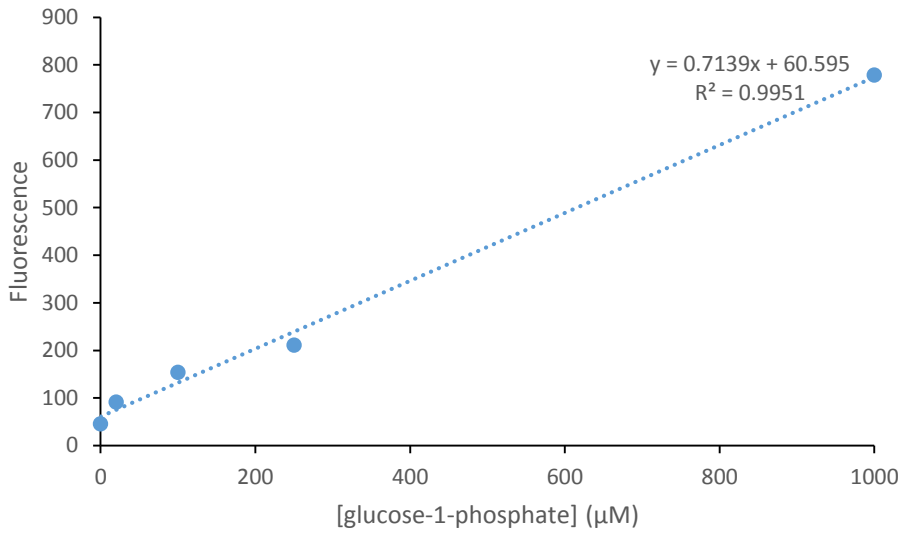

Figure 31. Calibration curve of fluorescence generated by purified RmlA and UGT72B1 with varying concentrations of glucose-1-phosphate in *E. coli* lysates by IVC-FACS. A linear relationship between initial glucose-1-phosphate concentration and mean fluorescence was observed.

This CEB-IVC-FACS screen can be used for directed evolution of several enzymes which produce UTP or glucose-1-phosphate. For example, UTP is produced by UDP phosphotransferase, nucleoside triphosphate-nucleoside monophosphate transphosphorylase, CTP aminohydrolase, and pyruvate 2-O-phosphotransferase. Glucose-1-phosphate is produced by phosphoglucomutase (PGM), sucrose phosphorylase, and glycogen phosphorylase.

3.2.3 Coupling PGM to RmlA and UGT72B1

Glucose-1-phosphate can be reversibly converted to glucose-6-phosphate by the enzyme PGM. This reaction was chosen for a number of reasons. First, glucose-6-phosphate is a fantastic intermediate for a CEB. We envisioned detection of oxidized cofactors through further coupling of glucose-6-phosphate to enzymatic oxidation of sorbitol-6-phosphate.

Second, PGM is commercially available, and thus its use provides additional evidence for the benefits of the modularity of CEBs for IVC-FACS. As more enzymes are commercialized due to the growing interest in multi-step *in vitro* biocatalysis, development of CEBs will become increasingly easier.

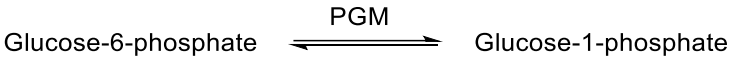

Figure 32. PGM catalyzes the interconversion of glucose-6-phosphate and glucose-1-phosphate.

This was also an opportunity to test whether CEBs could be effective across potentially inefficient intermediate catalytic steps. The thermodynamic equilibrium between glucose-1-phosphate and glucose-6-phosphate strongly favors glucose-6-phosphate (17.8 to 1).[76] Therefore the glucose-1-phosphate concentration will always be low, possibly exacerbating the reduced sensitivity of the downstream CEB toward glucose-1-phosphate relative to UTP. Successful, sensitive detection of glucose-6-phosphate by FACS with the PGM/RmlA/UGT72B1 biosensor would highlight the potential range and robustness of CEB-IVC-FACS screens.

Exchanging the commercial PGM solution into 50 mM Tris, pH 8.0, 20% glycerol buffer to remove the ammonium sulfate was necessary in order to detect activity with the PGM/RmlA/UGT72B1 biosensor. The purified enzymes were assayed in wells of a 96-well microtiter plate with varying concentrations of glucose-6-phosphate (Fig. 33). It was confirmed that endpoint fluorescence values were dependent on initial glucose-6-phosphate concentration

(Fig. 34). The sensitivity of this CEB toward glucose-6-phosphate (Fig. 34) was comparable to RmlA/UGT72B1 detection of glucose-1-phosphate (Fig. 27).

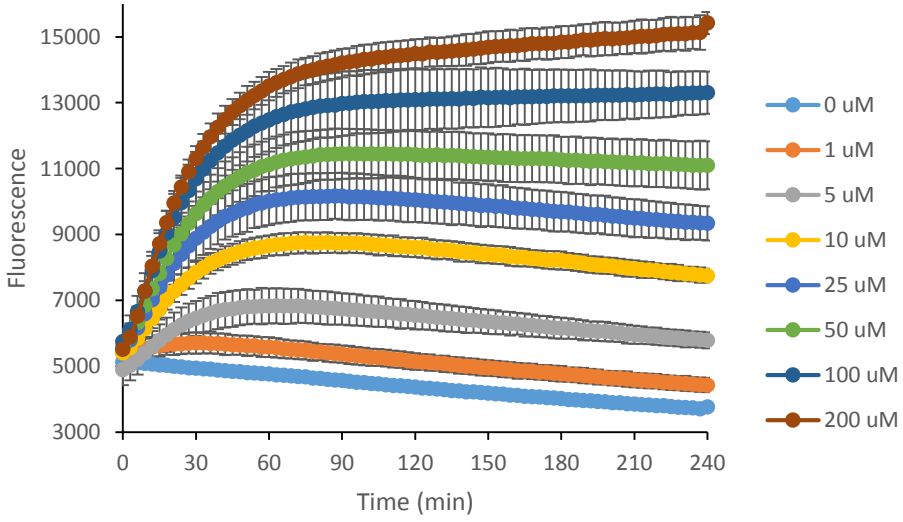

Figure 33. Timecourse of PGM/RmlA/UGT72B1 generated fluorescence with varying concentrations of glucose-6-phosphate. Error bars represent standard deviation of the mean (*n* = 3).

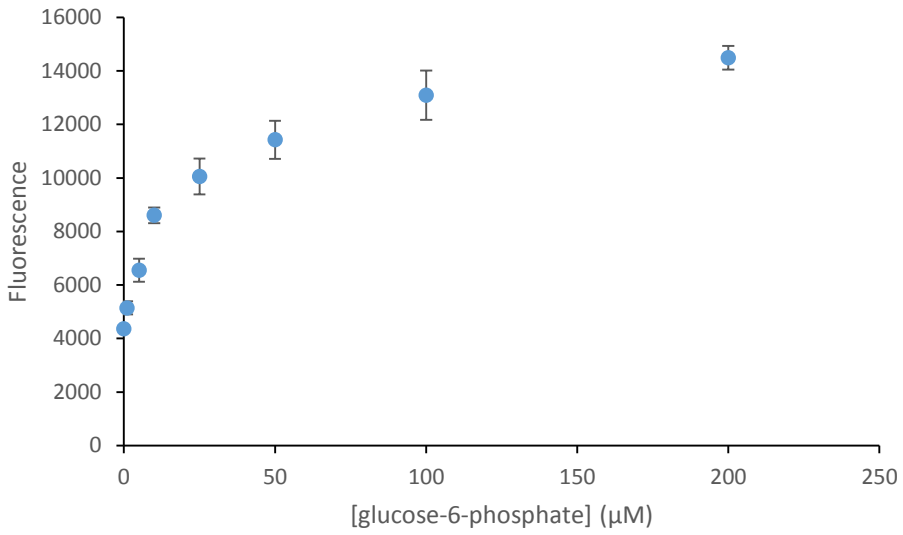

Figure 34. Fluorescence of PGM/RmlA/UGT72B1 reactions with varying concentrations of glucose-6-phosphate after two hours. Error bars represent standard deviation of the mean (*n* = 3).

W/o emulsions were generated as previously described from *E. coli* BL21(DE3) pETduet-*sfGFP* supplemented with purified PGM, RmlA and UGT72B1. The internal aqueous phase also contained 1 mM esculetin, 1 mM UTP, and varying concentrations of glucose-6-phosphate. These emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 90 minutes at 37 °C to allow the enzymatic reactions to proceed. The droplets were re-emulsified to generate the w/o/w emulsions and analyzed by FACS (Fig. 35). The relationship between initial glucose-6-phosphate concentration and mean fluorescence was linear (Fig. 36).

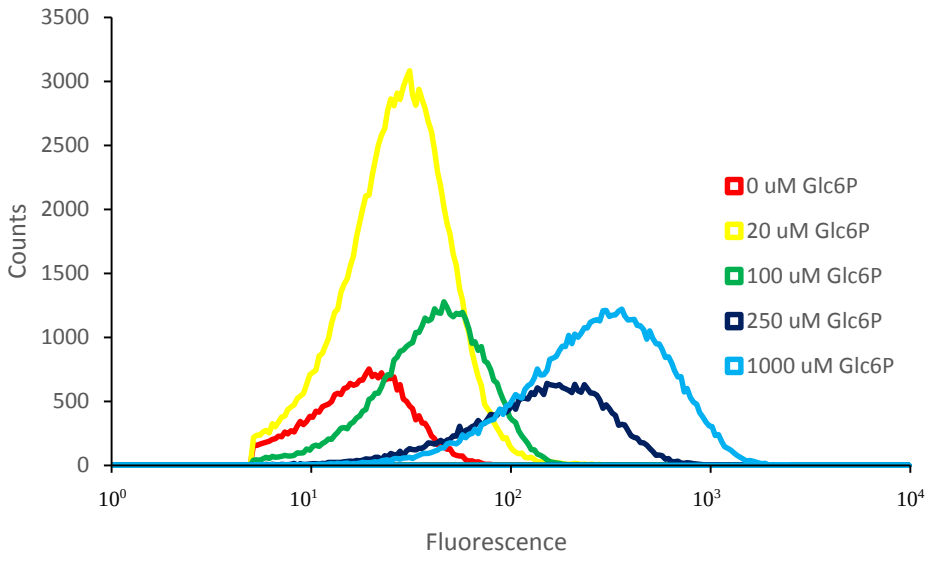

Figure 35. IVC-FACS populations of *E. coli* lysates supplemented with purified PGM, RmlA, and UGT72B1 and varying glucose-6-phosphate concentration.

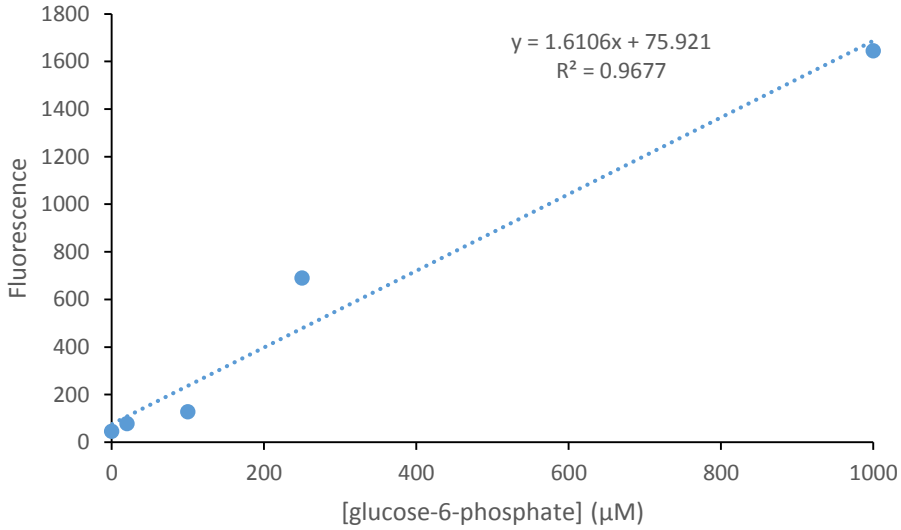

Figure 36. Calibration curve of fluorescence generated by purified PGM, RmlA, and UGT72B1 and varying concentrations of glucose-6-phosphate in *E. coli* lysates by IVC-FACS. A linear relationship between initial glucose-6-phosphate concentration and mean fluorescence was observed.

These results demonstrate that glucose-6-phosphate can be detected at concentrations relevant for directed evolution in *E. coli* lysates by the PGM/RmlA/UGT72B1 IVC-FACS screen. This screen can be used to evolve the following glucose-6-phosphate producing enzymes: aldose-6-phosphate reductase, hexokinase, and phosphoglucose isomerase.

3.2.4 Coupling A6PR to PGM, RmlA, and UGT72B1

Aldose-6-Phosphate Reductase (A6PR) from *Malus domestica* is known to catalyze the reversible NADPH-dependent reduction of glucose-6-phosphate to sorbitol-6-phosphate.[77] We exploited this reversibility to detect the $NADP^+$ dependent oxidation of sorbitol-6-phosphate. Many biomolecules are reduced by NADPH dependent reductases or oxidized by NADPH dependent monooxygenases. We envision broad application of the A6PR/PGM/RmlA/UGT72B1 CEB pathway for detection of natural product tailoring events, as well as the development of additional CEBs that contain those tailoring enzymes to detect successful assembly of natural products.

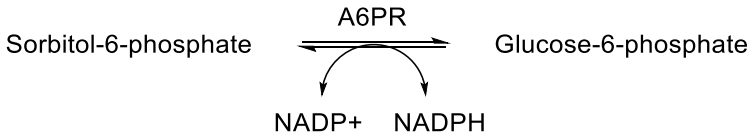

Figure 37. A6PR catalyzes the reversible, $NADP^+$ dependent oxidation of sorbitol-6-phosphate to glucose-6-phosphate.

Purified A6PR, PGM, RmlA, and UGT72B1 were assayed in wells of a 96-well microtiter plate with varying concentrations of sorbitol-6-phosphate or $NADP^+$ (Fig. 38 and 39). Endpoint fluorescence values were dependent on initial concentration of the limiting

substrate (Fig. 40). As was described for PGM, extension of the CEB pathway to include A6PR for sorbitol-6-phosphate or $NADP^+$ detection did not significantly reduce fluorescent output in a microtiter plate.

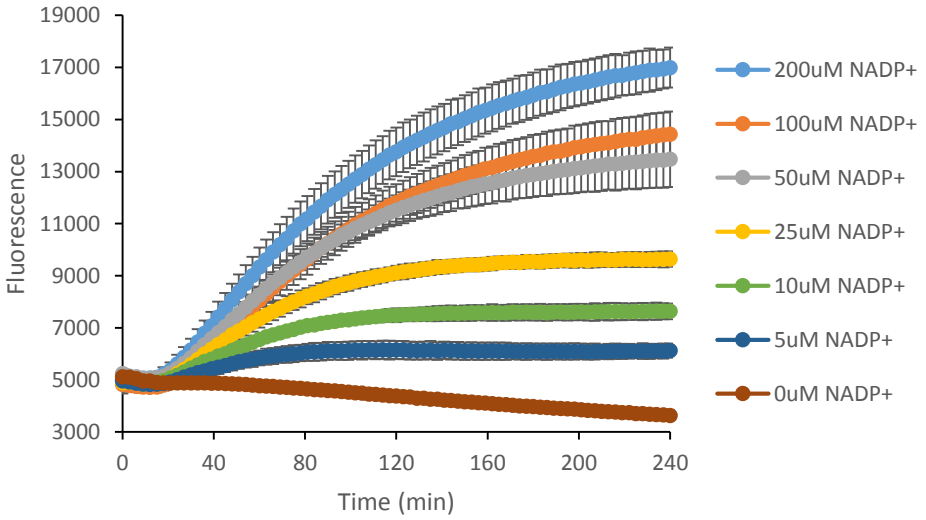

Figure 38. Timecourse of A6PR/PGM/RmlA/UGT72B1 generated fluorescence with varying concentrations of $NADP^+$. Error bars represent standard deviation of the mean (n = 3).

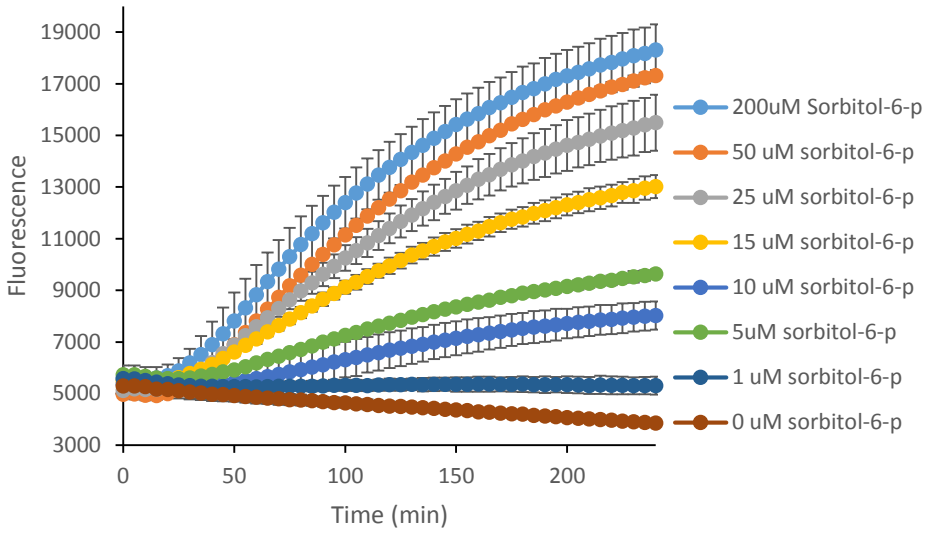

Figure 39. Timecourse of A6PR/PGM/RmlA/UGT72B1 generated fluorescence with varying concentrations of sorbitol-6-phosphate. Error bars represent standard deviation of the mean (*n* = 3).

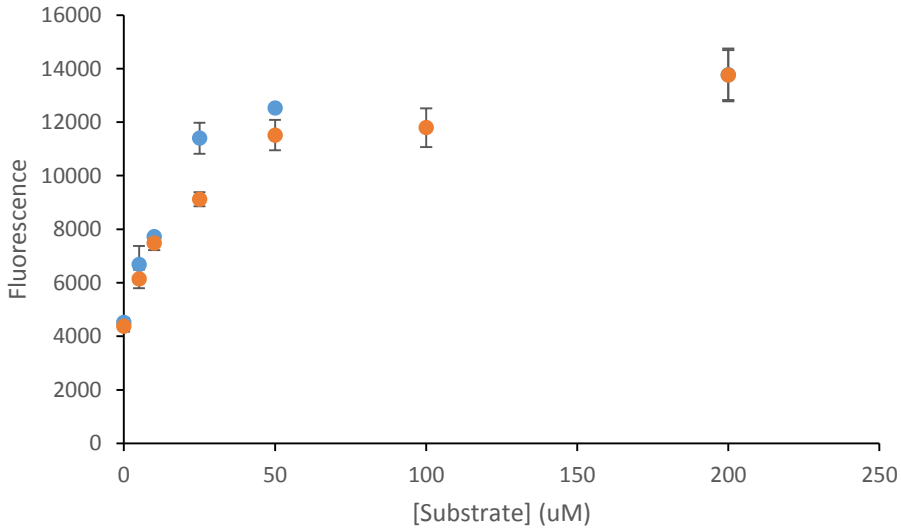

Figure 40. Fluorescence of A6PR/PGM/RmlA/UGT72B1 reactions detecting varying concentrations of $NADP^+$ (orange) or sorbitol-6-phosphate (blue) after two hours. Error bars represent standard deviation of the mean (*n* = 3).

W/o emulsions were generated as previously described from *E. coli* BL21(DE3) pETduet-*sfGFP* supplemented with purified RmlA, PGM, RmlA and UGT72B1. Additionally, the internal aqueous phase contained 1 mM esculetin, 1 mM UTP, 1 mM sorbitol-6-phosphate and varying concentrations of $NADP^+$. These emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 120 minutes at 37 °C to allow the enzymatic reactions to proceed. The droplets were re-emulsified to generate the w/o/w emulsions and analyzed by FACS (Fig. 41). The fluorescence output of the CEB was dependent on initial $NADP^+$ concentration (Fig. 42). Presumably due to interference from the lysate, the sensitivity was greatly diminished compared to the detection of the other analytes by the preceding CEBs. Future efforts to improve the sensitivity are discussed in Chapter 6.

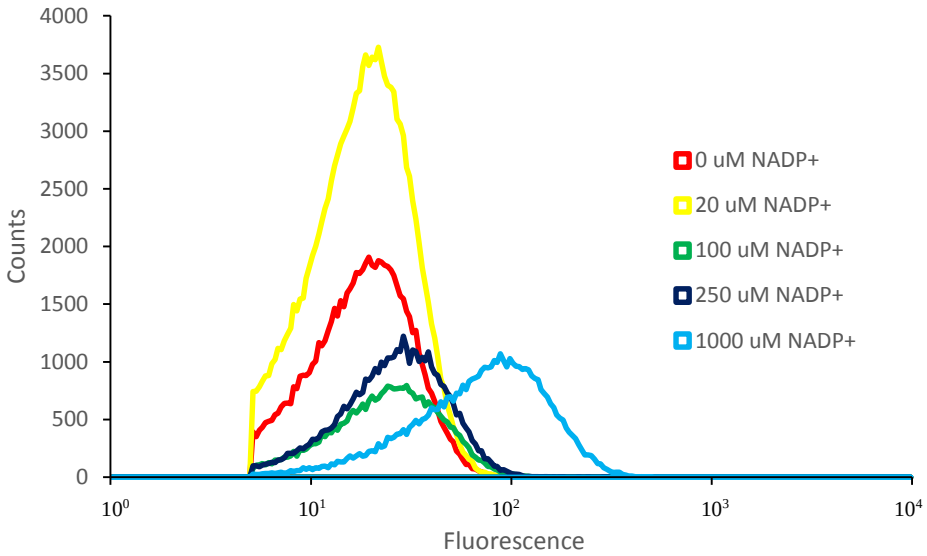

Figure 41. IVC-FACS populations of purified A6PR, PGM, RmlA, and UGT72B1 with increasing $NADP^+$ concentration.

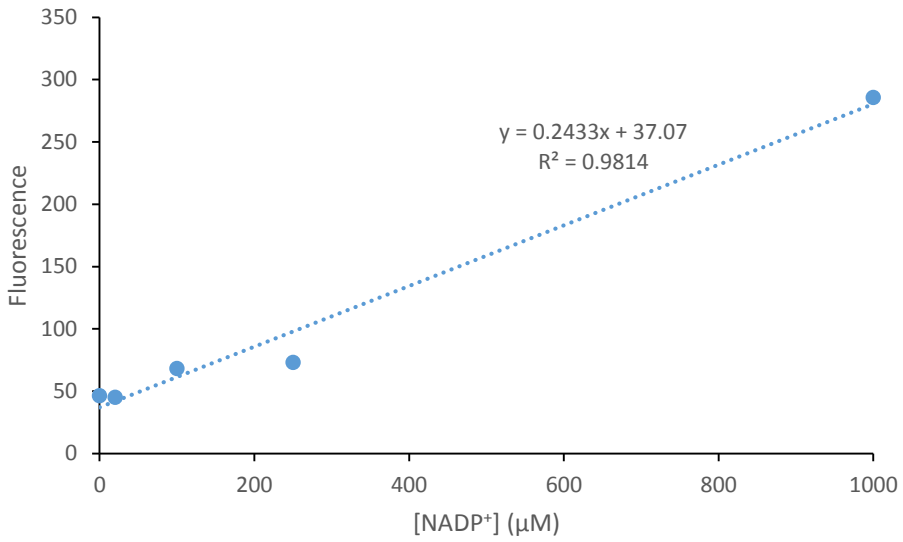

Figure 42. Calibration curve of purified A6PR, PGM, RmlA, and UGT72B1 generated fluorescence with varying concentrations of $NADP^+$ by IVC-FACS.

While each of the preceding CEB-IVC-FACS screens are suitable for directed evolution of several enzymes, significantly more opportunities are afforded by a CEB-IVC-FACS screen for $NADP^+$. Monooxygenases and reductases could be screened directly with this system or coupled to it for detection of a wide range of additional biomolecules. The intended applications of this CEB-IVC-FACS screen are discussed in Chapter 6.

3.3. Conclusions

The A6PR/PGM/RmlA/UGT72B1 CEB pathway and truncated variations provide fluorescent output dependent on the concentration of several biologically produced metabolites and cofactors. Importantly, these CEBs are IVC-FACS compatible because the intermediates

and product are sufficiently polar to guarantee a strong genotype-phenotype link in emulsion microdroplets.

The sensitivity and dynamic range of CEB-IVC-FACS screens are well suited for directed evolution. Like genetic biosensors, the components of a CEB are inherently evolvable for altered analyte recognition. Additionally, CEBs composed of purified enzymes are highly modular and, as shown, can be constructed rationally based on existing knowledge of enzymatic properties. Perhaps most importantly, CEBs can be validated in microtiter plates and can be used as the secondary screen in microtiter plates after library enrichment by FACS. FACS screens that work by product entrapment but without an inherently fluorogenic reaction require an additional microtiter plate screen to be validated for secondary screening.

The discovery or engineering of new IVC-FACS compatible fluorogenic enzymatic reactions will provide alternate routes for CEBs besides nucleotide sugar biosynthesis. The rapid development of molecular probes for microtiter plate lysate screening suggests that if IVC-FACS is widely regarded as a potentially near universal screening platform, new reactions for IVC-FACS CEBs will rapidly be developed. Unlike nearly all existing examples of fluorogenic reactions for FACS, development of a CEB for IVC-FACS enables the development of additional screens by shortening the path to other biomolecule analytes.

3.4. Experimental Section

Unless otherwise stated, all materials and reagents were of the highest grade possible and purchased from Sigma (St. Louis, MO). Isopropyl β-D-thiogalactoside (IPTG) was from Calbiochem (Gibbstown, NJ). Primers were ordered from Integrated DNA Technologies

(Coralville, IA). The *RmlA* gene was received in pET-28a from the Thorson lab (University of Kentucky, US). The *A6PR* gene was received in pET-19b from the Iglesias lab (Universidad Nacional Del Litoral, Argentina). Phosphoglucomutase was purchased from Sigma.

Protein Expression and Purification

UGT72B1 was overexpressed and purified as described in Chapter 2. RmlA and A6PR were overexpressed and purified from *E. coli* BL21(DE3) cells. After transformation, sequence verification, and replating on LB-agar plates supplemented with 30 μg/mL kanamycin (RmlA) or 100 μg/mL ampicillin (A6PR), a single colony was transferred to 3 mL LB supplemented with the appropriate antibiotic and grown at 37 °C and 250 rpm overnight. 1 mL of the saturated culture was transferred to 1 L LB supplemented with the appropriate antibiotic and grown at 37 °C and 250 rpm until the optical density reached a value of 0.6-0.7. At this point, IPTG was added to a concentration of 1mM, the temperature was adjusted to 28 °C for RmlA and 25 °C for A6PR, and protein expression proceeded for 18 hours. The cells were collected by centrifugation at 5,000 *g* for 20 min and resuspended in 10 mL of the wash buffer (20 mM phosphate, 500 mM NaCl, 20 mM imidazole, pH 7.4). Cell suspensions were subjected to sonication and then centrifuged at 14,000 rpm for 30 minutes. The soluble extract was purified by fast protein liquid chromatography using a 1 mL HisTrap HP column (GE Healthcare, NJ). The enzymes were eluted with 20 mM phosphate, 500 mM NaCl, 200 mM imidazole, pH 7.4. The purified proteins were concentrated using an Amicon Ultra 10,000 MWCO centrifugal filter (Millipore Corp., MA) and stored as 20% glycerol stocks at -20 °C. Protein purity was verified by SDS-PAGE, and the Bradford Protein Assay Kit from Bio-Rad was used to estimate

protein concentration. PGM was buffer exchanged into 50 mM Tris, pH 8.0, supplemented with 20% glycerol using an Amicon Ultra 10,000 MWCO centrifugal filter.

Microplate Assay

Reactions were performed with 5 μg of each enzyme catalyst. Reactions were performed in 50 mM Tris buffer, pH 8.0, with 10 mM $MgCl_2$. The total reaction volume for each well was 200 μL. The concentrations of variable substrates can be found in the relevant figure legends. UGT72B1 reactions consisted of UGT72B1, 100 μM esculetin, and varying concentrations of UDP-glucose. RmlA/UGT72B1 reactions consisted of RmlA, UGT72B1, 100 μM esculetin, and 500 μM of either UTP or glucose-1-phosphate, with the other substrate's concentration varied. PGM/RmlA/UGT72B1 reactions consisted of PGM, RmlA, UGT72B1, 100 μM esculetin, 500 μM UTP, and varying concentrations of glucose-6-phosphate. A6PR/PGM/RmlA/UGT72B1 reactions consisted of A6PR, PGM, RmlA, UGT72B1, 100 μM esculetin, 500 μM UTP, and 500 μM of either $NADP^+$ or sorbitol-6-phosphate, with the other substrate's concentration varied. Fluorescence emission at 454 nm following excitation at 336 nm was monitored over time on a BioTek Hybrid Synergy 4 plate reader (Winooski, VT). All reactions were prepared in triplicate, and each enzymatic reaction was also tested without the initial enzyme (data not shown).

FACS Assay

W/o/w samples were prepared and analyzed by FACS as described in Chapter 2 and the conditions described above. In all reactions, 5 μg of each enzyme was used. Each enzymatic reaction was also tested without the initial enzyme (data not shown). The A6PR/PGM/RmlA/UGT72B1 reaction was also tested with 1 mM NADPH (data not shown).

CHAPTER 4

Mathematical Models for Directed Evolution Library Construction and Enrichment

4.1. Introduction

Directed evolution requires strategic choices regarding mutagenesis and screening or selection to minimize cost and maximize the chance of success. In many ways the increasing effectiveness of directed evolution has been driven by advancements that are conceptual rather than technological. These concepts include fitness landscapes of sequence space,[16] genotype-phenotype links,[65] neutral fitness networks,[78] and various concepts related to library quality, which is the frequency of 'hits' within a library. The most pervasive of these concepts help to guide researchers in academia and industry and serve as the basis for decision making in directed evolution. Mathematical modelling of library quality is a common method for comparing directed evolution strategies. However, conceptual flaws in directed evolution, often widely well-regarded artifacts of ignorance, result in flaws in our mathematical models that ultimately support suboptimal directed evolution strategy.

Currently, most mutant library preparation is conducted in-house by academic and industrial directed evolution laboratories. However, commercial mutagenesis products are becoming increasingly sophisticated and affordable, therefore mathematical models of directed evolution strategies which guide commercial product development and consumer choice will likely increase in importance. Already these commercial products have significant advantages over homemade libraries, such as higher transformation cloning, mutation rates that are consumer-defined at each amino acid position, and trinucleotide building block libraries with no rare or stop codons and near-perfect codon diversity.[79] Other commercial options are likely

to arise, such as libraries that combine multisite saturation with computationally designed noncontiguous chimeragenesis.[33] Preliminary library enrichment through screening or selection may become part of commercial library preparation. Mathematical models and conceptual advances for library design and these preliminary screening strategies will likely help determine which commercial mutant library products become available in the future.

Error-prone polymerase chain reaction (epPCR) is still the leading method for diversity generation in directed evolution due to its technical simplicity. Average epPCR mutation rates can be aimed for through careful manipulation of amplification reaction conditions. Currently, leading experts in the field suggest mutation rates be kept quite low, approximately 1 to 2 average nucleotide mutations per gene, because at these mutation rates mutants with single amino acid mutations are most common.[80] Mutant enzymes with multiple amino acid mutations are viewed as inferior for two reasons: the "numbers problem" and the largely deleterious effects of mutagenesis.

The "numbers problem" is the result of combinatorial expansion of sequence space.[34] While a protein N amino acids long has 19N possible single amino acid mutants, it has $19^2((N^2-N))/2)$ double amino acid mutants, $19^3((N^3-3N^2+2N)/6)$ triple amino acid mutants, etc. For a protein of 300 amino acids, that corresponds to 5700 single amino acid mutations, 1.6×10^7 double amino acid mutants, and 9.2×10^{10} triple amino acid mutants.[81] Since most directed evolution campaigns are limited to screening fewer than 10^4 mutants per round, it is rarely possible to screen a large fraction of the theoretical double, triple, or greater mutants.[34,80] Therefore multiple rounds of low mutation rate mutagenesis and screening are preferred, with incremental improvements found at each step.

A substantial fraction of individual amino acid mutations, approximately 30% for most proteins, are deleterious to function, folding, or expression.[82] In practice, this is calculated from the nucleotide mutation frequency and activity data, so this value includes transcripts produced by epPCR which can not be translated into functional variants due to frameshifts or the introduction of premature stop codons. Although stabilizing mutations can "rescue" the function of enzymes destabilized by some deleterious mutations, such broadly stabilizing mutations are considerably rarer.[83] Increasing mutation rates therefore results in an exponential decline in the number of functional variants. The assumption which prevails in the field is that library quality declines proportionally.[80]

In this chapter, I attempt to correct the conceptual framework for choosing optimal mutation rates for directed evolution through mathematical modelling. I demonstrate the relative library quality curves for epPCR libraries, the value of purifying selections, and the value of larger rather than smaller libraries. I also compare multi-step screening strategies, concluding that the use of generic purifying selections for neutral drift followed by functional screening is the best broadly applicable screening strategy currently available.

4.2 Principles of the Fitness Landscape of Sequence Space

In order to understand how to exploit enzyme evolvability through directed evolution, it is helpful to understand how individual enzymes evolve in nature. Paramount to this understanding has been the concept of the fitness landscape of sequence space.[16] A fitness landscape of an enzyme of interest is a visual metaphor for its potential evolutionary paths. All possible genotypes that could result from mutagenesis of the template gene are arranged in two

dimensional space, with more similarly genotypes closer together than less similar genotypes. The height of this landscape represents the fitness value of each genotype. In the case of natural evolution, these values represent whole-organism reproductive potential and change over time due to changes in the environment and mutations on other genes. For directed evolution, fitness is user-defined, often by a single protein property, and can remain static indefinitely.

Directed evolution is modeled after Darwinian Evolution, that is, iterative single uphill steps along the fitness landscape. In nature, evolution escapes these local maxima through periods of neutral drift, in which non-functional variants are eliminated but functional variants are retained. The population spreads across the neutral landscape until an uphill step can be found.[84] Without similar strategies, directed evolution will always inevitably reach a mutant at a local maximum fitness.[16]

Directed evolution strategies which can overcome local fitness maxima will ultimately provide superior catalysts to those which can not.[17] Neutral ridges connecting local maxima to areas with further uphill climbs are readily visualized in three dimensions, and the true multidimensionality of fitness landscapes suggests that most if not all enzymes at local maxima may be improved through the exploration of neutral sequence space. This has led to a strategy known as neutral drift (discussed in greater detail in section 4.5), which mimics the natural evolutionary process of the same name.[83,85]

4.3 The Numbers Problem and Oversampling

Concern over the "numbers problem" of directed evolution has led to misuse of the concept of 'library coverage,' which is improved through intentional 'oversampling.'[34]

Reetz et al. provide Equation 1 for calculating library coverage, P_i, from the oversampling factor, T/V, with T being the number of colonies screened and V being the library size.[34]

$$\text{Equation 1: } \mathrm{T/V} = -\ln(1 - Pi)$$

These calculations conclude that 3-fold oversampling ensures that 95% of the sequences have been screened (Fig. 43).

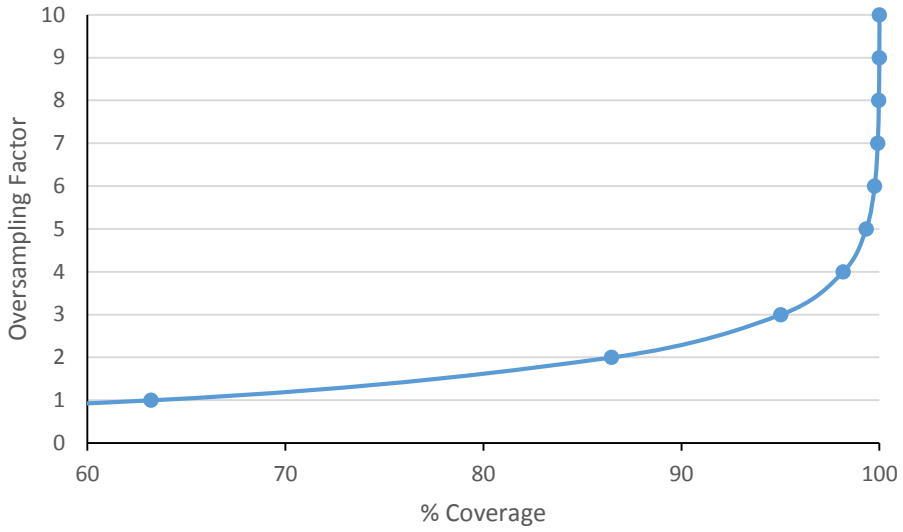

Figure 43. Oversampling factor, the ratio of colonies screened to theoretical library size, yields diminishing returns on library coverage. The line is a plot of Equation 1.

When screening is simple and inexpensive and library construction is complex and expensive, it makes sense to dig as deeply as possible into each library. In such a case, maximizing library coverage would be useful. Additionally, when targeted mutagenesis and

screening are used to study structure-function relationships, oversampling also makes sense. However, this concept is often used to guide epPCR mutation rates so that the resulting library can be more thoroughly screened.[38] Since the medium-throughput screens typically used for directed evolution are incapable of oversampling all two amino acid mutation mutants, researchers aim for libraries composed mostly of single amino acid mutation mutants.[80]

Intentionally screening the same sequence more than once is costly and unproductive. Consider two libraries, A and B, made up of one thousand sequences and one million sequences, respectively. In each case, a 'hit' sequence is present at a frequency of 0.1%. The library quality, therefore, is the same. The small size of library A actually hinders the discovery of this hit (Fig. 44).

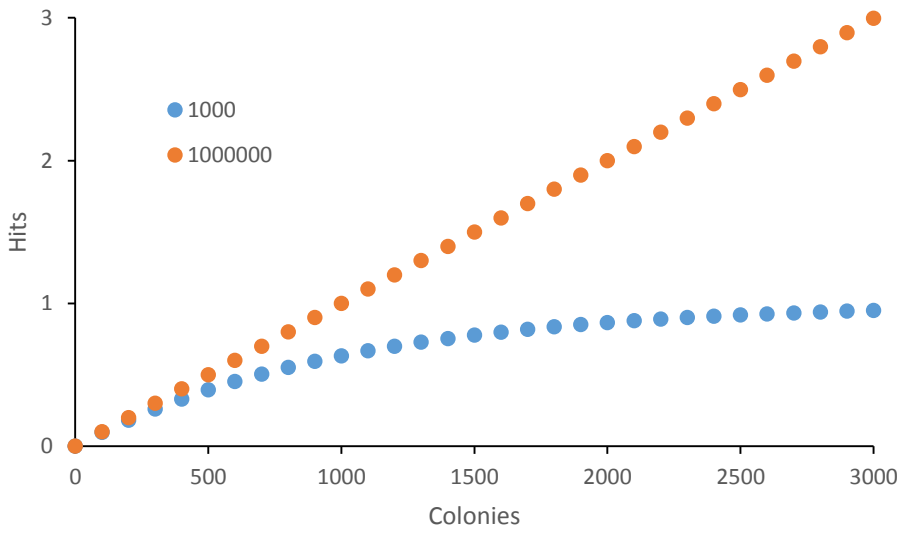

Figure 44. Number of hits found in hypothetical libraries A (1,000 members) and B (1,000,000 members) with 0.1% of mutations providing enhanced activity. As the number of colonies screened increases, the oversampling of library A leads to decreased efficiency.

After 3000 colonies, 950 unique sequences from library A have been screened versus 2996 unique sequences from library B. This corresponds to 0.95 and 3.00 average 'hits,' respectively. As a rule, severe oversampling of mutant libraries should be avoided, when possible. Library size bottlenecks, such as cloning of the mutant library, should be carefully avoided to ensure as many sequences are cloned as possible. Truly optimal mutation rates (calculated in 4.4), that is, those chosen to maximize the frequency of hits within the population, provide sufficient diversity so that reduced efficiency from oversampling is rarely a problem.

An illustrative example of the flaw of oversampling is the successful directed evolution campaign by Evans et al. for non-ribosomal peptide synthetase adenylation domain mutants capable of producing unnatural andrimid derivatives.[86] While ten residues are known to be responsible for substrate selectivity of these adenylation domains, the authors chose to limit library size by mutating only three of these positions through partial saturation which favored non-polar residues. Limiting the mutagenesis to three residues allowed them to achieve 10-fold oversampling, and over 99.99% coverage, of their theoretical 1404 mutants with a multiplexed mass spectrometry screen. The best mutant incorporated the non-natural substrates isoleucine and leucine into adrimid variants at 123.3% and 24.8% yield relative to wild-type production of andrimid (corresponding to valine incorporation). However, this mutant had an additional mutation at a fourth residue, presumably an artifact from PCR, encoding an arginine to lysine mutation. The authors suggest that the extra mutation may be neutral; however, if this were the case one would expect the triple mutant without that extra mutation to have been found as well, due to the 10-fold oversampling. Had the authors chosen to increase their library

size to include other residues and not restricted themselves to such a limited codon alphabet, perhaps a much larger number of interesting variants would have been isolated. If their library size were ten-fold larger, screening the same number of colonies would have examined approximately 7469 (532%) more mutant sequences.

Nov et al. pointed out that the difference between 95% coverage of a library and 95% chance to find one of the best mutants within the library is immense.[87] When considering the relative rarity of some amino acid combinations, the number of colonies which must be screened to ensure 95% coverage of all possible amino acid sequences is considerably higher than when each combination is assumed to be present in equal frequency. Simultaneous saturating two residues requires 8,128 colonies be screened to ensure 95% coverage of the 400 amino acid sequences, while only 2,130 colonies must be screened for a 95% probability of isolating the best mutant, and only 875 colonies must be screened for a 95% probability of isolating one of the best two mutants. While these calculations are useful, they should not be used to plan the mutation rate of an epPCR library or the number of saturation sites in a multi-site saturation library. Purposefully limiting sequence space so that the throughput of a screen can find one of the top mutants within that sequence space is counterproductive and can actually decrease library quality.

Since small library size leads to inefficiency through oversampling when screening throughput is high enough, it might seem appropriate to increase mutation rates specifically to avoid oversampling. In practice, for optimal mutation rates, theoretical library size is typically limited by the transformation yield of the gene library into the host organism rather than the

mutation rate. Therefore theoretical library size simply does not need to be considered when determining optimal mutation rates for epPCR libraries.

4.4 Optimal Mutation Rates

The fraction of amino acid mutations tolerated by an enzyme is its mutational robustness, ρ. For all mutants with *n* amino acid mutations, the fraction which are properly expressed with retention of their structural integrity is equal to ρ^n.[88,89] What, therefore, is the value of increasing mutation rates if the fraction of functional sequences declines exponentially? The leading concept in the literature is that mutation rate should be as low as possible while avoiding too many wild-type sequences, approximately 1-2 nucleotide mutations per gene.[80] An exception to this is when an ultra-high throughput screen or selection is conducted, in which case increased library diversity leads to reduced oversampling and 3 nucleotide mutations per gene is widely considered acceptable.[36]

Bershtein et al. point out a flaw in that concept.[35] Each additional amino acid mutation on a functional mutant might be "the one", and therefore the relative value of functional mutants increases linearly with the amino acid mutation rate. For example, a mutant with two amino acid mutations is less likely to be functional than a mutant with one amino acid mutation, but if functional the two amino acid mutation has twice the chance of being a hit. Therefore the quality, Q, of the average mutant with *n* amino acid mutations can be described according to Equation 2, adapted from Bershtein et al.[35]

Equation 2: $Q_n = n\rho^n$

Clearly, Q depends heavily on mutational robustness. Although mutational robustness varies, it is often found to be between 0.6 and 0.7.[82] Unrelated proteins sharing similar mutational robustness can be explained by the lack of evolutionary pressure for stability above a minimal threshold and the tendency of neutral drift to remove additional stability over time due to the high frequency of slightly deleterious mutations.[90] The relative quality of mutants according to their number of mutations can be predicted for the expected range of mutational robustness values (Fig 45).

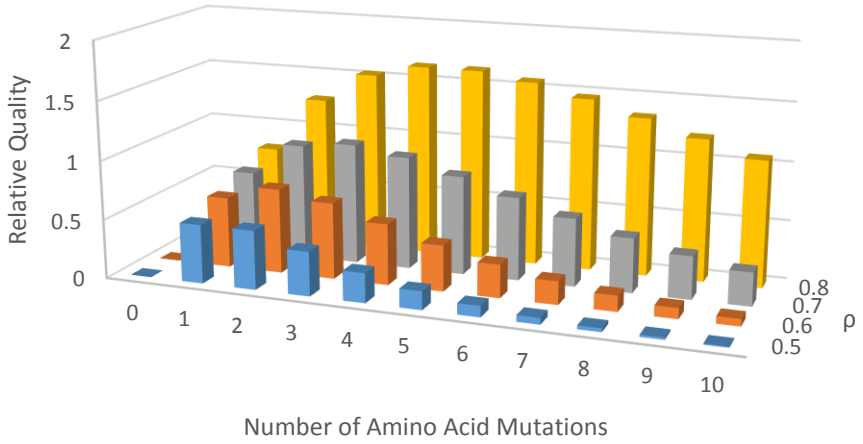

Figure 45. Relative quality of mutants, from Equation 2, within the expected range of mutational robustness values. In general high mutational robustness improves relative quality, particularly for mutants more heavily mutated than is customary for epPCR libraries.

Optimal epPCR library mutation rates should not be determined from only Equation 2 and Fig. 45. EpPCR is incapable of creating a tightly defined mutation rate, and library quality is determined by the relative frequency of each mutation level as well as the relative quality of these levels. Drummond et al. describe a model for calculating the standard deviation of library

mutation rate in epPCR.[36] The calculation depends on the mean mutation rate, μ, and the mutagenicity, *r*, of the epPCR in substitutions per duplication.

Equation 3: $\sigma = \left(\mu(1 + r)\right)^{\frac{1}{2}}$

The mutagenicity of epPCR with the most common commercial kit (GeneMorph II) is approximately 1 nucleotide substitution per kb per duplication according to the manufacturer. Therefore the standard deviation of the mutation rate increases according to gene size, although this is insignificant except for extremely large and extremely small genes (data not shown). Since 74% of nucleotide mutations encode a new amino acid,[91] we can plot the distribution of mutants with various amino acid mutations at defined average nucleotide mutation rates (Fig. 46).

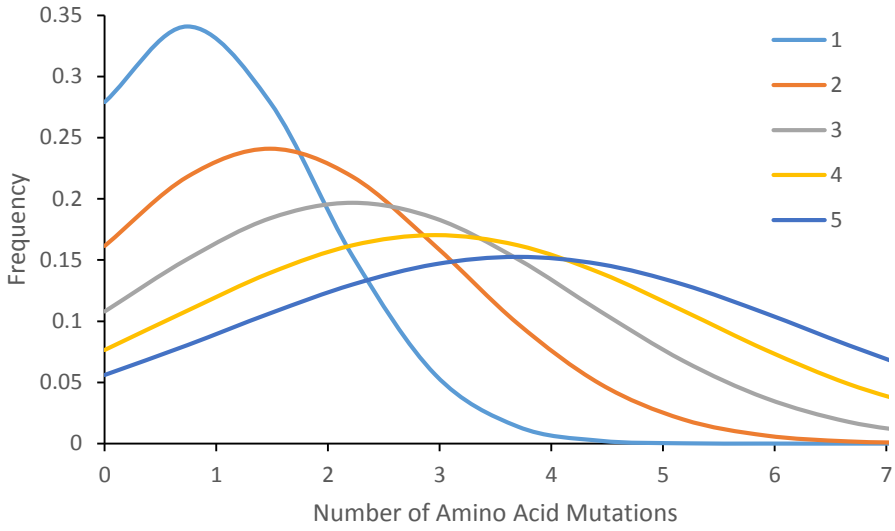

Figure 46. The lines are derived from Equation 3, and they represent the frequency of various amino acid mutation mutants within mutant libraries with varying mean nucleotide mutation rates.

With relative quality values for each amino acid mutation (Equation 2 and Fig. 45) and the frequency values, *f*, of each amino acid mutation at a particular mean mutation rate and gene length (Equation 3 and Fig. 46), we can calculate library quality at various nucleotide mutation rates (Equation 4 and Fig. 47).

Equation 4: $Q_L = \sum_{n=0}^{\infty} Q_n f_n$

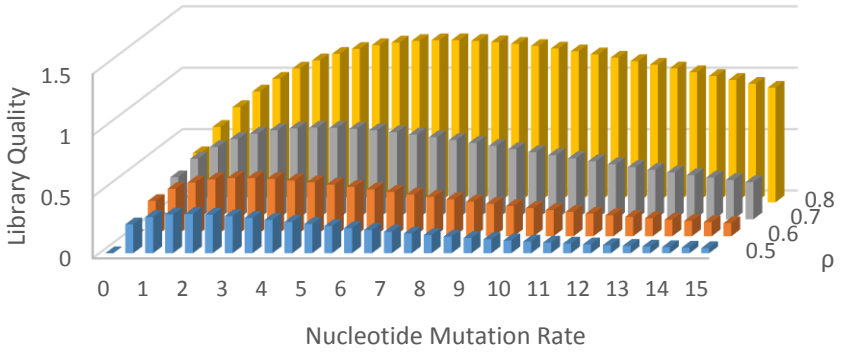

Figure 47. Simulated library quality for a 1.5kb gene at various nucleotide mutation rates and common mutational robustness values, from Equation 4.

According to these calculations, 2 to 5 nucleotide mutations is optimal for most enzymes. Interestingly, however, the slope of the decline in library quality is quite shallow. This suggests that the cost is minimal for mutation rates that are "too high" by traditional wisdom. At such high mutation rates, the theoretical library size is limited by transformation efficiency and oversampling is likely minimal and can be ignored when calculating library quality.

It should be noted that all of these calculations ignore epistatic mutations, which have a different effect when combined with other mutations than without. Epistatic interactions are rare enough to be ignored when calculating library quality, although their increased presence in libraries with higher mutation rates should be noted.

4.5 Purifying Selections and Neutral Drift

FACS screens like those described and developed in Chapters 2 and 3 can be used as purifying selections for neutral drift. The neutral drift strategy uses a screen or selection to collect mutants with some measurable enzymatic activity above a user-defined threshold.[78,83,85] Purifying selections don't enrich populations specifically for improved properties but rather for retention of existing properties, creating genetically polymorphic libraries that can be screened for new behaviors in a secondary round of screening. Purging a library of poorly expressed, unstable, or inactive variants results in a new library without the exponentially deleterious effects of increasing mutation rates.

Although considerable progress must be made to achieve a truly universal *in vitro* FACS directed evolution screening platform, a number of generic purifying selections are available which remove variants that are not expressed or folded properly. The most notable of these tools are a FACS screen of GFP-fusion proteins[92] and a selection on chloramphenicol for CAT-fusion proteins.[93] Approximately 90 to 95% of mutations that do not deleterious affect expression, folding, or stability are not deleterious to catalysis.[94] Generic purifying selections remove variants with these non-catalytic deleterious properties, yielding enriched libraries with mutational robustness values of 0.9 to 0.95. For such libraries, mutants with high mutation rates actually have the highest relative value (Fig. 48).

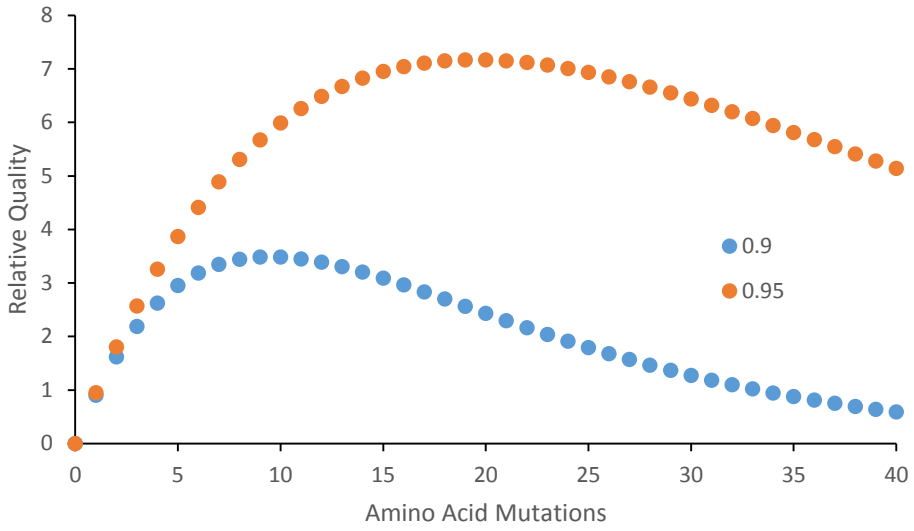

Figure 48. Relative quality of mutants isolated from a purifying selection within the expected range of mutational robustness values, from Equation 2.

When PCR distribution of mutations is taken into account, optimal nucleotide mutation rate can be calculated (Equation 4 and Fig 49).

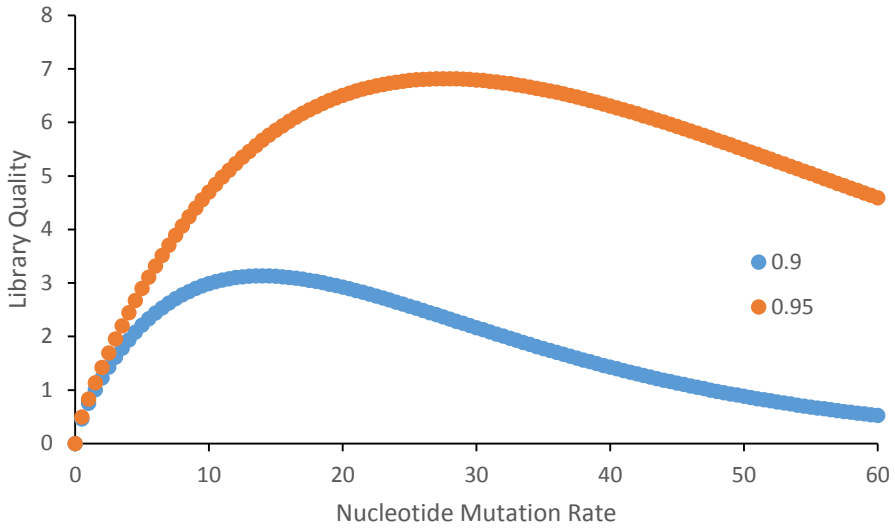

Figure 49. Simulated library quality from Equation 4 for a 1.5 kb gene at various nucleotide mutation rates and common mutational robustness values after purifying selection.

The optimal nucleotide mutation rate for epPCR when using a purifying selection appears to be between 10 and 30 nucleotide mutations per gene. However, as the nucleotide mutation rate rises to such high levels, the size of the enriched library decreases significantly, and this risks efficiency loss from oversampling in the subsequent screen. The fraction of the original library size remaining after purifying selection, F, can be calculated via Equation 5, where ρ_0 is the mutational robustness of the template gene, *m* is the nucleotide mutation rate, and the 0.74 multiplier is the amino acid mutation per nucleotide mutation conversion.

Equation 5: $F = \rho_0{}^{.74m}$

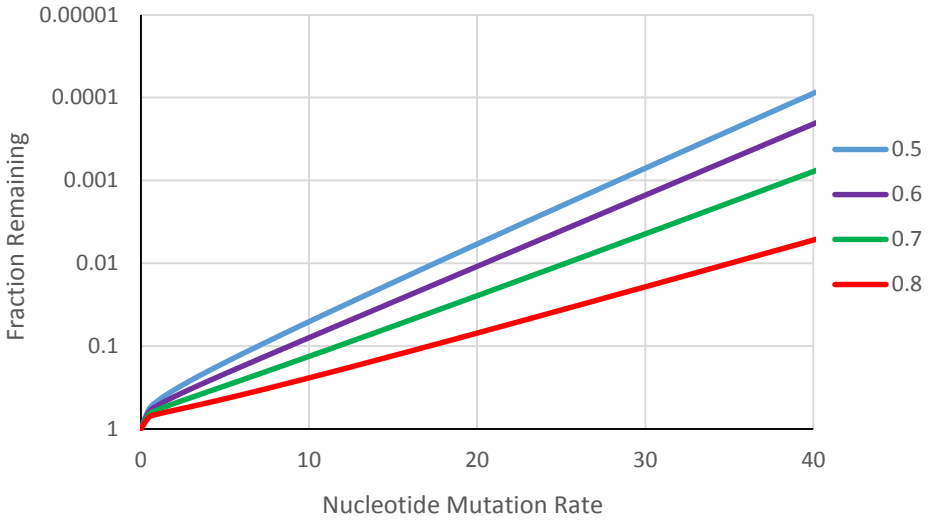

Figure 50. Fraction of mutants remaining within a library after a purifying selection for expected values of initial mutational robustness. Lines are plots of Equation 5.

Purifying a library with 30 initial average nucleotide mutations per gene will typically yield between 70 and 2,000 sequences per 100,000 initial sequences, while a library with 10 initial average nucleotide mutations will yield between 5,000 and 24,000 sequences per 100,000 initial sequences (Fig. 50). A carelessly chosen mutation rate might lead to a population bottleneck and large inefficiencies due to oversampling in subsequent screening. Considering the previous two figures, aiming for approximately 15 average nucleotide mutations per gene, which should preserve at least 1% of the original library size, can be considered the optimal strategy for purifying selections.

The library quality values listed in Fig. 46 and Fig. 49 can be compared to determine the value of purifying selections as generic tools enabling more efficient directed evolution. An initial purifying selection from a library with 15 average nucleotide mutations can improve

hit rates in subsequent screening approximately 10- to 20- fold, 6- to 12- fold, 4- to 8- fold, and 2.5- to 5- fold for enzymes with mutational robustness values of 0.5, 0.6, 0.7, and 0.8 respectively compared to libraries with 4 average nucleotide mutations per gene. Due to the relatively low cost of these generic selections, more widespread use of these tools would lead to large improvements in efficiency for nearly all directed evolution projects.

4.6 Potentially Beneficial Mutations, Evolvability, and the Arnold Strategy

Directed evolution not only exploits the knowledge of evolutionary biology, it also contributes to it. One example of this is the development of our understanding of epistatic mutations, combinations of mutations which individually provide no advantage but, when combined, provide increased fitness. The most common mechanism of apparent mutational epistasis is now believed to be the combination of a potentially beneficial but destabilizing mutation with a neutral or deleterious stabilizing mutation.[17] Importantly, a relatively large number of mutations may provide the necessary improved stability for the potentially beneficial mutation to be tolerated.

Not enough is known about potentially beneficial mutations and their access through the fitness landscape of sequence space. How common are potentially beneficial mutations compared to structurally tolerable beneficial mutations? Active site mutations are known to be predominantly destabilizing, suggesting that these largely unexploited potentially beneficial mutation are relatively common. How common are stabilizing mutations which allow potentially beneficial mutations to be tolerated? Perhaps fairly common as well, since many of the early successes of directed evolution involved improving protein stability.[95] There may be

many uphill paths on the fitness landscape which each connect to the template sequence through multiple single step neutral ridges.

These neutral ridges are particularly mutationally robust enzymes that can access more single-step uphill routes on the fitness landscape than the template sequence. The Arnold group has begun directed evolution projects by first screening for thermostability in hopes of improving initial mutational robustness.[96] The models for epPCR library quality in the preceding sections clearly show the improvement in library quality with increasing mutational robustness, but these models do not account for the existence of potentially beneficial mutations. By reaching a more mutationally robust enzyme through directed evolution, the evolvability, ε, is increased.[94] Evolvability is the percentage of single nucleotide mutations that encode a beneficial amino acid mutation tolerated by the enzyme. Evolvability was left out of the library quality calculations because it is typically an unknown constant and it doesn't vary with mutation rate. Therefore, for the purpose of determining optimal mutation rates in epPCR libraries, evolvability can be ignored. However, the directed evolution strategy of evolving thermostability prior to further mutagenesis and functional screening makes evolvability quite important, because evolvability is almost certainly improved in the more thermostable mutant. Without empirical evidence for how evolvability changes with evolved stability, estimating the efficiency of this strategy are difficult. Qualitatively, however, the library quality is improved proportionally to the improvement in evolvability as well as by the mutational robustness dependent improvement modeled in Fig 47.

How does this strategy compare to the neutral drift strategy explained in Section 4.5? It's probably less efficient. The estimates for improvement in library quality from purifying

selections are significantly higher than improvements in library quality as mutational robustness increases (Fig. 47), and are unlikely to be offset by improved evolvability. Additionally, the strategy of improving thermostability is more difficult, as it includes thermostability and activity analysis of first round hits and requires a second round of mutagenesis.

4.7 Multi-site Saturation Mutagenesis

Multi-site saturation mutagenesis is an increasingly popular alternative to epPCR thanks to commercialization of robust kits, improved software for protein sequence and structure analysis, and wide-spread acceptance of the "numbers problem" in directed evolution.[34] The quality of a multi-site saturation library can be predicted through Equation 2 using local mutational robustness, the average mutational robustness across all sites to be saturated. This explains the risk of multi-site saturation mutagenesis, since deleterious mutations are much more common within the active site of an enzyme than on its surface.

Multi-site saturation mutagenesis is therefore only optimal for mutating residues from variable regions within a family of highly similar sequences. Multiple sequence alignment software can be used to find these variable regions, and in silico substrate docking studies with computationally or crystallographically derived structural models can help determine which of these variable regions are most likely to bring about the desired chance in activity. The local mutational robustness of the individual chosen residues can be tailored by altering the nucleotide alphabet to more highly represent mutations that are tolerated according to the

multiple sequence alignments.[30,86] From these efforts, an average local mutational robustness can be estimated and the ideal number of residues for saturation chosen according to Fig 47.

Altering the nucleotide alphabet in multi-site saturation libraries can lead to a major improvement in screening efficiency. Reetz et al. screened 5000 colonies each from epoxide hydrolase libraries saturated at three residues with either NNK (N = all four bases and K = guanine and thymine) or NDT (D = adenine, guanine, and thymine and T = thymine), generating libraries with theoretical sizes of 32,768 and 1,728 codons respectively.[34] 38 hits were found from the NNK library while 511 hits were found from the NDT library. More restrictive screening led to 10 hits for NNK and 180 hits for NDT. The authors mistakenly credits smaller library size and higher oversampling to the increased hit rate. However, properties of the NDT alphabet can be used to partially explain this phenomenon.

NNK libraries encode 20 amino acids at each position with 32 codon sequences while NT libraries encode 12 amino acids at each position with 12 codon sequences. Therefore the NNK library contains 32,768 nucleotide sequences encoding 8,000 distinct amino acid sequences, while the 1,728 nucleotide sequences of the NDT library encode 1,728 distinct amino acid sequences. Based on Poisson calculations (Equation 1), we can estimate that 14.15% of the NNK library, approximately 1132 amino acid sequences, were screened. In contrast, 94.5% of the NDT library, approximately 1632 amino acid sequences, were screened. Clearly, the severe consequences of oversampling on screening efficiency are nearly offset by the redundancy of the NNK alphabet. The failure of this calculation to explain the 13- and 18-fold higher hit rate of NDT libraries in the screen is extremely interesting. It suggests an additional inherent efficiency of screening NDT libraries compared to NNK libraries. The 12

amino acids at each position in an NDT library are diverse and representative of most properties of the full 20 amino acids. If we assume that these 12 represent all of the properties of the full 20 amino acids, we can imagine that the 32,768 nucleotide sequences of the NNK library encode the same 1,728 amino acid sequences as the NDT library. In such an assumption, only 181 sequences were screened from the NNK library, approximately 9-fold fewer than in the NDT library.

The final explanation for the improved quality of the NDT library is that the mutational robustness at each site is significantly higher with NDT than with NNK. Some of this is predictable and universal, for example 9% of the NNK library contained an undesired stop codon. However, this is also evidence of the value of bioinformatics, which the researchers used to guide their choice in nucleotide alphabets. From just the wild-type and nine mutants that were characterized as most active from the NDT library and sequenced, eight distinct residues are found in first position, five in the second position, and seven in the third position. This high diversity amongst such a small number of sequences suggests that the mutational robustness of the NDT library at this position is higher than would be expected for active site residues.

Other nucleotide alphabets encode other restricted amino acid alphabets. The choice of the best alphabet to maximize mutational robustness and reduce oversampling should consistently yield improvements in directed evolution efficiency. Of particular note is the "22c" library strategy which uses a primer mixture of NDT (12 unique codons), VHG (9 unique codons), and TGG.[97] This mixture encodes all 20 amino acids in just 22 codons with no stop codons. However, while this strategy is easily applied when the sites to be saturated can be

encoded on separate primers, it is impractical for nearby mutations which must be encoded on the same primer.

4.8 Conclusions and Future Outlook

Efficient directed evolution from epPCR libraries requires properly chosen mutation rates. However, the traditional wisdom guiding the choice of mutation rate is in some cases, we believe, misleading. Mutation rates should be chosen to balance the tradeoff between phenotypic diversity and functionality that gives the highest frequency of improved variants. Oversampling should be avoided through proper library construction. Purifying selections offer a reliable way to improve the efficiency of most directed evolution projects.

A number of new ideas arose through the investigation into these concepts. For example, it should be possible to use predictive modelling software such as Rosetta to estimate the mutational robustness of each individual residue within an enzyme's active site.[98] Importantly, the mutational robustness values can be determined with various saturation nucleotide alphabets which only produce a subset of the 20 natural residues. When used in conjunction with a substrate docking model to identify active site residues which don't directly participate in the catalytic mechanism, optimal multi-site saturation mutagenesis can be performed in the absence of data from multiple sequence alignments. The user would be informed on the optimal number and location of residues to saturate and the optimal nucleotide alphabet for each position.

Another interesting strategy is iterative use of high rate mutagenesis and purifying selection to create extremely phenotypically diverse libraries. Notably, purifying selections

have been used iteratively with low rate mutagenesis and the results were promising.[85] Using the strategy of employing high mutagenic rates and purifying selections, it may be possible to achieve success in more ambitious directed evolution projects than have been previously reported.

Perhaps the greatest potential of the neutral drift strategy arises from the possibility of using Phage Assisted Continuous Evolution (PACE) as a generic purifying selection.[99] PACE is arguably the strongest example of a directed evolution platform to date, allowing simultaneous and continuous rounds of mutagenesis and selection across enormous libraries, but requires that the phenotype of interest be coupled *in vivo* to phage propagation. PACE was first used to select for RNA Polymerase variants capable of transcribing a protein essential to phage propagation regulated by a non-natural promoter sequence. However, it should be possible to select for proper expression and folding of a library of interest by fusing the library directly to an essential phage propagation protein.

Empirically derived mutational robustness values for specific residues of an enzyme would be extremely powerful information for directed evolution campaigns. Since mutually robust residues surrounding the active site are likely important for substrate recognition, this information might greatly enhance our ability to reengineer substrate specificity for applications in industrial biocatalysis and combinatorial biosynthesis of natural products. ProSAR is a prominent method for determining sequence-function relationships of enzymes that could be extremely effective at finding residue specific mutational robustness values.[32] ProSAR uses a statistical analysis of individual mutations from sequence and activity data to determine mutations which benefit the desired activity. Applying high rate random

mutagenesis and purifying selections iteratively would produce an extremely high mutation rate library purged of mutations that deleteriously affect folding, expression, or stability. ProSAR analysis of sequencing data from these surviving clones would provide a comprehensive list of mutations that are never tolerated, sometimes tolerated, or always tolerated. The cost of extensive sequencing is greatly reduced by the high mutation rate of the average mutant, since the identity of each mutation is an important data point. The cost of extensive screening is greatly reduced by the affordability of generic purifying selections. The empirically derived mutational robustness of specific residues would be valuable for reengineering all enzymes with high homology to the studied enzyme.

4.9. Experimental Section

All calculations were performed in Microsoft Excel. Amino acid mutation rates were set at 0.74 amino acid mutations per nucleotide mutation (real values will vary depending on the nucleotide sequence). Likewise, bias of error-prone DNA polymerases towards transitions over transversions (and its effect on the nucleotide mutation rates to amino acid mutation rates) was ignored. Library quality calculations included mutants with up to 100 nucleotide mutations.

CHAPTER 5

Neutral Drift Directed Evolution of UGT72B1

5.1. Introduction

Neutral drift purges inactive variants from high error rate mutant libraries, yielding a highly active and genetically polymorphic library with higher hit rates in subsequent screening than a naïve library. After being subjected to neutral drift, in which the purifying selection demanded activity toward a single substrate recognized by the wild-type, cytochrome P450 BM3 mutants demonstrated altered activities toward five other substrates.[100] 300 mutants of serum paraoxonase PON1 with natural lactonase activity similar to the wild-type were examined for activity with other known substrates, and over half of these mutants displayed altered phenotypes.[78] It has also been demonstrated that active cytochrome P450 mutants from larger (higher mutation rate) libraries display higher mutational robustness than those found in smaller libraries.[88]

Together, these results indicate that neutral drift is an impressive directed evolution strategy. Additionally, it is one best served by an ultra-high throughput screen or selection, because higher throughput permits higher mutation rates without the risk of reducing library size to the point where oversampling in a secondary screen becomes an issue. It has been noted that directly screening for an activity that is not detected by the wild-type enzyme is very unlikely to yield active mutants. Peisojovich and Tawfik quip that while the first rule of directed evolution is "you get what you screen for," the second rule might be "you should screen for what is already there."[101] Whether neutral drift enrichment for an existing activity can overcome this barrier is the focus of this Chapter.

In order to test the capability of neutral drift in this regard, UGT72B1 was subjected to neutral drift and the enriched library was tested for activity against a panel of UDP-sugars not recognized by the wild-type enzyme. The UGT72B1 mutants subjected to neutral drift were subsequently screened in lysates in 96-well microtiter plates for enhance catalytic activity.

5.2. Results and Discussion

5.2.1 Mutagenesis

A high-mutagenesis rate epPCR library of UGT72B1 mutants was prepared using the Genemorph II random mutagenesis kit and cloned into the pETduet-*sfGFP* vector. The library size was calculated to be greater than 4 x 10^5 based on a plated sample of the transformation. The mutation rate of this library was found to be 7.0 ± 3.7 nucleotide mutations per gene, corresponding to a calculated average amino acid mutation rate of 5.2 ± 2.7 mutations per mutant. Although 17 mutants were sequenced, 9 were found to consist of an unusual cloning artifact in which the vector religated across supposedly incompatible restriction sites (*Nco*I and *Hin*dIII). Efforts to remove these sequences from the population through electrophoresis were unsuccessful and ultimately deemed unnecessary since they would be removed through neutral drift.

5.2.2 Neutral Drift of UGT72B1 epPCR Library

W/o emulsions were generated using *E. coli* BL21(DE3) cells harboring the UGT72B1 mutant library, and the aqueous phase included 1 mM esculetin and 2.5 mM UDP-glucose. The high donor substrate concentration was chosen so that mutants with relaxed K_M toward UDP-

glucose would not be disadvantaged. These w/o emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 45 minutes at 37 °C to allow the enzymatic reaction to proceed. The droplets were reemulsified to generate the w/o/w emulsions and sorted by FACS as described in Chapter 2. 10.3% of droplets appeared to contain an active mutant, corresponding to 22% of the properly cloned mutants (Fig. 51). 78% inactivation of a library with an average mutation rate of 5.2 amino acid mutations per mutant corresponds to a mutational robustness value of 0.75, well within the expected range. Subsequently, 5.2 $x10^6$ droplets were sorted and 5.3 $x10^5$ droplets were collected. The DNA was extracted from the sorted droplets, PCR amplified, and recloned into the pETduet-*sfGFP* vector. DNA sequencing random colonies revealed that 7 out of 10 sequenced colonies contained the unusual cloning artifact, but the total number of transformants was high enough ($>10^6$) to ensure minimal loss of library diversity. As noted during initial library construction, neutral drift removes the empty vector from the population during either the FACS sorting or gene amplification steps. W/o/w emulsions were prepared in the same manner as the first round and were subsequently subjected to another round of purifying selection. In this second round of purifying selection, approximately 1.2 $x10^7$ droplets were sorted and 1.1 $x10^6$ droplets were collected, corresponding to half of the active population. The distribution of the active fluorescent population (Fig. 52) was entirely different from that of the naïve mutant library (Fig. 51). This altered distribution indicates successful enrichment despite the large negative population due to the cloning artifact.

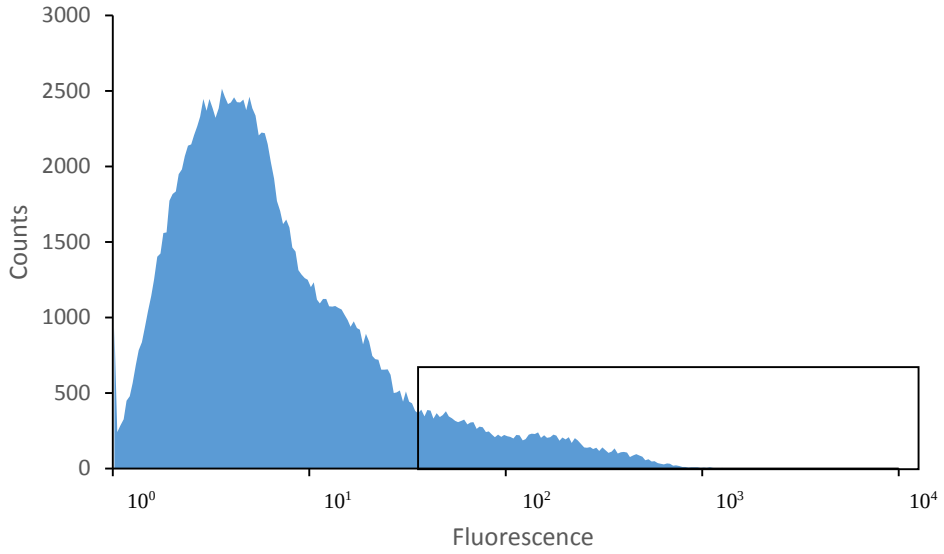

Figure 51. FACS histogram of the UGT72B1 library during the first purifying selection. The droplets within the gated area were collected.

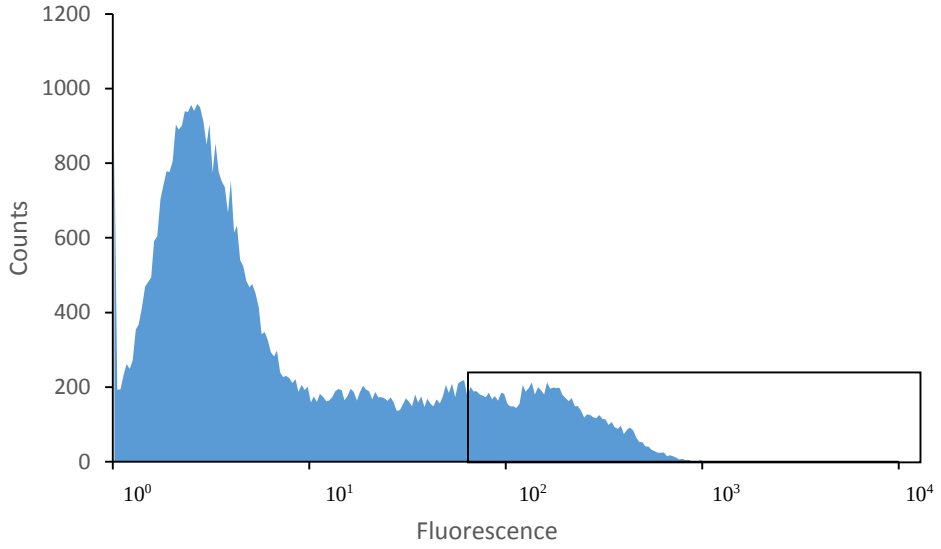

Figure 52. FACS histogram of the UGT72B1 library during the second purifying selection. The droplets within the gated area were collected.

5.2.3 Secondary Screening of the Enriched UGT72B1 Library

The twice purified mutant DNA library was extracted from the sorted droplets, amplified by PCR, and cloned into pETduet-*sfGFP*. W/o emulsions were generated from *E. coli* BL21(DE3) cells harboring the enriched mutant library supplemented with 1 mM esculetin, and 2.5 mM each UDP-N-acetylglucosamine, UDP-N-acetylgalactosamine, and UDP-glucuronic acid. These w/o emulsions were frozen in an ethanol/CO_2 bath, thawed at room temperature, and incubated for 60 minutes at 37 °C to allow the enzymatic reaction to proceed. The droplets were reemulsified to generate the w/o/w emulsions and sorted by FACS as described in Chapter 2. Disappointingly, FACS analysis failed to detect a significantly active population compared to the wild-type FACS histogram (Fig. 53).

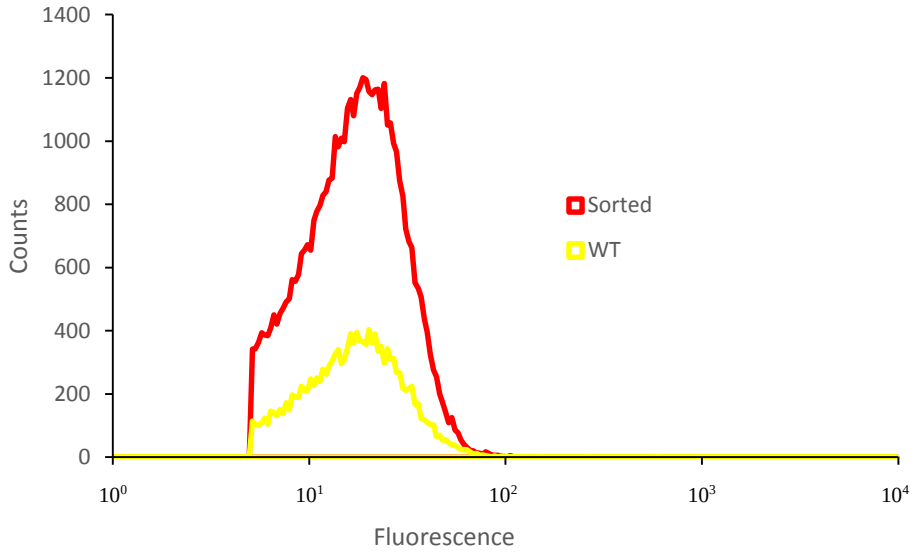

Figure 53. Histogram overlay of FACS data from droplets constructed a mixture of UDP-sugars and cells expressing either WT UGT72B1 (yellow) or the enriched mutant library (red).

This result was disappointing but not entirely unexpected. As an alternative to discovering UGT72B1 mutants with diverse UDP-sugar activities, the enriched library was instead searched for improved activity toward UDP-glucose and esculetin. First, the library was subcloned into pET-28a, from which the mutants could be expressed as C-terminally His-tagged fusion proteins for more facile downstream characterization.

Lysates of the FACS enriched UGT72B1 mutants were prepared in four deep well 96-well microtiter plates (as described in detail in section 5.4) and assayed with 100 μM esculetin and 500 μM UDP-glucose (Fig. 54-57). The first plate was analyzed for 60 minutes, however the negligible additional activity of top mutants near the end of this hour was noted, and subsequent plates were screened for 45 minutes.

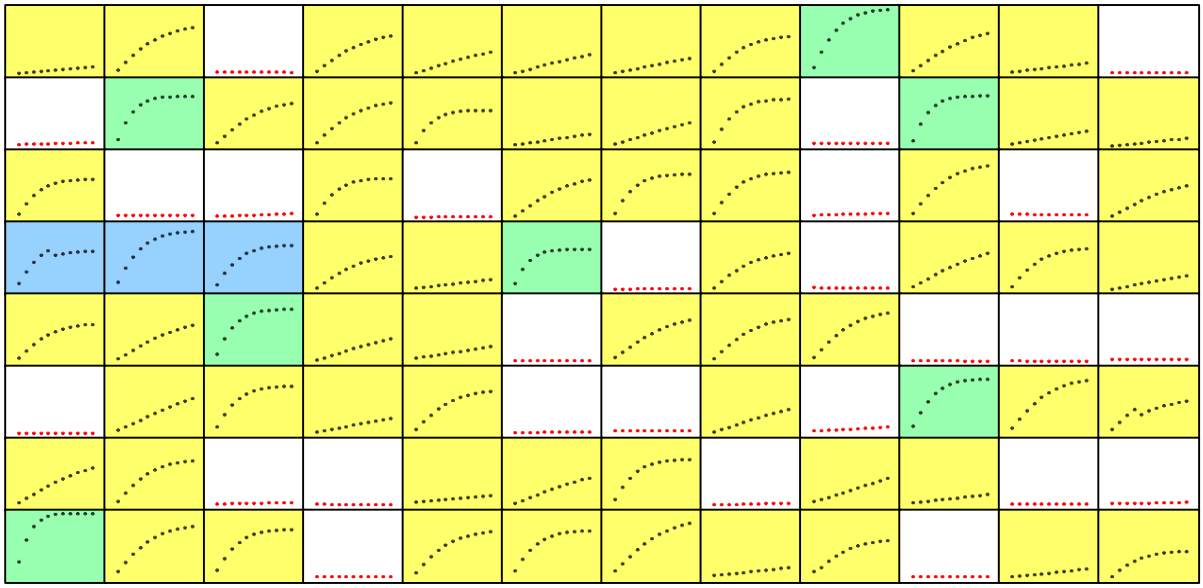

Figure 54. Microtiter plate (#1) screening of UGT72B1 mutant lysates from neutral drift with UDP-glucose. The wild-type controls are shown in blue (wells D1, D2, and D3). Potential hits are colored green, neutral or slightly deleterious mutants are colored yellow, and inactive variants are colored white.

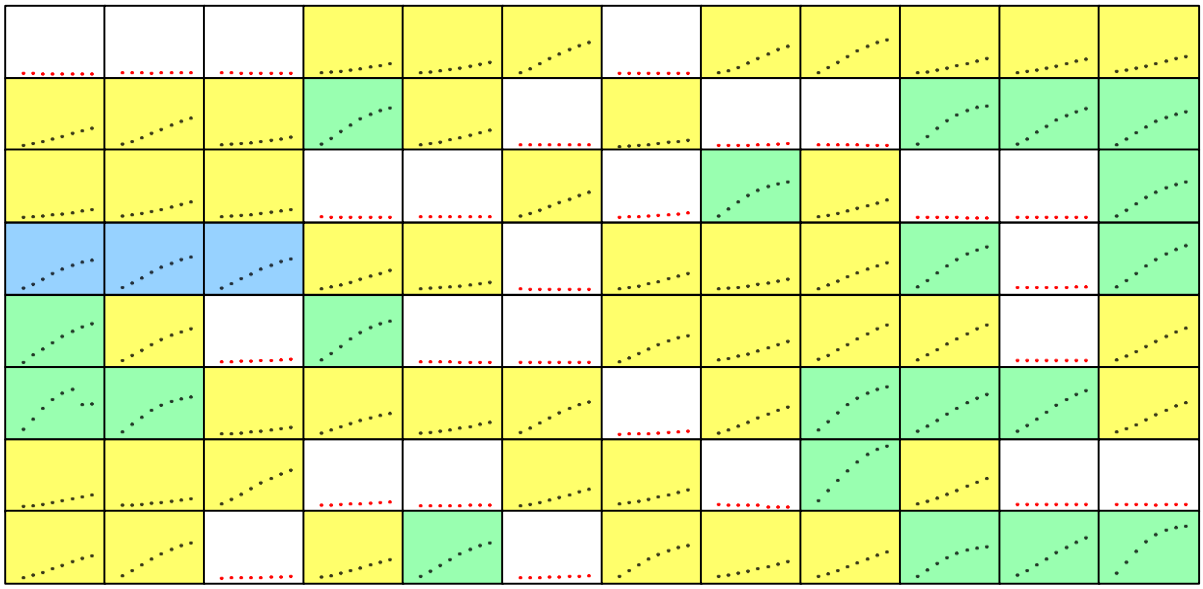

Figure 55. Microtiter plate (#2) screening of UGT72B1 mutant lysates from neutral drift with UDP-glucose. The wild-type controls are shown in blue (wells D1, D2, and D3). Potential hits are colored green, neutral or slightly deleterious mutants are colored yellow, and inactive variants are colored white.

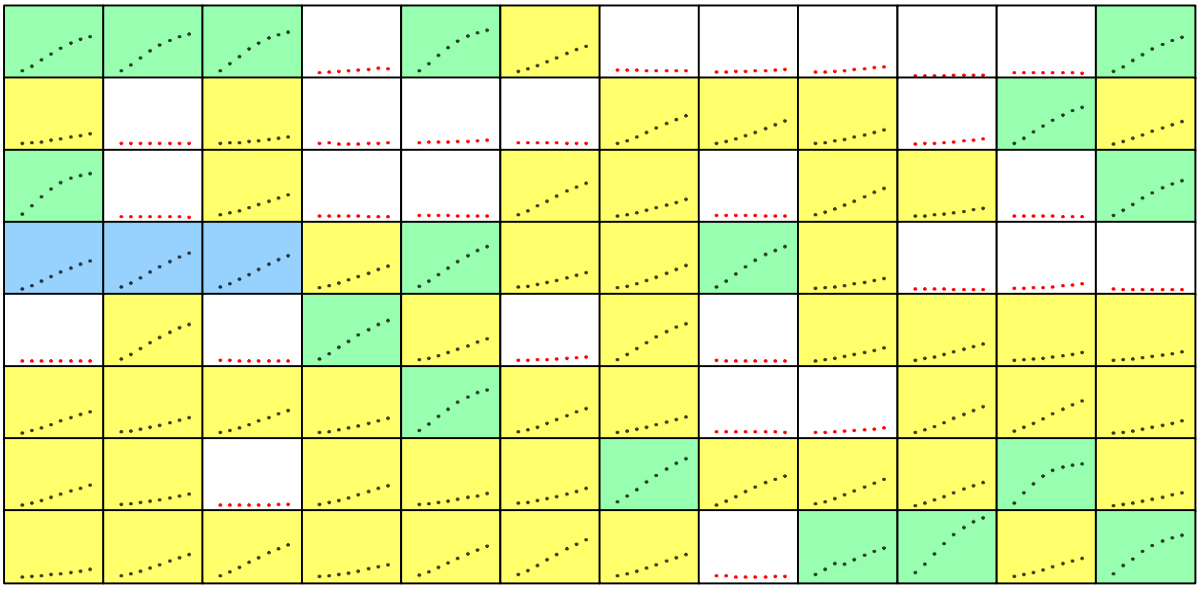

Figure 56. Microtiter plate (#3) screening of UGT72B1 mutant lysates from neutral drift with UDP-glucose. The wild-type controls are shown in blue (wells D1, D2, and D3). Potential hits are colored green, neutral or slightly deleterious mutants are colored yellow, and inactive variants are colored white.

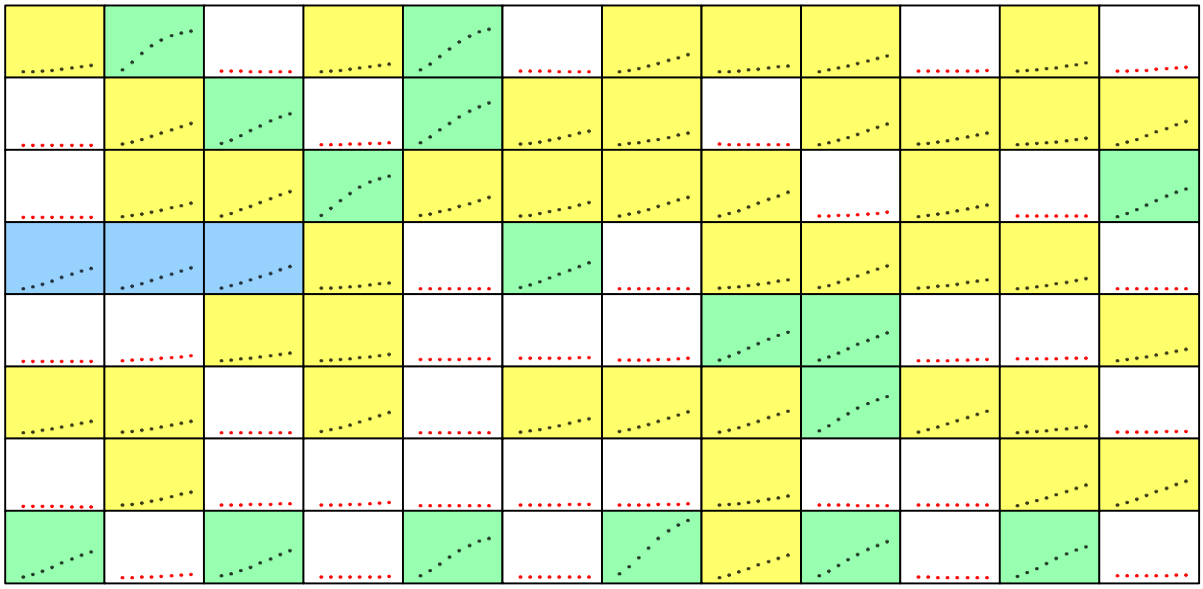

Figure 57. Microtiter plate (#4) screening of UGT72B1 mutant lysates from neutral drift with UDP-glucose. The wild-type controls are shown in blue (wells D1, D2, and D3). Potential hits are colored green, neutral or slightly deleterious mutants are colored yellow, and inactive variants are colored white.

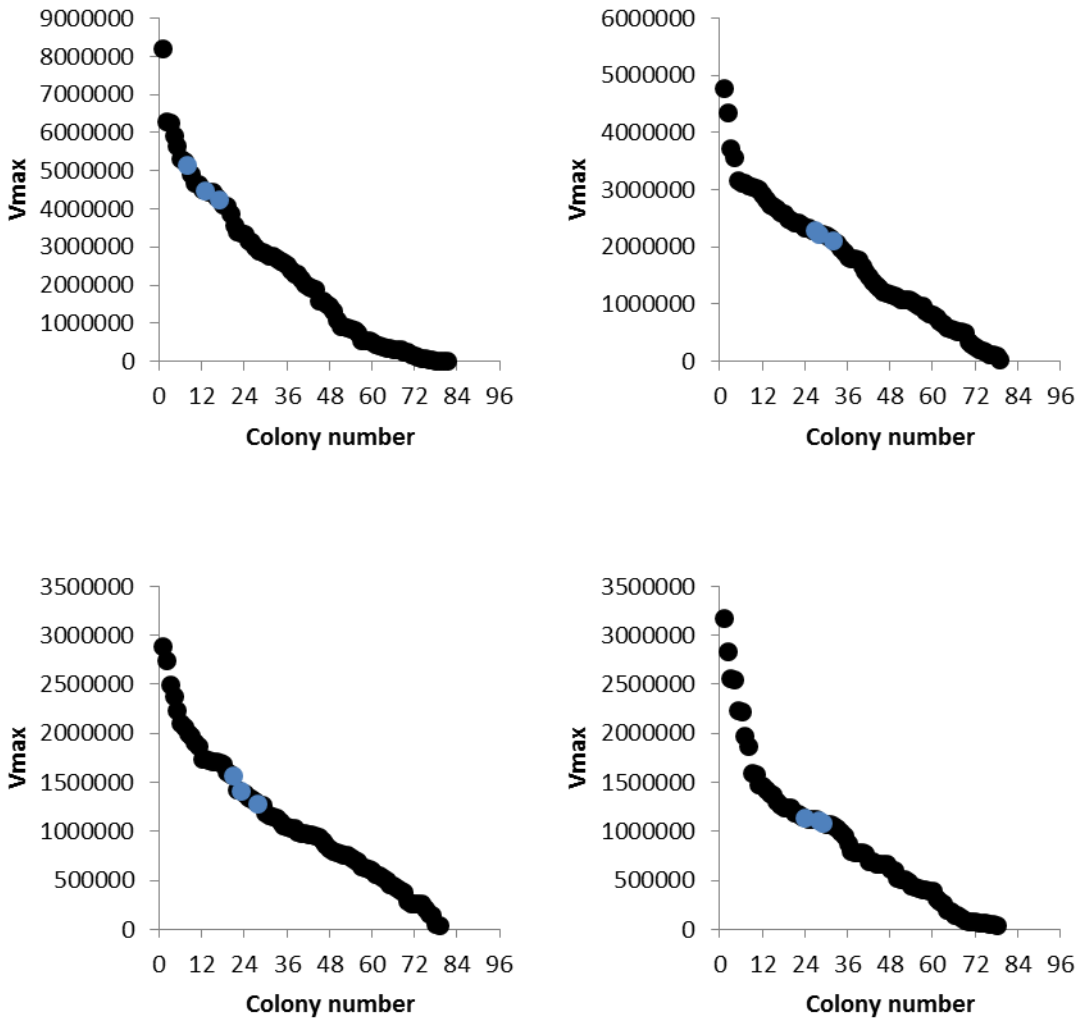

Figure 58. Initial rates from screening UGT72B1 mutants from neutral drift with UDP-glucose (top left: plate 1, top right: plate 2, bottom left: plate 3, bottom right: plate 4). The wild-type controls are shown in blue.

Gratifyingly, 44 of 372 wells (11.8%) displayed significantly higher UGT72B1 activity than the wild-type controls (wells D1-3) and were rescreened under identical conditions (data not shown). 15 mutants were selected for further analysis by DNA sequencing and LC-MS based on their performance during both rounds of microtiter plate screening.

Lysates of these selected mutants were prepared from small scale cultures and assayed with 100 μM esculetin and 500 μM UDP-glucose. The quenched reactions were analyzed by LC-MS, and the ratio of the mass ion counts for the product and substrate were compared to those generated by the wild-type enzyme (Table 1). All mutants performed better than the wild-type, but the mutant originally screened in plate 1 well H1 performed best (Fig. 59).

Table 1. The ratios of esculetin to esculin extracted mass ion counts for all mutants analyzed by LC-MS.

enzyme	339 m/z to 177 m/z ratio
WT UGT72B1	0.651
Plate 1 A9	2.783
Plate 1 B2	1.933
Plate 1 B10	1.535
Plate 1 E3	2.269
Plate 1 F10	0.947
Plate 1 H1	4.341
Plate 2 D12	1.068
Plate 2 F1	0.977
Plate 2 G9	2.596
Plate 3 A2	1.199
Plate 3 A5	1.119
Plate 3 H7	2.324
Plate 4 F5	2.020
Plate 4 G11	1.472
Plate 4 H10	1.396

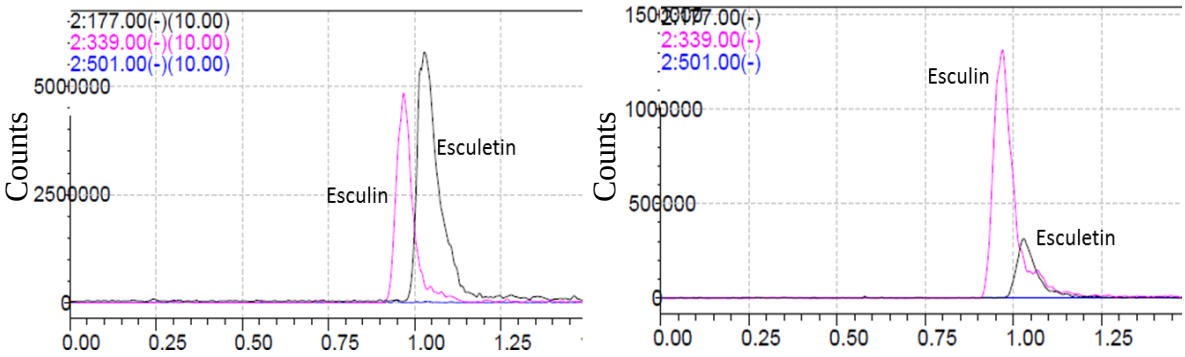

Figure 59. LC-MS analysis of lysates of UGT72B1(left) and the top performing mutant, originally screened in well H1 of plate 1 (right).

Confident that improvements to the kinetic parameters of UGT72B1 had been found, the four best mutants based on LC-MS data were partially purified from the remaining portion of their small scale lysates via batch Ni-NTA affinity purification. Protein quantification via Bradford Assay suggested that the improved activity in lysates was at least partially due to improved protein expression levels; the purification of these mutants yielded 348% (plate 1, well A9), 382% (plate 1, well H1), 173% (plate 2, well G9), and 123% (plate 3, well H7) more protein than the wild-type.

In order to determine if these mutants displayed improved kinetic parameters, 200 µL reactions containing 1 µg of enzyme, 100 µM esculetin, and 500 µM UDP-glucose were analyzed over time, but no improvements to enzymatic activity were detected (Fig. 60). This suggests that the improved reaction rates in lysates are due to increases in heterologous expression under the assay conditions of otherwise neutral UGT72B1 mutants in *E. coli*. SDS-PAGE analysis of the *E. coli* cells overexpressing these mutants confirmed higher expression (Fig. 61).

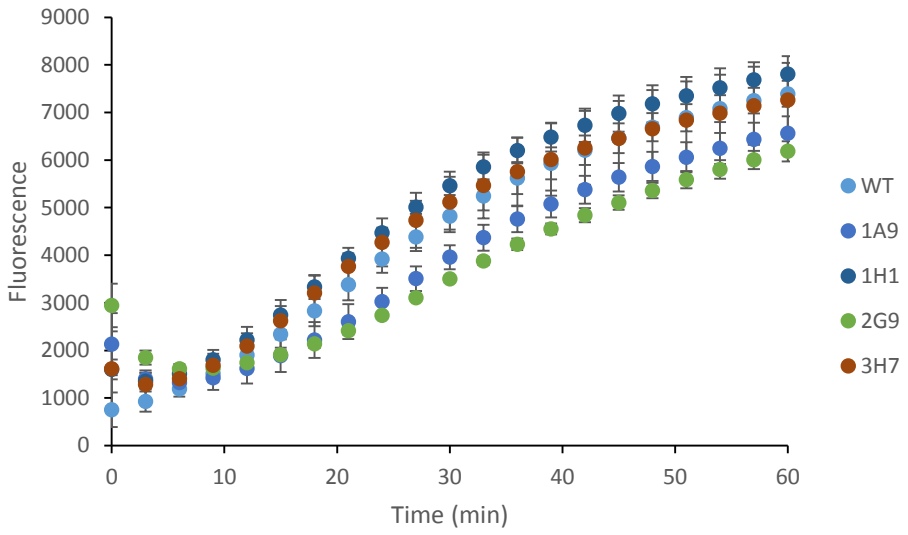

Figure 60. Timecourse experiment for purified UGT72B1 and top performing mutants. Error bars represent standard deviation of the mean (*n* = 3).

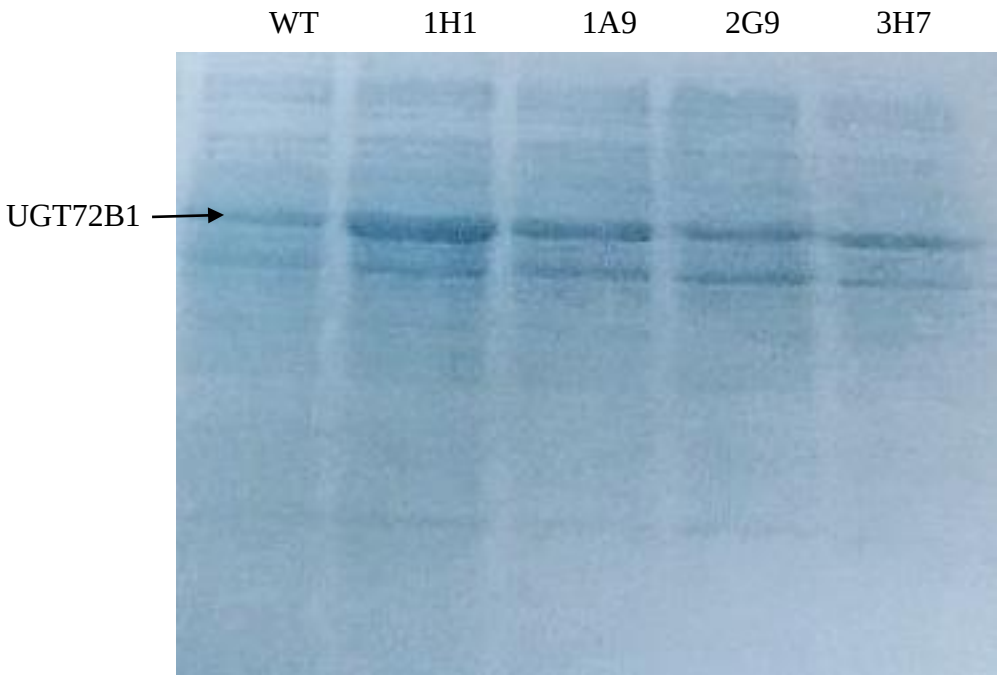

Figure 61. SDS-PAGE analysis of *E. coli* expressing WT UGT72B1 or the mutants from the most active lysates by LC-MS.

It is not feasible to distinguish between improvements in protein expression and improvements in k_{cat} from *E. coli* lysates analyzed in microtiter plates. Additionally, the poor heterologous expression of UGT72B1 in *E. coli* leaves room for considerable further improvement to heterologous expression. Although performing multiple rounds of IVC-FACS and microtiter plate screening would likely ultimately optimize UGT72B1 expression and begin to yield mutants with improved catalytic activity, the cost and time burden of such a project was deemed too high.

Sequencing results for the mutants chosen for LC-MS screening were very interesting. The average nucleotide mutation rate was 4.2 ± 2.9 mutations per gene and the average amino acid mutation rate was 1.7 ± 2.3 mutations per enzyme. As expected, the mutation rates were

significantly lower than those of the initial library. The number and identity of the nucleotide and amino acid mutations varied greatly for the four best mutants (Table 2).

Table 2. Sequencing analysis from the four selected UGT72B1 mutants.

mutant	# nucleotide mutations	# amino acid mutations	nucleotide mutations	amino acid mutations
plate 1 well A9	10	8	A109G, C554T, C691A, T733C, A742G, G1240A, A1302C, T1348A, A1337G, C1391A	T37A, P185L, L231I, I245T, Q248R, V414I, L450M, A464D
plate 1 well H1	7	2	G255C, G315A, C400T, T531C, T591C, C1296T, C1441T	P134S, T491I
plate 2 well G9	2	1	T647A, G723T	F216Y
plate 3 well H7	4	1	C498T, C592G, C642T, A966G	L198V

5.3. Conclusions

The attempt to alter UGT72B1's donor substrate specificity via directed evolution with the IVC-FACS screen failed, despite library enrichment via neutral drift. However, the neutral drift enrichment potential of the IVC-FACS screen was validated, as indicated by the altered population diversity from the unsorted library to the enriched library and by the phenotypic diversity displayed in subsequent microtiter plate analysis.

The attempt to indentify UGT72B1 mutants with improvements in catalytic activity by secondary screening the enriched library in microtiter plates also failed. However, the screen

did not fail due to an inability to measure increased product yields, but rather due to its inability to distinguish between improved protein expression and improved k_{cat}. Despite failing to achieve the desired enzyme engineering goal, mutants with improved heterologous expression in *E. coli* were isolated. Future UGT72B1 engineering efforts should work from a gene codon optimized for expression in *E. coli*.

Neutral drift is a technique well suited to IVC-FACS screens, because the low accuracy of FACS measurements, exacerbated by the polydispersity of bulk generated emulsion microdroplets, isn't particularly detrimental to purifying selections. As described in Chapter 4, neutral drift can be used to enrich libraries with even higher mutation rates to achieve high quality mutant libraries.

5.4. Experimental Section

Unless otherwise stated, all materials and reagents were of the highest grade possible and purchased from Sigma (St. Louis, MO). Isopropyl β-D-thiogalactoside (IPTG) was from Calbiochem (Gibbstown, NJ). Primers were ordered from Integrated DNA Technologies (Coralville, IA).

Mutant Library Construction

The error-prone PCR library was prepared using the Stratagene GeneMorph II Random Mutagenesis Kit, as described by the manufacturer, using 10 ng of pETduet-*UGT72B1-sfGFP* as template. The *UGT72B1* gene was amplified with the 72B1_F and 72B1_R primers. Amplified product was digested with *Nco*I and *Hin*dIII, separated by agarose gel electrophoresis (0.8% w/v agarose), purified with the QIAquick Gel Extraction Kit (QIAgen,

Valenica, CA), and ligated into similarly treated pETduet-*sfGFP* using T4 DNA ligase (New England Biolabs, MA). The ligation mixture was transformed into E. cloni 10G ELITE electrocompetent cells (Lucigen, Middleton, WI). After one hour, 5 μL of the recovered cells were plated on LB-agar supplemented with 100 μg/mL ampicillin and incubated overnight at 37 °C. The remaining cells were diluted to 10 mL with LB supplemented with 100 μg/mL ampicillin and cultured at 37 °C and 250 rpm overnight, subsequently yielding the mutant library via plasmid mini-preparation. Approximately 2,000 colonies grew on the LB-agar plate, indicating a total library size of 4 x 10^5. Plasmid DNA was prepared from single colonies for DNA sequencing, which revealed that the library had an acceptably high mutation rate of 7.0 ± 3.7 nucleotide mutations per gene. However, 9 of 17 sequenced plasmids were empty vector artifacts, religated across the *Nco*I and *Hind*III restriction sites to yield the hybrid sequence: 5'-CCATGCTT-3'. Adjustment of the ligation conditions failed to yield more properly cloned mutants. Since this artifact would be removed by both FACS sorting and *UGT72B1* specific PCR amplification of DNA from sorted droplets, the artifact was ignored. The library was transformed into E. cloni EXPRESS BL21(DE3) electrocompetent cells (Lucigen, Middleton, WI) and the transformants were cultured in 10 mL LB supplemented with 100 μg/mL ampicillin at 37 °C and 250 rpm overnight. 20 μL of this culture was transferred to 3 mL LB supplemented with 100 μg/mL ampicillin and the freshly inoculated culture was grown to an optical density of 0.6 at 37 °C and 250 rpm. Protein expression was induced by adding IPTG to a final concentration of 1 mM. The culture was shaken at 22 °C for 18 hours, at which point the cells were prepared for emulsification as described in Chapter 2.

Neutral Drift

Emulsions were prepared as described in Chapter 2. The samples were sorted on a MoFlo XDP (Beckman Coulter, CA) cell sorter. Events were triggered on fluorescence excited by the 100 mW 488 nm laser in order to ignore droplets which do not contain a cell lysate. Esculin was excited by a 150 mW 355 nm laser. Events were gated for forward and side scattering to partially offset droplet polydispersity. Approximately 60% of droplets were excluded based on these gating parameters. Sorting was gated according to a negative control expressing pETduet-*KO*-*sfGFP*. The population was sorted as described above, and positive events were collected in 10 mL polypropylene tubes. DNA was purified from the collected droplets by isopropanol precipitation. The DNA was amplified using the *UGT72B1* cloning primers, and the amplified gene was cloned into pETduet-*sfGFP* via *Nco*I and *Hin*dIII restriction sites. The plasmid library was prepared as before, and the high density of colonies grown from 5 μL of the transformed cells suggested a yield of at least 10^6 transformants. Only 3 of 10 sequences contained the properly cloned gene. FACS samples were prepared as before, and the library was further enriched by FACS. After isopropanol precipitation and PCR amplification of the sorted DNA, the enriched library was cloned into pET-28a via the *Nco*I and *Hin*dIII restriction sites. The ligation reaction was transformed into E. cloni 10G ELITE electrocompetent cells (Lucigen, Middleton, WI). After one hour, 5 μL of the recovered cells were plated on LB-agar supplemented with 30 μg/mL kanamycin and incubated overnight at 37 °C. The remaining cells were diluted to 10 mL with LB supplemented with 30 μg/mL kanamycin and cultured at 37 °C and 250 rpm overnight, subsequently yielding the mutant library via plasmid mini-preparation. Approximately 300 colonies grew on the LB-agar plate, indicating a library size

of 6 x 10^4, which was deemed large enough to avoid oversampling inefficiencies in microplate screening. . The library was transformed into E. cloni EXPRESS BL21(DE3) electrocompetent cells (Lucigen, Middleton, WI) and varying volumes of the recovered transformants were plated onto LB-agar plates supplemented with 30 μg/mL kanamycin. Only plates inoculated with 1 or 2 μL of recovered transformants colonies appropriately separated.

Microplate Screening

An Eppendorf epMotion liquid handling machine (Hauppauge, NY) was used for liquid transfer steps. The wells of a round-bottomed 96-deep-well plate (VWR) containing 1 mL LB medium supplemented with 30 μg/mL kanamycin were inoculated with individual colonies of *E. coli* BL21(DE3) pET28a-*UGT72B1* mutants. Wells D1, D2, and D3 on each plate were inoculated with pET28a-*UGT72B1* wildtype. Culture plates were tightly sealed with AeraSeal™ breathable film (Research Products International Corp.) and incubated at 37 °C and 350 rpm for 18 h. 20 μL of each culture was transferred to the corresponding wells of a fresh deep-well-plate containing 1 mL of LB medium supplemented with 30 μg/mL kanamycin. Initial deep well plates were stored as 10% glycerol stocks at -20 °C. The freshly inoculated plate was incubated at 37 °C and 350 rpm for 4 h. Protein expression was induced with 1 mM IPTG and the plate was incubated for 18 h at 22 °C and 350 rpm. Cells were harvested by centrifugation at 5,000 *g* for 20 min and resuspended in 350 uL of activity buffer (50mM Tris-HCl, 10 mM $MgCl_2$, pH 8.0) supplemented with 2 mg/mL lysozyme. The plates were frozen at -80 °C, thawed at room temperature, and the cell debris pelleted by centrifugation at 5,000 *g* for 20 min. 20 μL of lysate from each well was transferred to the corresponding well of a black-bottom microtiter plate and kept on ice. Immediately prior to

screening, 180 μL of master mix was added to each well to give 100 μM esculetin and 500 μM UDP-glucose. The fluorescence emission at 454 nm upon excitation at 336 was monitored over time with a BioTek Hybrid Synergy 4 plate reader (Winooski, VT). Each timepoint was separated by a 5 minute incubation period and a 30 second shake.

Hit Characterization

Glycerol stocks of putative hits were used to inoculate 3 mL LB supplemented with 30 μg/mL kanamycin. These cultures were grown overnight at 37 °C and 250 rpm. Fresh 3 mL LB cultures supplemented with 30 μg/mL kanamycin were inoculated from the overnight cultures. Additionally, plasmid was prepared from the overnight cultures and sent for sequencing. The fresh cultures were grown to optimal density 0.6-0.7, protein expression was induced with IPTG to a concentration of 1 mM, and the cultures were shaken at 22 °C and 250 rpm for 18 hours. 200 μL of each culture was collected, the cells were pelleted by centrifugation and washed with PBS buffer pH 7.4, and the pellets were resuspended in 50 μL 6x SDS-PAGE loading dye. 20 μL of each sample was analyzed by SDS-PAGE. To the remaining portion of each culture was added 5 mg of lysozyme. The cultures were frozen at -80 °C, thawed at room temperature, and centrifuged to yield supernatants. These supernatants were used to prepared the 100 μL reactions descried above. These reactions contained 100 μM esculetin and 500 μM UDP-glucose and were quenched with 100 μL ice-cold methanol after 20 minutes at room temperature. These reactions were analyzed on a LCMS-2020 from Shimadzu (Columbia, MA, US). The remaining supernatants were purified by batch purification using PerfectPro Ni-NTA Agarose from 5Prime (Gaithersburg, MD, US). The Bradford Protein Assay Kit from Bio-Rad was used to estimate protein concentration.

Microplate Assay of Purified Mutants

100 μL reactions were performed in a microtiter plate with 100 μM esculetin, 500 μM UDP-glucose, and 1 μg UGT72B1 or UGT72B1 mutants. The reactions were performed in triplicate and fluorescence readings were collected every 3 minutes for an hour with a BioTek Hybrid Synergy 4 plate reader (Winooski, VT) at 336 nm excitation and 454 nm emission.

CHAPTER 6

Summary and Future Work

6.2. Summary

Biocatalyst engineering will be increasingly important to the future of synthetic biology.[1] The development of biocatalysts with improved or altered properties is best accomplished through directed evolution, and therefore progress in synthetic biology is dependent on improvements in directed evolution tools and strategies. The work presented in the preceding chapters has focused on these improvements.

IVC provides *in vitro* assay conditions compatible with FACS sorting, greatly reducing per sample cost and increasing throughput compared to larger volume *in vitro* formats such as lysate analysis in microtiter plates. The two limiting problems of preceding IVC-FACS technology, low protein expression yields and limited scope of compatible fluorogenic reactions, have been overcome through a compartmentalized cell lysis strategy (Chapter 2) and the development of IVC-FACS compatible CEBs (Chapter 3). The viability of this improved IVC-FACS format was confirmed through two model enrichment experiments (Chapter 2) as well as neutral drift enrichment of a UGT72B1 epPCR library and subsequent secondary screening in microtiter plates and by LC-MS (Chapter 5).

CEB development enabled successful detection of endpoint fluorescence dependent on initial analyte concentration for a diverse set of analytes: UDP-glucose, UTP, glucose-1-phosphate, glucose-6-phosphate, sorbitol-6-phosphate, and $NADP^+$. Concentrations of these analytes relevant for directed evolution were detectable by CEB-IVC-FACS. Possible target analytes for future CEBs include diphosphate (terpene synthases and ATP dependent ligases),

adenosine monophosphate (non-ribosomal peptide synthases), adenosine diphosphate (kinases), and S-adenosylhomocysteine (methyltransferases).

While other FACS screens have exploited cellular product entrapment, none of these entrapment reactions are also inherently fluorogenic. Therefore secondary screening must be performed with a separate assay. The inherently fluorogenic UGT72B1 reaction provides a method for secondary microtiter plate screening of all CEB-IVC-FACS detectable analytes.

The mathematical models in Chapter 4 provide optimal mutation rates for epPCR libraries screened directly or after neutral drift. The calculations suggest that neutral drift provides sufficiently large improvements in library quality to be recommended for most directed evolution projects. The development of a universal continuous neutral drift platform for protein folding and stability based off of existing PACE technology could provide a tool for generating libraries with unprecedented quality with minimal effort.

6.2. Future Work

6.2.1 Introduction to Polyketide Synthase Synthetic Biology

Polyketide synthases (PKSs) catalyze the construction of polyketide natural products, a remarkable class of molecules rich in structural diversity and therapeutic value as antimicrobial, antifungal, anticancer, and immunosuppressing drugs.[102] Sales of pharmaceutical polyketides exceed $20 billion per annum.[103] Semisynthetic derivatization has successfully yielded improved pharmaceutical properties,[104] but is limited in scope due to structural features of polyketides which can not be selectively modified through modern

organic chemistry.[105] Combinatorial biosynthesis efforts have focused on engineering PKSs to produce polyketides with modified structural features.

Modular type I PKSs (Fig. 62), which act as enzymatic assembly lines with individual PKS modules responsible for each iteration of polyketide intermediate elongation and processing, have been the focus of most PKS engineering efforts for polyketide combinatorial biosynthesis.[103] Each module consists of a number of catalytic domains,[106] including the acyltransferase (AT) domain responsible for selectively transferring a malonyl group from a molecule of malonyl-CoA to the acyl carrier protein (ACP) domain, which directs polyketide substrates and intermediates to the appropriate catalytic domains. The malonyl-ACP intermediate is recognized by the ketosynthase (KS) domain which condenses the malonyl-ACP intermediate with the ACP-bound polyketide intermediate on the ACP of the preceding module via a decarboxylative Claisen condensation yielding a new ACP-bound polyketide intermediate elongated by two carbons. While the minimal type I PKS module must contain AT, KS, and ACP domains, one or more other domains are often present. Ketoreductase domains reduce the newly formed β-keto group, setting the stereochemistry of the corresponding alcohol, and, when present, the α-alkyl group.[107] The resulting alcohol can be eliminated to yield an alkene by dehydratase (DH) domains, and this alkene can be further reduced to the alkane via enoyl-reductase (ER) domains. Cyclization of reduced polyketide products is performed by a terminal thioesterase (TE) domain. Finally, many polyketides are further decorated by *trans*-acting enzymes including oxidoreductases, glycosyltransferases, and methyltransferases.

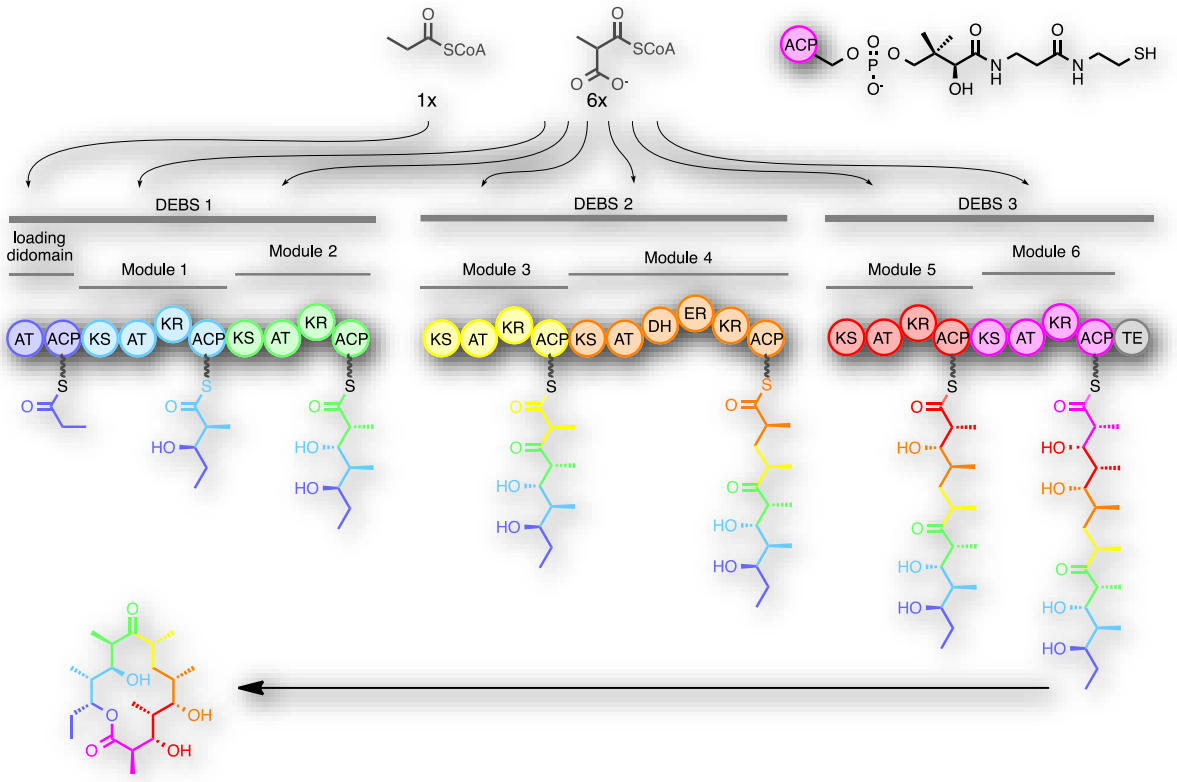

Figure 62. Scheme of 6-deoxyerythronolide B synthase (DEBS), the prototypical type I PKS, including its modular structure, natural substrate choice, protein-bound intermediates, and cyclized product. Substrates and intermediates are attached to the ACP domains via phosphopantetheinyl arms (wavy lines).

Polyketide combinatorial biosynthesis has progressed in parallel with our understanding of substrate specificities for PKS component enzymes. Early work indicated that these type I PKSs typically demonstrate absolute substrate specificity for malonyl-CoA extender units produced naturally within their host, suggesting that one or more domains within a module are unable to process other malonyl-CoAs. Therefor initial efforts to engineer PKSs focused on swapping domains and modules for other PKS domains and modules with different

activity.[108] These efforts largely resulted in undetectable or greatly reduced activity, suggesting protein-protein interactions which are still poorly understood.[109]

The next paradigm in polyketide combinatorial biosynthesis was precursor directed biosynthesis and mutasynthesis, which involved loading initial modules with non-natural substrates or deleting early modules and supplementing cultures with synthetic analogues of simple diketide thioester intermediates.[110] These strategies relied on new understanding of promiscuity in KS and loading module AT domains toward unnatural substrate analogues, and their relative success compared to domain and module swapping strategies led investigators to analyze the substrate specificity of other PKS domains.

The third paradigm in polyketide combinatorial biosynthesis was developed within the Williams Lab. For example, Dr. Irina Koryakina engineered a malonyl-CoA synthetase, MatB, to accept a wide range of unnatural extender units.[111] She then investigated the substrate specificity of several AT domains toward these unnatural substrates, finding in all cases that AT domains which naturally act on the largest malonyl-CoA present within their natural host organism are broadly promiscuous toward larger, unnatural substrates.[112] Early engineering attempts have been limited to low-throughput assays of purified Mod6 mutants, but despite this several point mutations have been found to greatly shift Mod6 AT substrate specificity.

My experimental role in this project was to chemically synthesize unnatural substrates for MatB and PKSs. Additionally, I assisted in the development of the analytical methods for detecting substrate selectivity shifts of mutant ATs. However, the future of this ongoing project may be greatly assisted by my development of the $NADP^+$ dependent CEB-IVC-FACS, which should provide the first suitable screening tool for PKS directed evolution. Without such a

screen, efforts to engineer a panel of PKS ATs with orthogonal, unnatural activities are greatly limited.

6.2.2. Potential Role for CEB-IVC-FACS for PKS Engineering

AT and KS catalysis with assistance from ACPs are the minimal activities necessary for a PKS module. However, these activities yield an extended product with a β-ketone which can subsequently be reduced by a KR domain (Fig. 63). Five of six modules in DEBS have active KR domains, and each of these domains uses NADPH as a hydride source.[107]

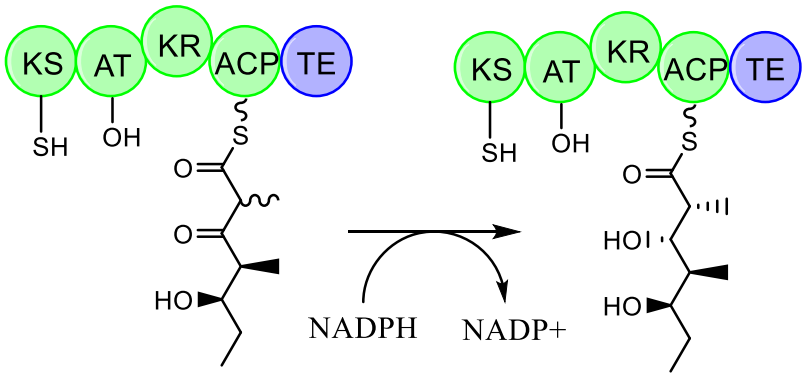

Figure 63. The KR domains of PKSs reduce β-keto thioester intermediates formed by KS catalyzed decarboxylative Claisen condensation. This reduction sets the stereochemistry at the α and β carbons of the intermediate.

Oxidation of NADPH to $NADP^+$ results in an absorbance change that is easily monitored in microtiter plates.[68,69] The Khosla lab used this absorbance change to kinetically characterize individually purified components of *E. coli*'s fatty acid biosynthesis *in vitro*.[113] Recently, they similarly reconstituted DEBS *in vitro* from all six individually purified modules

and used this assay in an attempt to find the rate limiting module by titrating in individual components, although their results were not conclusive.[114] However, PKS activity has not yet been detected in any format suitable for directed evolution, such as *E. coli* lysates.

Screening a PKS module by FACS for the ability to incorporate a non-natural extender unit would be highly advantageous. Although current low-throughput methods are sufficient for identifying amino acid residues that play a role in substrate specificity, they are too costly and slow for the ambitious enzyme engineering goals required for combinatorial biosynthesis.

In addition to slow and costly screening, constructing multiple low diversity libraries is significantly challenging for PKSs due to the large size of their genetic components. All forms of mutagenesis are more challenging and prone to failure on such long genes. In addition to the obvious throughput advantage of FACS, fewer libraries with higher diversity could be generated and screened.

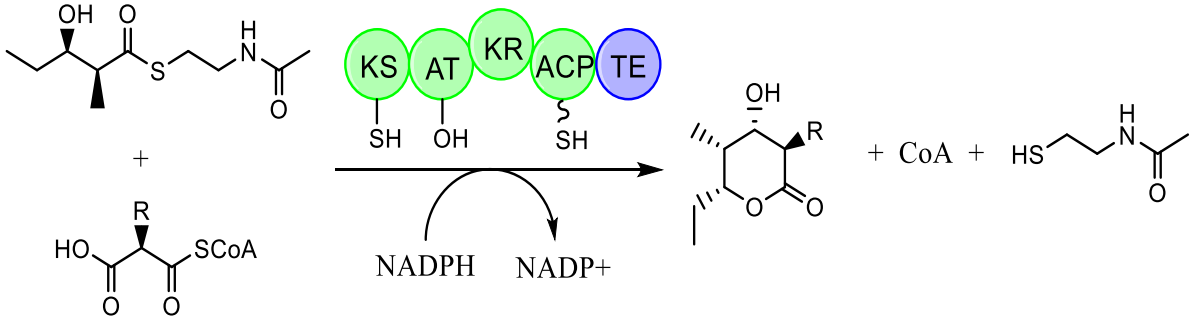

Figure 64. Reaction scheme for a PKS module with a diketide-SNAC surrogate substrate. The stereochemistry of the diketide-SNAC and the triketide lactone are those for DEBS Mod6.

In the triketide-generating reactions used to screen individual PKS modules (Fig. 64) there are up to four products: a cyclic triketide lactone, $NADP^+$, CoA, and N-acetylcysteamine (SNAC). Analysis of these substrates suggests that only *in vitro* detection of $NADP^+$ is a suitable basis for a FACS screen. SNAC and the triketide lactone are not sufficiently polar to be retained within cells or droplets. CoA is released by the AT domain, but the AT domain does not transfer the malonate directly from malonyl-CoA to the ACP. Rather, the AT first transfers the malonate onto itself, and this intermediate is prone to hydrolysis.[112] Evolving an AT domain by screening for CoA release would result in the enrichment of ATs which enhance this hydrolysis to rates faster than sluggish turnover of the correct product by the entire PKS module. $NADP^+$, however, is only consumed by the enzyme to reduce polyketide intermediates, and these intermediates only exist after successful AT and KS catalysis.

Despite the possibility of activating unnatural extender units *in vivo* via MatB mutants and the success of a recent genetic biosensor for detection of $NADP^+$, *in vivo* screening for $NADP^+$ will not work for enzymes as sluggish as PKS modules because the balance between reduced and oxidized forms of nicotinamide cofactors is carefully maintained by the cell. The k_{cat} of DEBS Mod6 for the suggested reaction is 0.5 min^{-1}.[115]

Detection of such slow activity by CEB-IVC-FACS will be challenging. The sensitivity of the A6PR/PGM/RmlA/UGT72B1 biosensor in bulk emulsion is currently insufficient for Mod6TE detection. A number of improvements can be made to correct this. The RmlA reaction can be made irreversible by the addition of inorganic pyrophosphatase, which is commercially available. Additionally, RmlA's K_M toward glucose-1-phosphate can be improved by directed evolution. The individual CEB components could be added at higher concentrations to

accelerate the reactions and optimal conditions for detection of low concentrations of $NADP^+$ could be explored.

The single largest improvement would likely come from switching to microfluidic production of the emulsion microdroplets. The increased polydispersity of bulk emulsions triples fluorescence variance by FACS relative to microfluidic droplets.[47] Additionally, microfluidic droplets are typically considerably larger than those generated by bulk emulsion, thus further diluting the lysate and potentially decreasing its negative impact on the sensitivity of the A6PR/PGM/RmlA/UGT72B1 CEB toward $NADP^+$. Although previous microfluidic developments were incredibly complex, recently relatively simple and inexpensive devices have been developed for the production of FACS compatible double emulsions.[47] After validating microfluidic droplet production, the A6PR/PGM/RmlA/UGT7B1 CEB-IVC-FACS screen can be extensively used to investigate PKS mutant libraries for altered substrate specificity.

References

1. Porcar, M. Beyond directed evolution: Darwinian selection as a tool for synthetic biology. *Syst. Syn. Biol.* **2010**, *4*, 1-6.

2. Skerker, J.; Lucks, J.; Arkin, A. Evolution, ecology and the engineered organism: lessons for synthetic biology. *Genome Biol.* **2009**, *10(11)*, 114.

3. Cameron, D. E.; Bashor, C. J.; Collins, J. J. A brief history of synthetic biology. *Nat. Rev. Micro.* **2014**, *12*, 381-390.

4. Lienert, F.; Lohmueller, J. J.; Garg, A.; Silver, P. A. Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 95-107.

5. Ro, D.; Paradise, E. M.; Ouellet, M.; Fisher, K. J.; Newman, K. L.; Ndungu, J. M.; Ho, K. A.; Eachus, R. A.; Ham, T. S.; Kirby, J.; Chang, M. C. Y.; Withers, S. T.; Shiba, Y.; Sarpong, R.; Keasling, J. D. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **2006**, *440*, 940-943.

6. Rogers, C.; Oldroyd, G. E. D. Synthetic biology approaches to engineering the nitrogen symbiosis in cereals. *J. Exp. Bot.* **2014**, *65*, 1939-1946.

7. Carlson, R. The U.S. Bioeconomy in 2012 reached $350 billion in revenues, or about 2.5% of GDP. *http://www.synthesis.cc/2014/01/the-us-bioeconomy-in-2012.html* (accessed 08/10, **2014**).

8. Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K. Engineering the third wave of biocatalysis. *Nature* **2012**, *485*, 185-194.

9. Liu, X.; Sheng, J.; Curtiss III, R. Fatty acid production in genetically modified cyanobacteria. *PNAS* **2011**.

10. Tibrewal, N.; Tang, Y. Biocatalysts for Natural Product Biosynthesis. *Annu. Rev. Chem. Biomol. Eng.* **2014**, *5*, 347-366.

11. Nakamura, C. E.; Whited, G. M. Metabolic engineering for the microbial production of 1,3-propanediol. *Curr. Opin. Biotechnol.* **2003**, *14*, 454-459.

12. Arnold, F.; Moore, J. In *Optimizing industrial enzymes by directed evolution;* Springer Berlin Heidelberg: **1997**; Vol. 58, pp 1-14.

13. Wang, C.; Oh, M.; Liao, J. C. Directed Evolution of Metabolically Engineered Escherichiacoli for Carotenoid Production. *Biotechnol. Prog.* **2000**, *16*, 922-926.

14. Alberstein, M.; Eisenstein, M.; Abeliovich, H. Removing allosteric feedback inhibition of tomato 4-coumarate:CoA ligase by directed evolution. *Plant J.* **2012**, *69*, 57-69.

15. Brustad, E. M.; Arnold, F. H. Optimizing non-natural protein function with directed evolution. *Curr. Opin. Chem. Biol.* **2011**, *15*, 201-210.

16. Romero, P. A.; Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866-876.

17. Fasan, R.; Meharenna, Y. T.; Snow, C. D.; Poulos, T. L.; Arnold, F. H. Evolutionary History of a Specialized P450 Propane Monooxygenase. *J. Mol. Biol.* **2008**, *383*, 1069-1080.

18. Esvelt, K. M.; Carlson, J. C.; Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **2011**, *472*, 499-503.

19. Savile, C. K.; Janey, J. M.; Mundorff, E. C.; Moore, J. C.; Tam, S.; Jarvis, W. R.; Colbeck, J. C.; Krebber, A.; Fleitz, F. J.; Brands, J.; Devine, P. N.; Huisman, G. W.; Hughes, G. J. Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **2010**, *329*, 305-309.

20. Goldsmith, M.; Tawfik, D. S. Directed enzyme evolution: beyond the low-hanging fruit. *Curr. Opin. Struct. Biol.* **2012**, *22*, 406-412.

21. Patrick, W. M.; Firth, A. E. Strategies and computational tools for improving randomized protein libraries. *Biomol. Eng.* **2005**, *22*, 105-112.

22. Angov, E.; Hillier, C. J.; Kincaid, R. L.; Lyon, J. A. Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host. *PLoS ONE* **2008**, *3*, e2189.

23. Williams, G. J.; Zhang, C.; Thorson, J. S. Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution. *Nat. Chem. Biol.* **2007**, *3*, 657-662.

24. Ruff, A. J.; Dennig, A.; Schwaneberg, U. To get what we aim for: progress in diversity generation methods. *FEBS J.* **2013**, *280*, 2961-2978.

25. Hogrefe, H. In *Fine-Tuning Enzyme Activity Through Saturation Mutagenesis;* Braman, J., Ed.; Humana Press: **2010**; Vol. 634, pp 271-283.

26. McCullum, E.; Williams, B. R.; Zhang, J.; Chaput, J. In *Random Mutagenesis by Error-Prone PCR;* Braman, J., Ed.; Humana Press: **2010**; Vol. 634, pp 103-109.

27. Coco, W. M.; Levinson, W. E.; Crist, M. J.; Hektor, H. J.; Darzins, A.; Pienkos, P. T.; Squires, C. H.; Monticello, D. J. DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotech.* **2001**, *19*, 354-359.

28. Hiraga, K.; Arnold, F. H. General Method for Sequence-independent Site-directed Chimeragenesis. *J. Mol. Biol.* **2003**, *330*, 287-296.

29. Hoebenreich, S.; Zilly, F. E.; Acevedo-Rocha, C.; Zilly, M.; Reetz, M. T. Speeding up Directed Evolution: Combining the Advantages of Solid-Phase Combinatorial Gene Synthesis with Statistically Guided Reduction of Screening Effort. *ACS Synth. Biol.* **2014**.

30. Reetz, M. T.; Wu, S. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun.* **2008**, 5499-5501.

31. Miyazaki, K. Chapter seventeen - MEGAWHOP Cloning: A Method of Creating Random Mutagenesis Libraries via Megaprimer PCR of Whole Plasmids. *Meth. Enzymol.* **2011**, *498*, 399-406.

32. Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotech.* **2007**, *25*, 338-344.

33. Smith, M.; Arnold, F. In *Noncontiguous SCHEMA Protein Recombination;* Gillam, E. M. J., Copp, J. N. and Ackerley, D., Eds.; Springer New York: **2014**; Vol. 1179, pp 345-352.

34. Reetz, M. T.; Kahakeaw, D.; Lohmer, R. Addressing the Numbers Problem in Directed Evolution. *ChemBioChem* **2008**, *9*, 1797-1804.

35. Bershtein, S.; Tawfik, D. S. Ohno's Model Revisited: Measuring the Frequency of Potentially Adaptive Mutations under Various Mutational Drifts. *Mol. Biol. Evol.* **2008**, *25*, 2311-2318.

36. Drummond, D. A.; Iverson, B. L.; Georgiou, G.; Arnold, F. H. Why High-error-rate Random Mutagenesis Libraries are Enriched in Functional and Improved Proteins. *J. Mol. Biol.* **2005**, *350*, 806-816.

37. Aharoni, A.; Griffiths, A. D.; Tawfik, D. S. High-throughput screens and selections of enzyme-encoding genes. *Curr. Opin. Chem. Biol.* **2005**, *9*, 210-216.

38. Dietrich, J. A.; McKee, A. E.; Keasling, J. D. High-Throughput Metabolic Engineering: Advances in Small-Molecule Screening and Selection. *Annu. Rev. Biochem.* **2010**, *79*, 563-590.

39. Yang, G.; Rich, J. R.; Gilbert, M.; Wakarchuk, W. W.; Feng, Y.; Withers, S. G. Fluorescence Activated Cell Sorting as a General Ultra-High-Throughput Screening Method for Directed Evolution of Glycosyltransferases. *J. Am. Chem. Soc.* **2010**, *132*, 10570-10577.

40. Griffiths, Andrew D. Tawfik,Dan S. Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. *EMBO J.* **2003**, *22*, 24-35.

41. Chen, I.; Dorr, B. M.; Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *PNAS* **2011**, *108*, 11399-11404.

42. Bernath, K.; Hai, M.; Mastrobattista, E.; Griffiths, A. D.; Magdassi, S.; Tawfik, D. S. In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal. Biochem.* **2004**, *325*, 151-157.

43. Mastrobattista, E.; Taly, V.; Chanudet, E.; Treacy, P.; Kelly, B. T.; Griffiths, A. D. High-Throughput Screening of Enzyme Libraries: In Vitro Evolution of a β-Galactosidase by Fluorescence-Activated Sorting of Double Emulsions. *Chem. Biol.* **2005** , *12*, 1291-1300.

44. Ghadessy, F. J.; Ong, J. L.; Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *PNAS* **2001**, *98*, 4552-4557.

45. Doi, N.; Kumadaki, S.; Oishi, Y.; Matsumura, N.; Yanagawa, H. In vitro selection of restriction endonucleases by in vitro compartmentalization. *Nucleic Acids Res.* **2004**, *32*, e95-e95.

46. Tawfik, D. S.; Griffiths, A. D. Man-made cell-like compartments for molecular evolution. *Nat. Biotech.* **1998**, *16*, 652-656.

47. Zinchenko, A.; Devenish, S. R. A.; Kintses, B.; Colin, P.; Fischlechner, M.; Hollfelder, F. One in a Million: Flow Cytometric Sorting of Single Cell-Lysate Assays in Monodisperse Picolitre Double Emulsion Droplets for Directed Evolution. *Anal. Chem.* **2014**, *86*, 2526-2533.

48. Kaltenbach, M.; Devenish, S. R. A.; Hollfelder, F. A simple method to evaluate the biochemical compatibility of oil/surfactant mixtures for experiments in microdroplets. *Lab Chip* **2012**, *12*, 4185-4192.

49. Hall, B. G. Experimental evolution of Ebg enzyme provides clues about the evolution of catalysis and to evolutionary potential. *FEMS Microbiol. Lett.* **1999**, *174*, 1-8.

50. Theberge, A.; Courtois, F.; Schaerli, Y.; Fischlechner, M.; Abell, C.; Hollfelder, F.; Huck, W. Microdroplets in Microfluidics: An Evolving Platform for Discoveries in Chemistry and Biology. *Angew. Chem. Int. Ed.* **2010**, *49*, 5846-5868.

51. Agresti, J. J.; Antipov, E.; Abate, A. R.; Ahn, K.; Rowat, A. C.; Baret, J.; Marquez, M.; Klibanov, A. M.; Griffiths, A. D.; Weitz, D. A. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *PNAS* **2010**, *107*, 4004-4009.

52. Fallah-Araghi, A.; Baret, J.; Ryckelynck, M.; Griffiths, A. D. A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab Chip* **2012**, *12*, 882-891.

53. Aharoni, A.; Amitai, G.; Bernath, K.; Magdassi, S.; Tawfik, D. S. High-Throughput Screening of Enzyme Libraries: Thiolactonases Evolved by Fluorescence-Activated Sorting of Single Cells in Emulsion Compartments. *Chem. Biol.* **2005** , *12*, 1281-1289.

54. Kintses, B.; Hein, C.; Mohamed, M.; Fischlechner, M.; Courtois, F.; Lainé, C.; Hollfelder, F. Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution. *Chem. Biol.* **2012**, *19*, 1001-1009.

55. Lim, E.; Baldauf, S.; Li, Y.; Elias, L.; Worrall, D.; Spencer, S. P.; Jackson, R. G.; Taguchi, G.; Ross, J.; Bowles, D. J. Evolution of substrate recognition across a multigene family of glycosyltransferases in Arabidopsis. *Glycobiology* **2003**, *13*, 139-145.

56. Miller, O. J.; Bernath, K.; Agresti, J. J.; Amitai, G.; Kelly, B. T.; Mastrobattista, E.; Taly, V.; Magdassi, S.; Tawfik, D. S.; Griffiths, A. D. Directed evolution by in vitro compartmentalization. *Nat. Meth.* **2006**, *3*, 561-570.

57. Riccardi, C.; Nicoletti, I. Analysis of apoptosis by propidium iodide staining and flow cytometry. *Nat. Protoc.* **2006**, *1*, 1458-1461.

58. Yang, G.; Rich, J. R.; Gilbert, M.; Wakarchuk, W. W.; Feng, Y.; Withers, S. G. Fluorescence Activated Cell Sorting as a General Ultra-High-Throughput Screening Method for Directed Evolution of Glycosyltransferases. *J. Am. Chem. Soc.* **2010**, *132*, 10570-10577.

59. Liu, L.; Li, Y.; Liotta, D.; Lutz, S. Directed evolution of an orthogonal nucleoside analog kinase via fluorescence-activated cell sorting. *Nucleic Acids Res.* **2009**, *37*, 4472-4481.

60. Cummins, I.; Steel, P. G.; Edwards, R. Identification of a carboxylesterase expressed in protoplasts using fluorescence-activated cell sorting. *Plant Biotechnol.* **2007**, *5*, 354-359.

61. Meyer, K.; Schonfeld, H. Uber die Unter sheidung des Enterococcus vom Streptococcus viridans und die Beziehunger beider zum Streptococcus lactis. *Zentralbl.Bakteriol.Parasitenkd.Infectionskr.Hyg.Abt.I.Orig* **1926**, *99*, 402-416.

62. Chuard, C.; Reller, L. B. Bile-Esculin Test for Presumptive Identification of Enterococci and Streptococci: Effects of Bile Concentration, Inoculation Technique, and Incubation Time. *J. Clin. Microbiol.* **1998**, *36*, 1135-1136.

63. Brazier-Hicks, M.; Offen, W. A.; Gershater, M. C.; Revett, T. J.; Lim, E.; Bowles, D. J.; Davies, G. J.; Edwards, R. Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. *PNAS* **2007**, *104*, 20238-20243.

64. Palmqvist, E. Mutagenesis of the sugar donor site of the *Arabidopsis thaliana* glycosyltransferase UGT72B1. **2010**.

65. Leemhuis, H.; Stein, V.; Griffiths, A. D.; Hollfelder, F. New genotype–phenotype linkages for directed evolution of functional proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 472-478.

66. Yoon, J.; Kim, B.; Lee, W. J.; Lim, Y.; Chong, Y.; Ahn, J. Production of a Novel Quercetin Glycoside through Metabolic Engineering of Escherichia coli. *Appl. Environ. Microbiol.* **2012**, *78*, 4256-4262.

67. Amstutz, P.; Forrer, P.; Zahnd, C.; Plückthun, A. In vitro display technologies: novel developments and applications. *Curr. Opin. Biotechnol.* **2001**, *12*, 400-405.

68. Smith, B. C.; Hallows, W. C.; Denu, J. M. A continuous microplate assay for sirtuins and nicotinamide-producing enzymes. *Anal. Biochem.* **2009**, *394*, 101-109.

69. Molnos, J.; Gardiner, R.; Dale, G. E.; Lange, R. A continuous coupled enzyme assay for bacterial malonyl–CoA:acyl carrier protein transacylase (FabD). *Anal. Biochem.* **2003**, *319*, 171-176.

70. Uhr, M. L. Coupled enzyme systems: Exploring coupled assays with students. *Biochem. Educ.* **1990**, *18*, 48-50.

71. Sugiarto, G.; Lau, K.; Qu, J.; Li, Y.; Lim, S.; Mu, S.; Ames, J. B.; Fisher, A. J.; Chen, X. A Sialyltransferase Mutant with Decreased Donor Hydrolysis and Reduced Sialidase Activities for Directly Sialylating Lewisx. *ACS Chem. Biol.* **2012**, *7*, 1232-1240.

72. Jiang, J.; Biggins, J. B.; Thorson, J. S. A General Enzymatic Method for the Synthesis of Natural and Unnatural UDP- and TDP-Nucleotide Sugars. *J. Am. Chem. Soc.* **2000**, *122*, 6803-6804.

73. Moretti, R.; Chang, A.; Peltier-Pain, P.; Bingman, C. A.; Phillips, G. N.; Thorson, J. S. Expanding the Nucleotide and Sugar 1-Phosphate Promiscuity of Nucleotidyltransferase RmlA via Directed Evolution. *J. Biol. Chem.* **2011**, *286*, 13235-13243.

74. Moretti, R.; Thorson, J. S. Enhancing the Latent Nucleotide Triphosphate Flexibility of the Glucose-1-phosphate Thymidylyltransferase RmlA. *J. Biol. Chem.* **2007**, *282*, 16942-16947.

75. Moretti, R.; Thorson, J. S. A comparison of sugar indicators enables a universal high-throughput sugar-1-phosphate nucleotidyltransferase assay. *Anal. Biochem.* **2008**, *377*, 251-258.

76. Gao, H.; Leary, J. A. Kinetic measurements of phosphoglucomutase by direct analysis of glucose-1-phosphate and glucose-6-phosphate using ion/molecule reactions and Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Biochem.* **2004**, *329*, 269-275.

77. Figueroa, C. M.; Iglesias, A. A. Aldose-6-phosphate reductase from apple leaves: Importance of the quaternary structure for enzyme activity. *Biochimie* **2010**, *92*, 81-88.

78. Amitai, G.; Gupta, R. D.; Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP Journal* **2007**, *1*, 67-78.

79. Prassler, J.; Thiel, S.; Pracht, C.; Polzer, A.; Peters, S.; Bauer, M.; Nörenberg, S.; Stark, Y.; Kölln, J.; Popp, A.; Urlinger, S.; Enzelberger, M. HuCAL PLATINUM, a Synthetic Fab Library Optimized for Sequence Diversity and Superior Performance in Mammalian Expression Systems. *J. Mol. Biol.* **2011**, *413*, 261-278.

80. Arnold, F. H.; Wintrode, P. L.; Miyazaki, K.; Gershenson, A. How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **2001**, *26*, 100-106.

81. Bosley, A. D.; Ostermeier, M. Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol. Eng.* **2005**, *22*, 57-61.

82. Guo, H. H.; Choe, J.; Loeb, L. A. Protein tolerance to random amino acid change. *PNAS* **2004**, *101*, 9205-9210.

83. Bershtein, S.; Goldin, K.; Tawfik, D. S. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *J. Mol. Biol.* **2008**, *379*, 1029-1044.

84. Kimura, M. *The neutral theory of molecular evolution;* Cambridge University Press: 1984; .

85. Gupta, R. D.; Tawfik, D. S. Directed enzyme evolution via small and effective neutral drift libraries. *Nat. Meth.* **2008**, *5*, 939-942.

86. Evans, B.; Chen, Y.; Metcalf, W.; Zhao, H.; Kelleher, N. Directed Evolution of the Nonribosomal Peptide Synthetase AdmK Generates New Andrimid Derivatives In Vivo. *Chem. Biol.* ***2011***, *18*, 601-607.

87. Nov, Y. Fitness Loss and Library Size Determination in Saturation Mutagenesis. *PLoS ONE* **2013**, *8*, e68069.

88. Bloom, J.; Lu, Z.; Chen, D.; Raval, A.; Venturelli, O.; Arnold, F. Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology* **2007**, *5*, 29.

89. Bloom, J. D.; Silberg, J. J.; Wilke, C. O.; Drummond, D. A.; Adami, C.; Arnold, F. H. Thermodynamic prediction of protein neutrality. *PNAS* **2005**, *102*, 606-611.

90. Taverna, D. M.; Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Genet.* **2002**, *46*, 105-109.

91. Nei, M.; Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **1986**, *3*, 418-426.

92. Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotech.* **1999**, *17*, 691-695.

93. Maxwell, K. L.; Mittermaier, A. K.; Forman-Kay, J. D.; Davidson, A. R. A simple in vivo assay for increased protein solubility. *Protein Sci.* **1999**, *8*, 1908-1911.

94. Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein stability promotes evolvability. *PNAS* **2006**, *103*, 5869-5874.

95. Giver, L.; Gershenson, A.; Freskgard, P.; Arnold, F. H. Directed evolution of a thermostable esterase. *PNAS* **1998**, *95*, 12809-12813.

96. Lauchli, R.; Rabe, K. S.; Kalbarczyk, K. Z.; Tata, A.; Heel, T.; Kitto, R. Z.; Arnold, F. H. High-Throughput Screening for Terpene-Synthase-Cyclization Activity and Directed Evolution of a Terpene Synthase. *Angew. Chem. Int. Ed.* **2013**, *52*, 5571-5574.

97. Kille, S.; Acevedo-Rocha, C.; Parra, L. P.; Zhang, Z.; Opperman, D. J.; Reetz, M. T.; Acevedo, J. P. Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis. *ACS Synth. Biol.* **2013**, *2*, 83-92.

98. Adams, P. D.; Baker, D.; Brunger, A. T.; Das, R.; DiMaio, F.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C. Advances, Interactions, and Future Developments in the CNS, Phenix, and Rosetta Structural Biology Software Systems. *Annu. Rev. Biophys.* **2013**, *42*, 265-287.

99. Esvelt, K. M.; Carlson, J. C.; Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **2011**, *472*, 499-503.

100. Bloom, J.; Romero, P.; Lu, Z.; Arnold, F. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2007**, *2*, 17.

101. Peisajovich, S. G.; Tawfik, D. S. Protein engineers turned evolutionists. *Nat. Meth.* **2007**, *4*, 991-994.

102. Staunton, J.; Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **2001**, *18*, 380-416.

103. Weissman, K. J.; Leadlay, P. F. Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Micro.* **2005**, *3*, 925-936.

104. Ruan, B.; Pong, K.; Jow, F.; Bowlby, M.; Crozier, R. A.; Liu, D.; Liang, S.; Chen, Y.; Mercado, M. L.; Feng, X.; Bennett, F.; von Schack, D.; McDonald, L.; Zaleska, M. M.; Wood, A.; Reinhart, P. H.; Magolda, R. L.; Skotnicki, J.; Pangalos, M. N.; Koehn, F. E.; Carter, G. T.; Abou-Gharbia, M.; Graziani, E. I. Binding of rapamycin analogs to calcium channels and FKBP52 contributes to their neuroprotective activities. *PNAS* **2008**, *105*, 33-38.

105. Williams, G.; Koryakina, I.; McArthur, J.; Draelos, M.; Muddiman, D.; Randal, S. In *Reprogramming the Biosynthesis of Natural Products by Directed Evolution;* American Chemical Society: 2013; Vol. 1125, pp 147-163.

106. Khosla, C.; Tang, Y.; Chen, A. Y.; Schnarr, N. A.; Cane, D. E. Structure and Mechanism of the 6-Deoxyerythronolide B Synthase. *Annu. Rev. Biochem.* **2007**, *76*, 195-221.

107. Piasecki, S.; Taylor, C.; Detelich, J.; Liu, J.; Zheng, J.; Komsoukaniants, A.; Siegel, D.; Keatinge-Clay, A. Employing Modular Polyketide Synthase Ketoreductases as Biocatalysts in the Preparative Chemoenzymatic Syntheses of Diketide Chiral Building Blocks. *Chem. Biol.* **2011**, *18*, 1331-1340.

108. Ruan, X.; Pereda, A.; Stassi, D. L.; Zeidner, D.; Summers, R. G.; Jackson, M.; Shivakumar, A.; Kakavas, S.; Staver, M. J.; Donadio, S.; Katz, L. Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives. *J. Bacteriol.* **1997**, *179*, 6416-6425.

109. Hans, M.; Hornung, A.; Dziarnowski, A.; Cane, D. E.; Khosla, C. Mechanistic Analysis of Acyl Transferase Domain Exchange in Polyketide Synthase Modules. *J. Am. Chem. Soc.* **2003**, *125*, 5366-5374.

110. Jacobsen, J. R.; Hutchinson, C. R.; Cane, D. E.; Khosla, C. Precursor-Directed Biosynthesis of Erythromycin Analogs by an Engineered Polyketide Synthase. *Science* **1997**, *277*, 367-369.

111. Koryakina, I.; McArthur, J.; Randall, S.; Draelos, M. M.; Musiol, E. M.; Muddiman, D. C.; Weber, T.; Williams, G. J. Poly Specific trans-Acyltransferase Machinery Revealed via Engineered Acyl-CoA Synthetases. *ACS Chem. Biol.* **2013**, *8*, 200-208.

112. Koryakina, I.; McArthur, J. B.; Draelos, M. M.; Williams, G. J. Promiscuity of a modular polyketide synthase towards natural and non-natural extender units. *Org. Biomol. Chem.* **2013**, *11*, 4449-4458.

113. Xiao, X.; Yu, X.; Khosla, C. Metabolic Flux between Unsaturated and Saturated Fatty Acids Is Controlled by the FabA:FabB Ratio in the Fully Reconstituted Fatty Acid Biosynthetic Pathway of Escherichia coli. *Biochemistry (N. Y.)* **2013**, *52*, 8304-8312.

114. Lowry, B.; Robbins, T.; Weng, C.; O'Brien, R. V.; Cane, D. E.; Khosla, C. In Vitro Reconstitution and Analysis of the 6-Deoxyerythronolide B Synthase. *J. Am. Chem. Soc.* **2013**, *135*, 16809-16812.

115. Rawlings, B. J. Type I polyketide biosynthesis in bacteria (Part A-erythromycin biosynthesis). *Nat. Prod. Rep.* **2001**, *18*, 190-227.